Science Advances
AAAS

# Supplementary Materials for

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

Yuto Ozaki *et al.*

Corresponding author: Yuto Ozaki, yuto_ozaki@keio.jp; Patrick E. Savage, patrick.savage@auckland.ac.nz

*Sci. Adv.* **10**, eadm9797 (2024)
DOI: 10.1126/sciadv.adm9797

**This PDF file includes:**

FEATURES AND HYPOTHESES

**Literature review 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 included this text here for completeness but placed it in the Supplementary Materials 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 research design.

***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 (*21*, *32–34*, *67*). Why have these features emerged specifically for singing? From the viewpoint of the social bonding hypothesis, slower production rate may help multiple singers synchronize, thus facilitating "formation, strengthening, and maintenance of affiliative connections" (*11*). 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 5kHz (*87*). 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 (*105*). Therefore, heightening $f_0$ can be considered as conveying pitch information via a channel that is as sensitive as possible. When it comes to language, its acoustic realization is controlled by formants. Formants are the resonance of upper harmonics that usually hit the loudness-sensitive frequency range. Thus, in both song and speech, we interpret that they draw upon high frequencies for effective communication, and the elevation of fundamental frequencies, typically in singing, is a consequence of emphasizing pitch contour 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 subglottal pressure at a high level to sustain phonation, which may facilitate raising pitch (*86*).

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 (*106*) 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 (*26*, *107*). Along with other features in infant-directed speech, this difference is argued to play an important role in linguistic and social development (*107*).

Pitch discreteness is often considered a key feature of music (*13*, *24*, *25*, *35*, *37*, *83*). 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 measure quantifies how fast $f_0$ modulates, although we admit this may not fully account for the characteristics of pitch discreteness. For example, recent studies indicate pitch discreteness might relate to the ease of memorization (*13*, *108*), but our measurement does not directly take such effects into account. 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 (*109*)), 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 and

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 (*110–113*).

In relation to the differentiation between song and speech, (*84*) 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 (*41*) 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 where 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 (*114*). In fact, a melodic character of music is often considered to function in communicating mental states (*12*, *42*), and infant-directed singing acts as the indication of emotional engagement (*115*). 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 (*116*, *117*).

***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 of assessing pitch interval size. Although there is a line of research studying musical intervals based on the limited notion of interval as defined within the Western twelve-tone equal-tempered scale (*112*, *118*, *119*); but cf. (*120*, *121*), our study treats intervals more generally as a ratio of frequencies to characterize intervals of song and speech in a unified way.

(*122*) reported that country singers use similar formant frequencies in both song and speech. This is consistent with our pilot analysis (Fig. S2). They further argued that the use of higher formant frequencies (e.g. singer's formant, see also (*123*)) in the Western classical music tradition stemmed from the necessity of the singer's voice to be heard over a loud orchestral accompaniment. Similarly, (*112*) confirmed that speech and song have a similar spectral structure. Although we can find studies showing higher brightness in singing performed by professional singers (*33*, *34*, *124*, *125*), our dataset does not necessarily consist of recordings by professional musicians and, as in the case of (*122*), the prominent use of the high formant frequencies in singing may depend on musical style (but see (*116*) 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 be different between song and speech (*38*).

Cross-species comparative studies identified that the shape of pitch contours is regulated by voice production mechanisms (*28*, *68*). Since both humans and birds use respiratory air pressure to drive sound-producing oscillations in membranous tissues (*28*), their pitch contours tend to result in descending towards the end of the phrase. Although previous studies only compared 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 (*126*, *127*).

**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 ($-|\Delta f_0|$) [*cents/sec*],
4) Timbral brightness (spectral centroid) [*Hz*],
5) Pitch interval size ($f_0$ ratio) [*cents*],
   - Absolute value of pitch ratio converted to the cent scale.
6) Pitch declination (sign of $f_0$ slope) [dimensionless]
   - Sign of the coefficient of robust linear regression fitted to the phrase-wise $f_0$ contour.

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. 8).

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. (*25*, *26*, *32*, *34*, *83*) see the "Literature review of hypotheses and potential mechanisms" section 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 below. 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").

***Pitch height ($f_0$):***
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 (*128*, *129*); 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 (*130*). 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 (one octave is 1200 cents), and the time resolution is 5 ms.

***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; (*131–136*)), 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.

***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 (*137*) 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 data. We use 20 ms as the

standard deviation parameter of the first-order derivative of Gaussian to smooth the noise. This corresponds to the scaling factor of the wavelet function.

***Pitch interval size ($f_0$ ratio) [cent]:***
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 intervals 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 that 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 $f_0$. Fig. S5 shows a schematic overview of our approach.

***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 (*138*), 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.

***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 $f_0$s 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 that the slope of the linear regression tends to be negative. MATLAB's fitlm() function was used to estimate the slope. Fig. 3 illustrates linear models fitted to each phrase.

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

7) Rhythmic regularity (IOI ratio (*39*) deviation) [*dimensionless*],

- This is calculated by the absolute difference between the observed IOI ratios and the nearest mode estimated from the observed IOI ratios. Similar IOI ratios are repeatedly observed if there is regularity in rhythm patterns, and IOI ratios form clusters. We quantify such regularity by measuring how much the IOI ratios are dispersed within clusters. This idea is similar to measuring the variance of the within-cluster. In this analysis, we apply modal clustering, that the cluster centroids are modes of density of data. Various methods for density modes (equivalently zero-dimensional density ridges or degree zero homological features) estimation have been recently proposed (*139–147*). Here, we adopted techniques of topological data analysis. In particular, we use the mean-shift algorithm (*143*) to detect the modes. Gaussian kernels are used and we choose to obtain a bandwidth parameter using (*148*)'s method that selects a bandwidth from the range within which the Betti number (number of modes in this case) is most stable (*148*, *149*). Note that this is not the only way, and other criteria also exist (*141*, *145*) for the bandwidth selection from the viewpoint of topological features. The search space of bandwidth is set to $\sigma\{\log(n)/n\}$ as a minimum following (*145*). The maximum bandwidth value is set to Silverman's rule-of-thumb (*150*) since this bandwidth selection is usually considered oversmoothing (*151*), and this idea was previously also used for ridge detection analysis (*152*). Removing low density data points (outliers) to infer the persistent homology features is recommended (*141*), 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 a 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. Fig. S12 illustrates our approach.

8) Phrase length (duration between two breaths/breaks) (onset-break interval) [*seconds*],
   - It is defined as 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) [*cents*],
   - Like the IOI-ratio deviation, this is calculated by 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) [*cents*],
   - 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 intensity of 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 future studies.

12) Pulse clarity [*dimensionless*],
    - Pulse clarity is calculated using MIRToolbox V1.8.1 (*153*).
13) Timbre noisiness (spectral flatness (*154*, *155*)) [*dimensionless*]
    - Spectral flatness is measured at each acoustic unit, namely inter-onset intervals and onset-break intervals, as in (*38*).

SUPPLEMENTARY METHODS

**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 (cf. "Recording protocol" section). 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 was not a coauthor on the manuscript. All recordings were made by the coauthors themselves singing/speaking/playing instruments.

In addition to the recordings, we collected the texts of recordings which were segmented into acoustic units (e.g., notes, syllables) according to their perceptual center (P-center) (*131–136*). 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 (*132*, *156*). The segmentation by the P-center was expected to reflect the vocalizer's perception of the beginning of acoustic units. Here, we used 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 (*156–159*). 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 ((*160*); 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 point in the audio to reflect the P-center. Those annotations were created by the first author (Ozaki) because the time required to train and ask each collaborator to create these annotations would not have allowed us to recruit enough collaborators for a well-powered analysis.

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

1) Each coauthor sent a text file segmenting their recorded song/speech into acoustic units and breathing breaks (see "Recording protocol" 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 checked Ozaki's annotations (by listening to the recording with "clicks" added to each acoustic unit) and corrected them and/or had Ozaki correct them as needed until the coauthor was satisfied with the accuracy of the annotation.

**Language sample**

***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) *[NB: This was later reduced to 75 coauthors as described in the main text. Here we have preserved the wording of the original Stage 1 Protocol awarded In Principle Acceptance by* Peer Community In Registered Reports (*2*)*]*. 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. "Open call for collaboration to the International Council for Traditional Music (ICTM) email list." section), 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 ((*48*); 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)

***Exclusion criteria and data quality checks***

If coauthors chose to withdraw their collaboration agreement at any point prior to formal acceptance after peer review, their recording set would be excluded (cf. "Collaboration agreement form" section). If their recording quality was too poor to reliably extract features, or if they failed to meet the formatting requirements in the protocol, we would 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 asked 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 would 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 allowed 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 would not share the account password needed to access these recordings with us until we had received In Principle Acceptance.

**Break annotation**

Break was defined as the end of a continuous sequence of sounds before relatively long pauses. Breaks were used to avoid creating inter-onset intervals that did 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 was up to the person who made the recording, but it was expected that pauses would be indicated during sound production.

**Robustness analyses**

***Exclusion of data generated after knowing the hypotheses***

One distinctive aspect of this study is that the authors ourselves generated 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 attempted to control for bias by withholding from analysis of audio data until we confirm the in-principle acceptance of this manuscript. We collected most recordings in a way that coauthors did 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 were 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 might 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 tested 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 tested the direction of effect sizes, so applying the same tests allows us to check if that holds under varying conditions. In case the results of this analysis and the original confirmatory analysis do not match, we interpret our results as not robust (whether due to potential confirmation bias or to other sampling differences) and thus not draw strong conclusions regarding our confirmatory hypotheses.

***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 (*161*). Thus, it is possible that our sample of speakers/singers may not represent independent data points. 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 (*162–164*). While our study included 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 re-ran the same confirmatory analyses using multi-level meta-analysis models (linear mixed-effects models; (*165*)) with each recording set nested in the language family. We performed model comparison using the Akaike Information Criterion (AIC) (*166*) 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 (*167*), although critical assessment of information criteria and model selection methods in light of domain knowledge is also important (*168*). If the choice of model technique qualitatively changes the results of our confirmatory hypothesis testing, we conclude that our results depend on the assumption of the language dependency.

**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 planned to explore questions such as the following through post-hoc exploratory analyses, which could then be used to inform confirmatory analyses in future research:

***More acoustic features:***

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

***Relative differences between features:***

Our confirmatory analysis formally tested 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 planned to rank the magnitude of effect sizes to investigate the most differentiating features and most similar features among the pairs of song and speech.

***Music-language continuum:***

To investigate how music-language relationships vary beyond just song and spoken description, we planned to 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.

***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 planned to explore such effects for all relevant features.

***Linguistic factors:***

We also planned to 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; (*56*)), etc.

***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)

***Reliability of annotation process:***
Each of Ozaki's annotations are based on segmented text provided by the coauthor who recorded it, and Ozaki's annotations were 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 planned to 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 focused 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 were used as the "Reference" datapoint as in Fig. S6, and Savage's annotations (also corrected by the original recorder) correspond to the "Another annotator" data points in Fig. S6. Note however that we predicted that Savage's corrected annotations are 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.

***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 planned to replicate our analyses with Hilton et al. (*26*)'s existing dataset, focusing on the subset of languages that can be directly compared. This subset of languages consists of 5 languages (English, Spanish, Mandarin, Kannada, Polish) represented by matched adult-directed song and speech recordings by ~240 participants (cf. Hilton et al.'s (*26*) Table 1).

Because our main analysis method requires time-intensive manual or semi-manual annotation involving the recorded individual that would not be feasible to apply to Hilton et al.'s (*26*) dataset, we instead relied on our reanalysis of Hilton et al.'s (*26*) data on purely automated features. We then re-analyzed our own data using these same purely automated features. This allowed 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 (*26*) 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 (*26*) data that we previously manually annotated (*37*), allowing us to simultaneously compare differences in our sample vs. Hilton et al.'s (*26*) 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 (cf. "Recording protocol" section) ensures minimal background noise, Hilton et al.'s (*26*) 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 (*26*) using both fully-automated and semi-automatedly 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.

**Applying zero-cell correction to the signs of $f_0$ slopes**

Signs of $f_0$ slopes are dichotomous outcomes (i.e. positive or negative). Therefore, statistical analysis requiring variance becomes uncomputable, including our hypothesis testing using the Gaussian random-effects model meta-analysis, when all values are positive or negative. Zero-cell correction is a workaround to handle such data ((*61*); see also "Alternative analysis approaches for pitch declination (hypothesis 6)" section in the main texts). By employing this method, 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. Our additional procedure is similar to zero-cell corrections but adding 1 instead of 0.5 to all cells.

**Computation of average $f_0$ contours of Fig. 7**

The extracted $f_0$ contours from recordings were normalized to the length of 128 samples using interpolation by Fourier transform and resampling (*169*, *170*). 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 $f_0$ 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.

**Computation of permutation importance**

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 Fig. 5, were used as data after normalization. The average of 1024 realizations of permutation importance values was reported here as the final output. Incidentally, in our experiments, all classifiers achieved average accuracy and F1 score higher than 90 (cf. Table S5).

<u>MANIPULATION OF FEATURES TO DEMONSTRATE OUR DESIGNATED SESOI (COHEN'S D = 0.4)</u>

Following Brysbaert's (*93*) 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 = 0.4. 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 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 us authors seem like 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).

PILOT DATA ANALYSIS

We collected recordings from five coauthors for pilot data analysis Each speaks a different 1st language: English, Japanese, Farsi, Marathi, and Yoruba. Please note that 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. Fig. S2 uses the analysis framework shown in Fig. 8 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 (*93*), while ample theory predicts that song should use more stable pitches than speech (*83*) but sung and spoken pitch interval size should be similar (*28*). However, we would 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 Fig. S3.

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 (*40*, *59*) 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-annotator randomness of annotation is also negligible, 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 superiority of manual annotation, it would be desirable to increase its efficiency in the future via semi-automated methods or crowd-sourcing (though there will likely be tradeoffs between data quality and quantity (*171*)).

## STATISTICAL MODELS AND POWER ANALYSIS PROCEDURE

### Statistical models

The Gaussian random-effects model used in meta-analysis is (*97*, *98*)

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

where $Y_i$ is the effect size (or summary statistics) from the $i$th study, $\theta_i$ is the study-specific population effect size, $\sigma_i^2$ is the variance of the $i$th effect size estimate (e.g., standard error of estimate) which is also called the within-study variance, $\mu_0$ is the population effect size, $\tau^2$ is the between-study variance, and $K$ is the number of studies. In our study, $Y_i$ is the relative effect and $\sigma_i^2$ is its variance estimator (*94*). In addition, the term "studies" usually used in meta-analysis corresponds to recording sets. This model can also be written as

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

### 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 five recording sets are too small to empirically derive reliable effect size estimates, our power analyses used an SESOI corresponding to d = 0.4 (see (*93*, *172*) 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 (*99*) 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 (*173*, *174*). 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 (*99*), and the required sample size does not differ substantially between false discovery rate methods and stepwise family-wise error control methods in certain cases (*175*), 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 Materials and Methods.

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 Table 1 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.

**Power analysis procedure**

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 the "Power analysis" section, 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 $\mu_0 = \Phi(0.4/\sqrt{2}) \approx 0.6114$ and $\Phi(\cdot)$ is the standard cumulative normal distribution.

The power of the Gaussian random-effects model is given by (*176*, *177*)

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

where $Z_\alpha$ satisfies $\Phi(Z_\alpha) = \alpha$ that $\alpha$ is the significance level of the test, and $\delta$ is a 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 $\tau^2$ (between-study variance) which is generally unknown. We use DerSimonian-Laird estimator (*98*, *101*) to estimate $\tau^2$ using pilot data. However, there is the issue that the within-study variance $\sigma_i^2$ 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., $+\infty$) 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 an assumption for $\sigma_i^2$ 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 $\tau^2$ 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

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

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 (*102*)

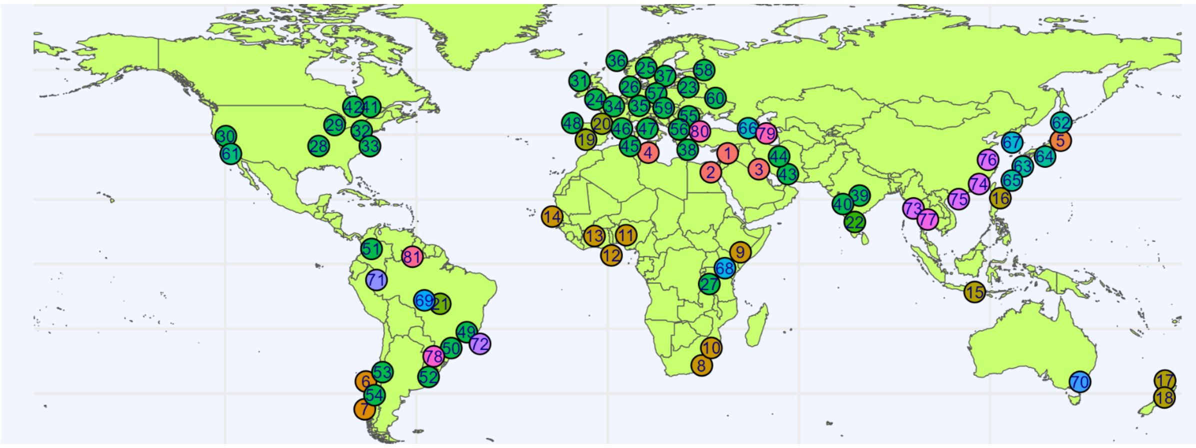
$$K^{1/2}|\bar{Y}_K| \leq C(\alpha, \delta, \sigma_\tau)$$

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 the sample estimate of the mean, and we use the estimated $\mu_0$ instead of the simple average of effect sizes. Here, $\delta$ 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.

SUPPLEMENTARY FIGURES



**Afro-Asiatic**
1 Modern Hebrew [Jerusalem]
2 Modern Hebrew [Tel Aviv]
3 Palestinian Arabic (South Levantine Arabic)
4 Tunisian Arabic

**Ainu**
5 Aynu (Hokkaido Ainu)

**Araucanian**
6 Mapudungun
7 Tsesungún (Huilliche)

**Atlantic-Congo**
8 IsiXhosa (Xhosa) — *Bantu*
9 Kiswahili (Swahili)
10 Ronga
11 Yoruba — *Defoid*
12 Fante (Akan) — *Tano*
13 Twi (Akan)
14 Wolof — *Wolof*

**Austronesian**
15 Balinese
16 Rukai
17 Te Reo Māori (Māori) [Auckland]
18 Te Reo Māori (Māori) [Wellington]

**Basque**
19 Euskara (Basque) [Errenteria]
20 Euskara (Basque) [Hondarribia]

**Cariban**
21 Língua Kuikuro (Kuikúro-Kalapálo)

**Dravidian**
22 Kannada

**Indo-European**
23 Lithuanian — *Baltic*
24 Gaeilge (Irish) — *Celtic*
25 Danish — *Germanic*
26 Dutch [Heemstede]
27 Dutch [Nairobi]
28 English [Indiana]
29 English [Michigan]
30 English [Nevada]
31 English [Newry]
32 English [Pennsylvania]
33 English [Washington D.C.]
34 Flemish (Dutch)
35 German
36 Norwegian
37 Svenska (Swedish)
38 Greek — *Greek*
39 Hindi — *Indic*
40 Marathi
41 Punjabi (Eastern Panjabi)
42 Urdu
43 Western Farsi [Isfahan] — *Iranian*
44 Western Farsi [Tehran]
45 Catalan — *Romance*
46 French
47 Italian
48 Portuguese [Porto]
49 Portuguese [São Paulo]
50 Portuguese [São Paulo]
51 Spanish [Bogotá]
52 Spanish [Montevideo]
53 Spanish [Santiago]
54 Spanish [Osorno]
55 Bulgarian — *Slavic*
56 Macedonian
57 Polish
58 Russian
59 Slovenian
60 Ukrainian

**Iroquoian**
61 Cherokee

**Japonic**
62 Japanese [Hokkaido]
63 Japanese [Hyogo]
64 Japanese [Tokyo]
65 Northern Amami-Oshima

**Kartvelian**
66 Georgian

**Koreanic**
67 Korean

**Nilotic**
68 Luo (dholuo) (Luo (Kenya and Tanzania))

**Nuclear-Macro-Jê**
69 Rikbaktsa

**Pama-Nyungan**
70 Ngarigu

**Pano-Tacanan**
71 Matís

**Puri-Coroado**
72 Puri Kwaytikindo (Puri)

**Sino-Tibetan**
73 Myanmar (Burmese) — *Burmese-Lolo*
74 Cantonese (Yue Chinese) — *Chinese*
75 HainanHua (Min Nan Chinese)
76 Mandarin Chinese

**Tai-Kadai**
77 Thai

**Tupian**
78 Mbyá-Guaraní

**Turkic**
79 North Azerbaijani
80 Turkish

**Yanomamic**
81 Yanomami (Yanomám)

**Fig. S1. Map of the linguistic varieties spoken by our [planned] 81 coauthors as 1st/heritage languages.** Each circle represents a coauthor singing and speaking in their 1st (L1) or heritage language [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. 1 for the final version of the map of 75 linguistic varieties and Acknowledgments section for details of the 6 planned coauthors]. The geographic coordinates represent their hometown where they learned that language. In cases when the language name preferred by that coauthor (ethnonym) differs from the L1 language name in the standardized classification in the Glottolog (*47*), the ethnonym is listed first followed by the Glottolog name in round brackets. Language family classifications (in bold) are based on Glottolog. Square brackets indicate geographic locations for languages represented by more than one coauthor. Atlantic-Congo, Indo-European and Sino-Tibetan languages are further grouped by genus defined by the World Atlas of Language Structures ((*48*); https://wals.info/languoid; Accessed: September 1, 2022; Version number: v2014.2-199-ga9d1a68).

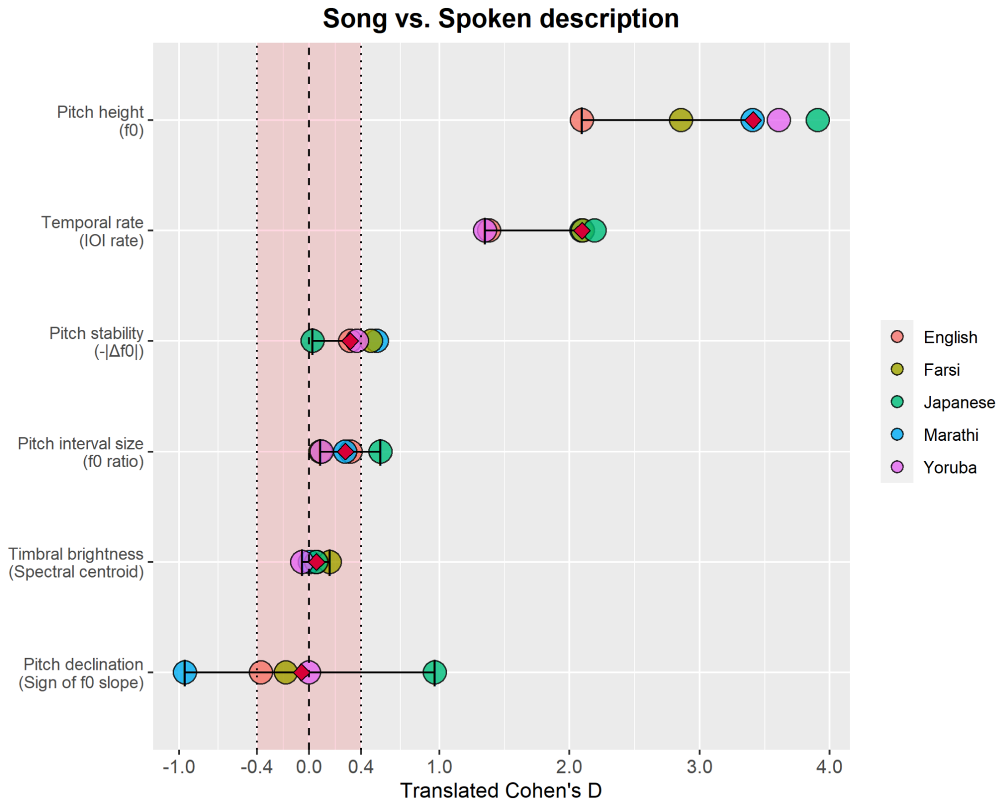



**Fig. S2. Pilot data showing similarities/differences between song and speech for each of the six hypothesized features across speakers of five languages (coauthors McBride, Hadavi, Ozaki, D. Sadaphal, and Nweke).** Red diamonds indicate the population mean and black bars are confidence intervals estimated 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 indicate 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 indicates the "smallest effect size of interest" (SESOI) suggested by (*93*). See Fig. 8 for a schematic of how each effect size is calculated from each pair of sung/spoken recordings.

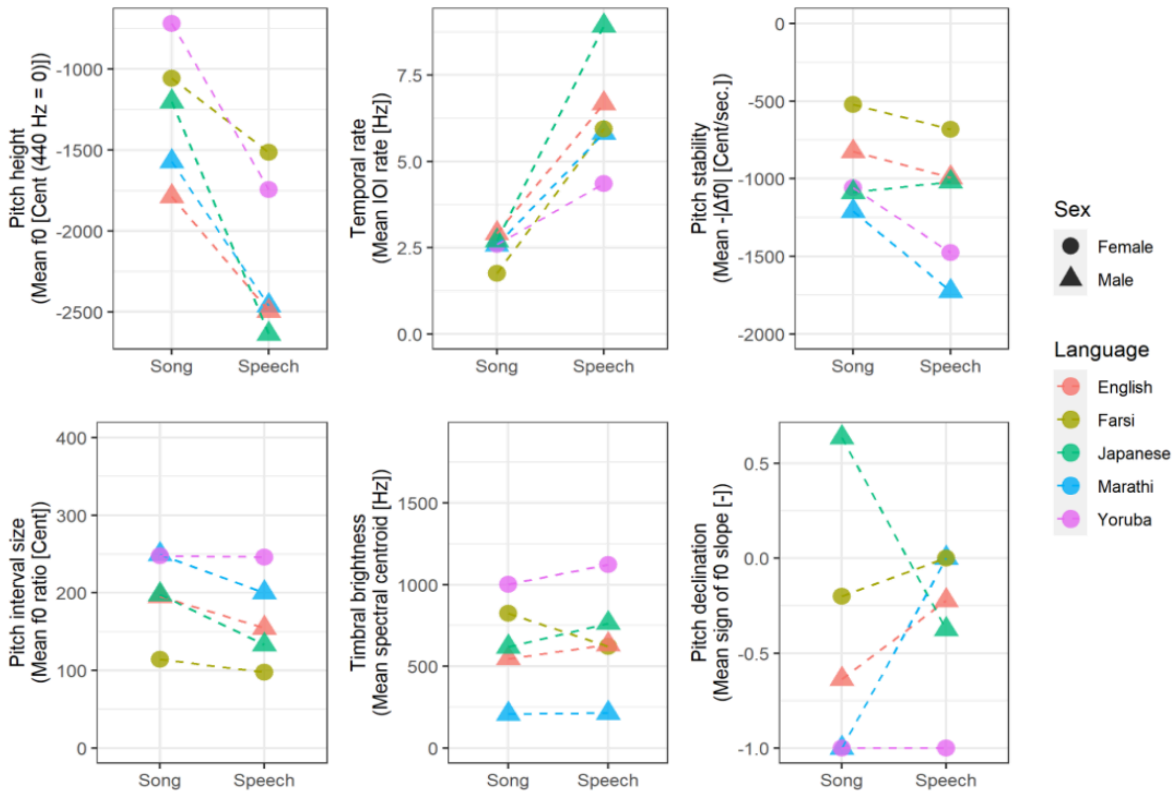



**Fig. S3**. **Alternative visualization of Fig. S2 showing mean values of each feature of song and speech, rather than paired differences.** "Speech" indicates spoken description (not lyric recitation). This figure allows us to visualize some trends not viewable from Fig. S2, such as absolute values 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.)

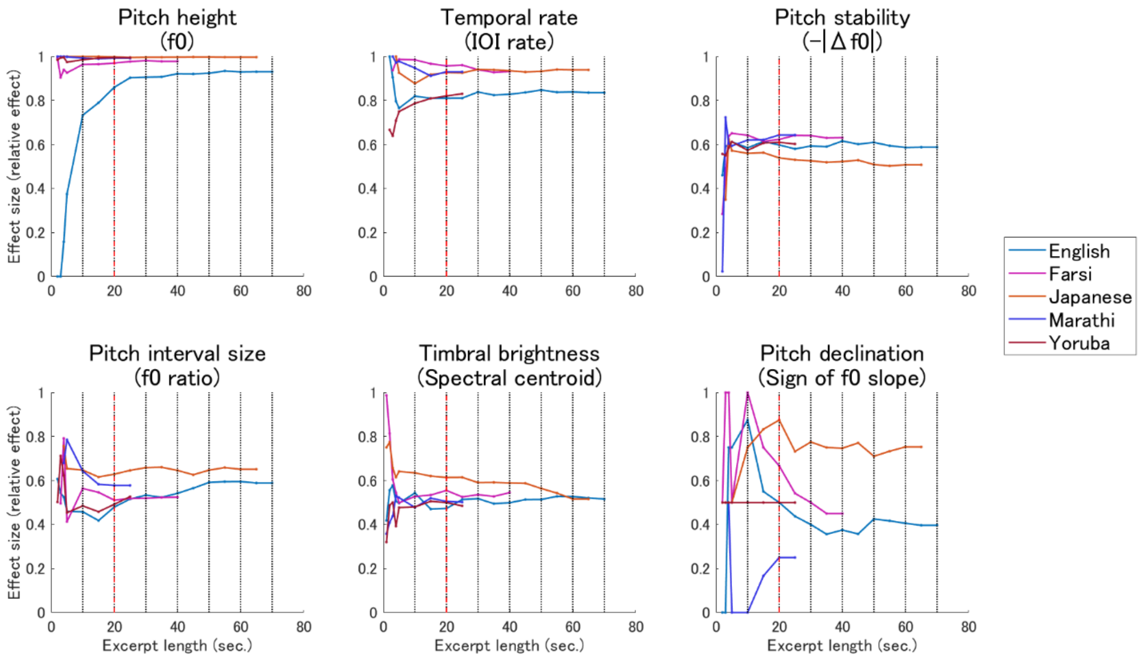



**Fig. S4**. **Relationship between the duration of recording excerpt analyzed and estimated effect size for the 6 features and 5 sets of pilot recordings analyzed in Fig. S2.** Since the length of the pilot recordings ranged from under 30s to over 70s, plots are truncated at the point when there is 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.

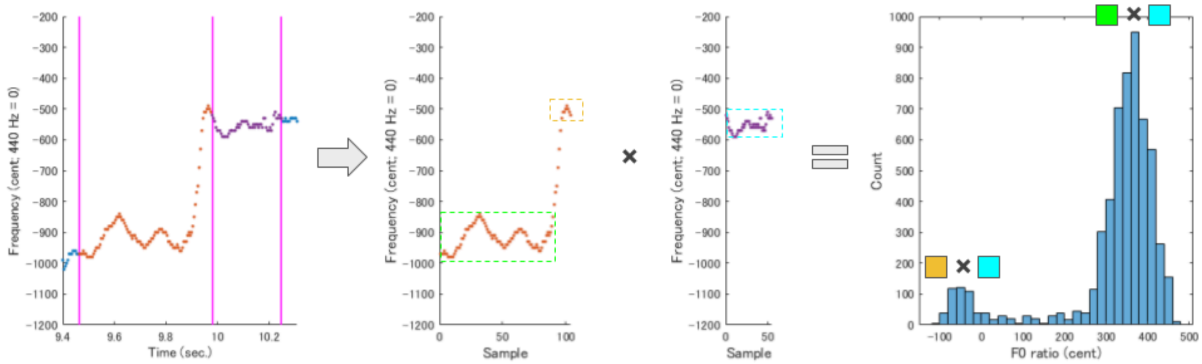



**Fig. S5**. **Process of computing $f_0$ ratios**. The leftmost figure shows an $f_0$ contour which is segmented by three onset times. Then, the pitch ratio of the antecedent segmented $f_0$ contour (orange) and the consequent $f_0$ contour (purple) is calculated by taking exhaustive pairs of samples 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.

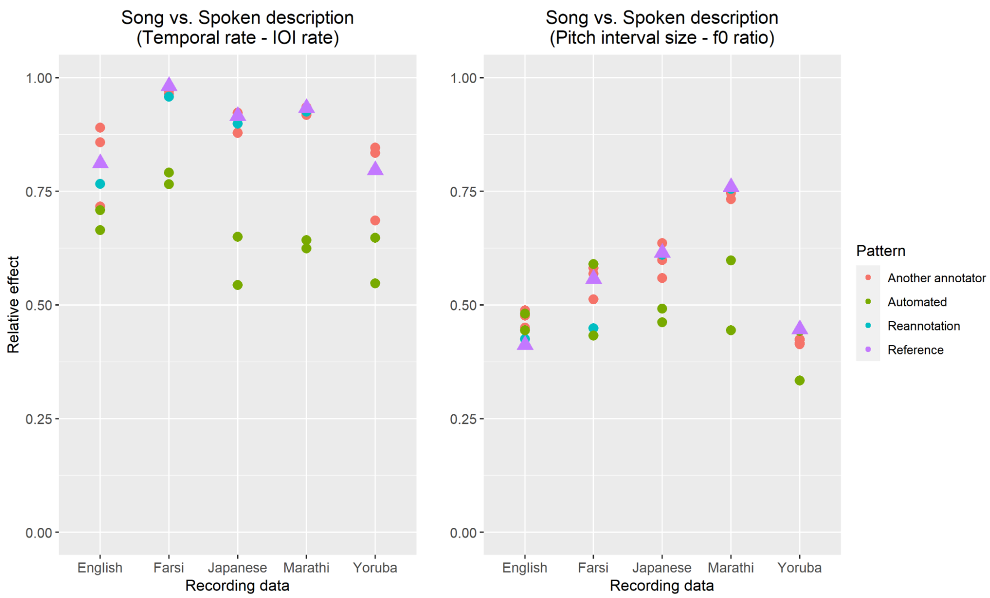



**Fig. S6. Within- and between-annotators randomness of onset annotations including automated methods** (*40*, *59*) **discussed in Section S1.4 "Pilot data analysis".** 10-second excerpts were used. "Reference" is the result of the annotation by the person who originally made the recording, and "Another annotator" is an annotation by other collaborators. "Reannotation" is annotation conducted again by the person who undertook the reference annotation.

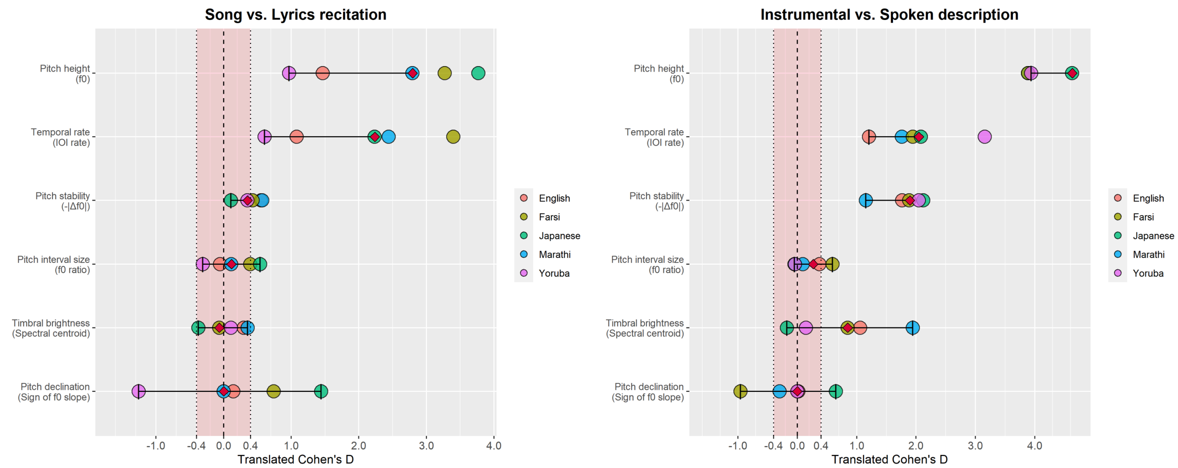



**Fig. S7. Effect sizes of each feature across five languages using the pilot data as in Fig. S2 but with exploratory comparisons with recitation and instrumental recording types.** Refer to Fig. S2 for the explanation of the figure description.

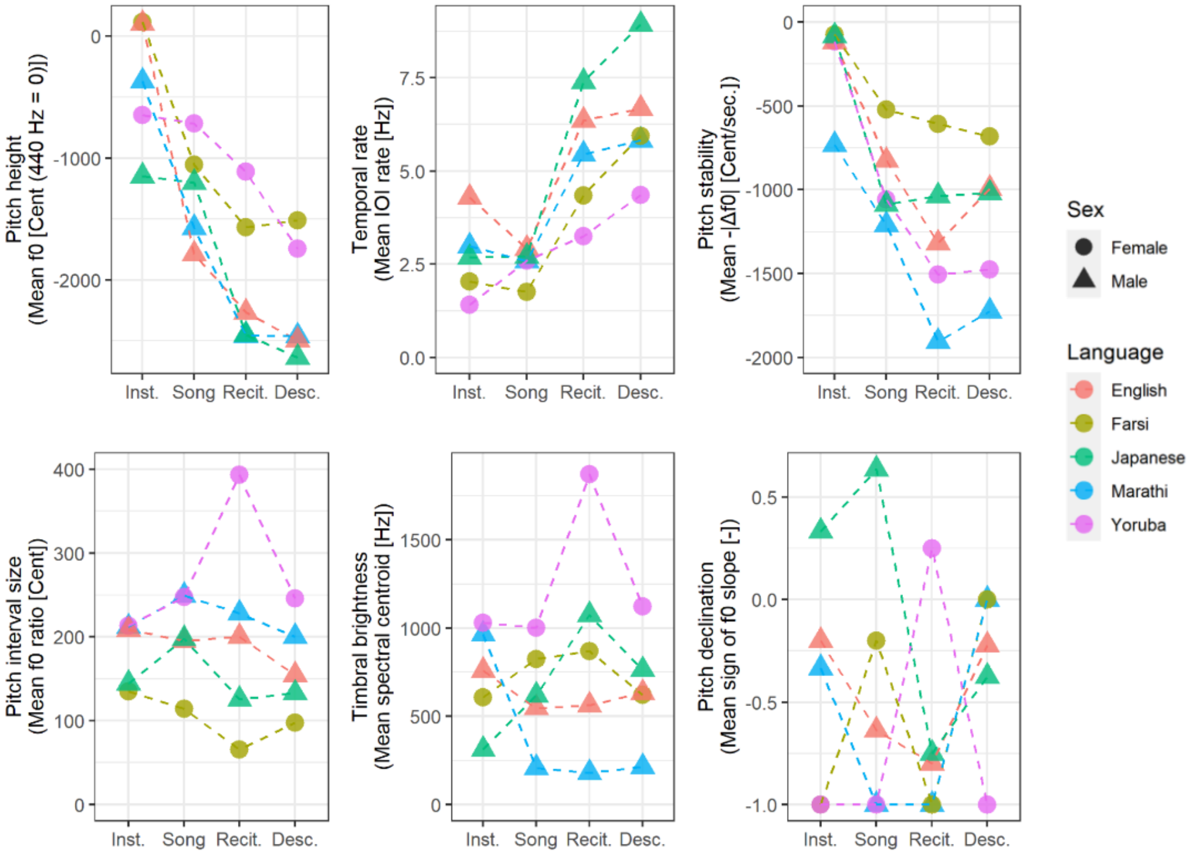



**Fig. S8**. **Mean values of each feature as in Fig. S3 but with all recording types (including recitation and instrumental)**. "Desc." means spoken description, "Recit." means recited lyrics.

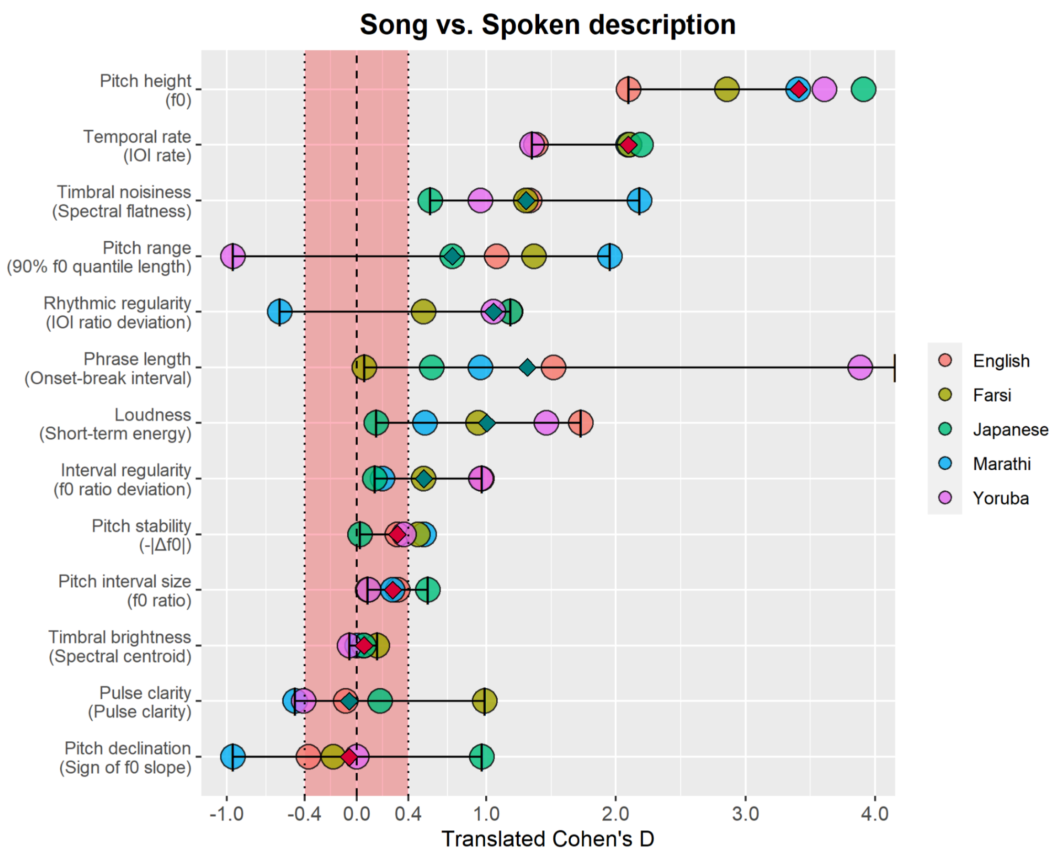



**Fig. S9**. **Effect sizes of each feature across five languages using the pilot data as in Fig. S2 with additional exploratory features.** Green-colored diamonds and two-sided confidence intervals are used for the features for which hypotheses are not specified.

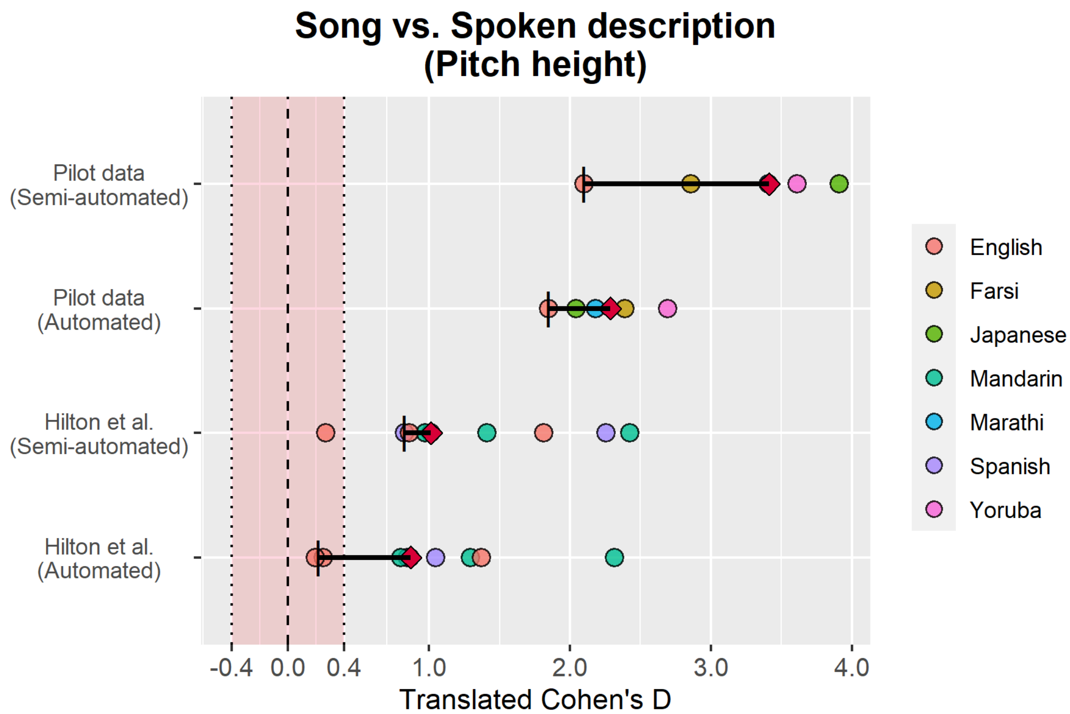



**Fig. S10**. **Pilot analysis of a subset of Hilton et al.'s** (*26*) **data (pairs of adult-directed singing/speaking recordings from n=9 participants speaking English, Spanish, or Mandarin) focusing on pitch height**. (*27*) previously analyzed this subset for preliminary analyses using the same method 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 (*26*) 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 the "Exploring recording representativeness and automated scalability" section for discussion).

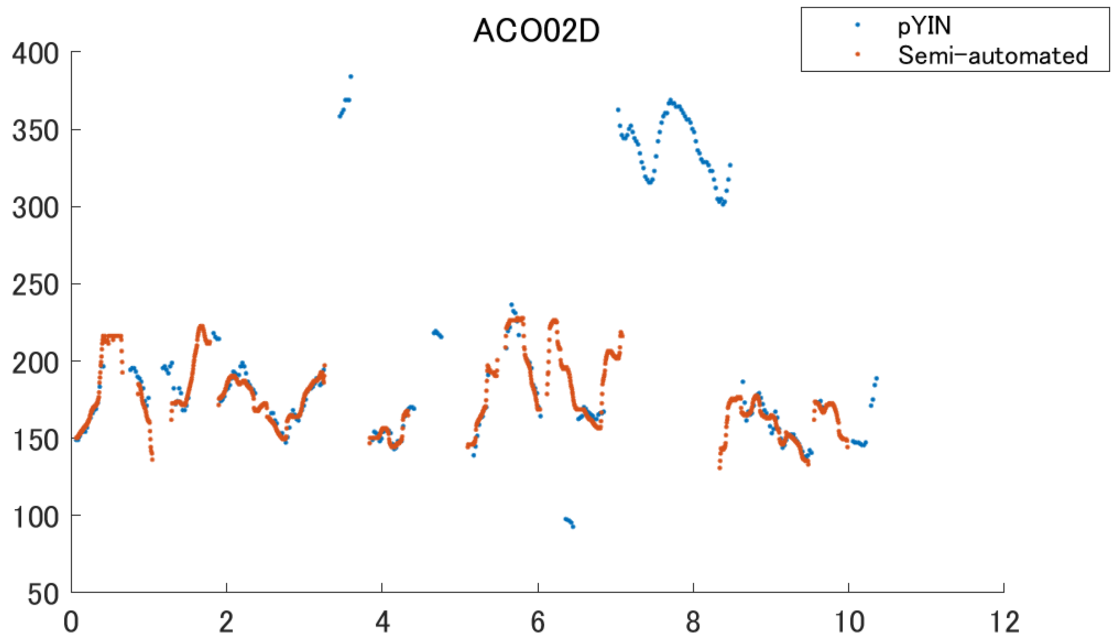



**Fig. S11. An example of fully-automated vs. semi-automated $f_0$ extraction underlying the analyses in Fig. S7 for one of the field recordings from Hilton et al.'s (*26*) dataset.** AC002D = adult-directed speech [D] from individual #02 from the Spanish-speaking Afro-Colombian [ACO] sample). While 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).

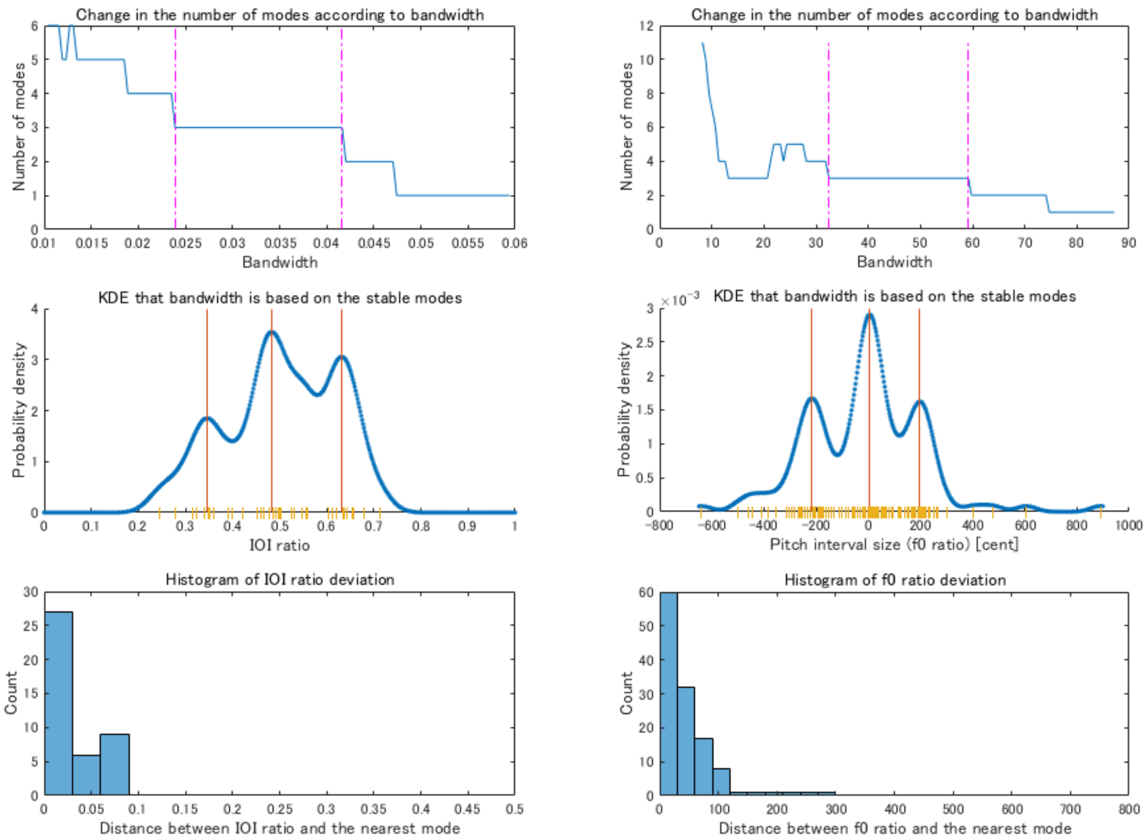



**Fig. S12**. **Illustration of the computation of IOI ratio deviation and $f_0$ ratio deviation.** KDE stands for kernel density estimation. The interval between the magenta lines is the range of the bandwidth parameter within which the Betti number (number of modes) is most stable which we interpret 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.

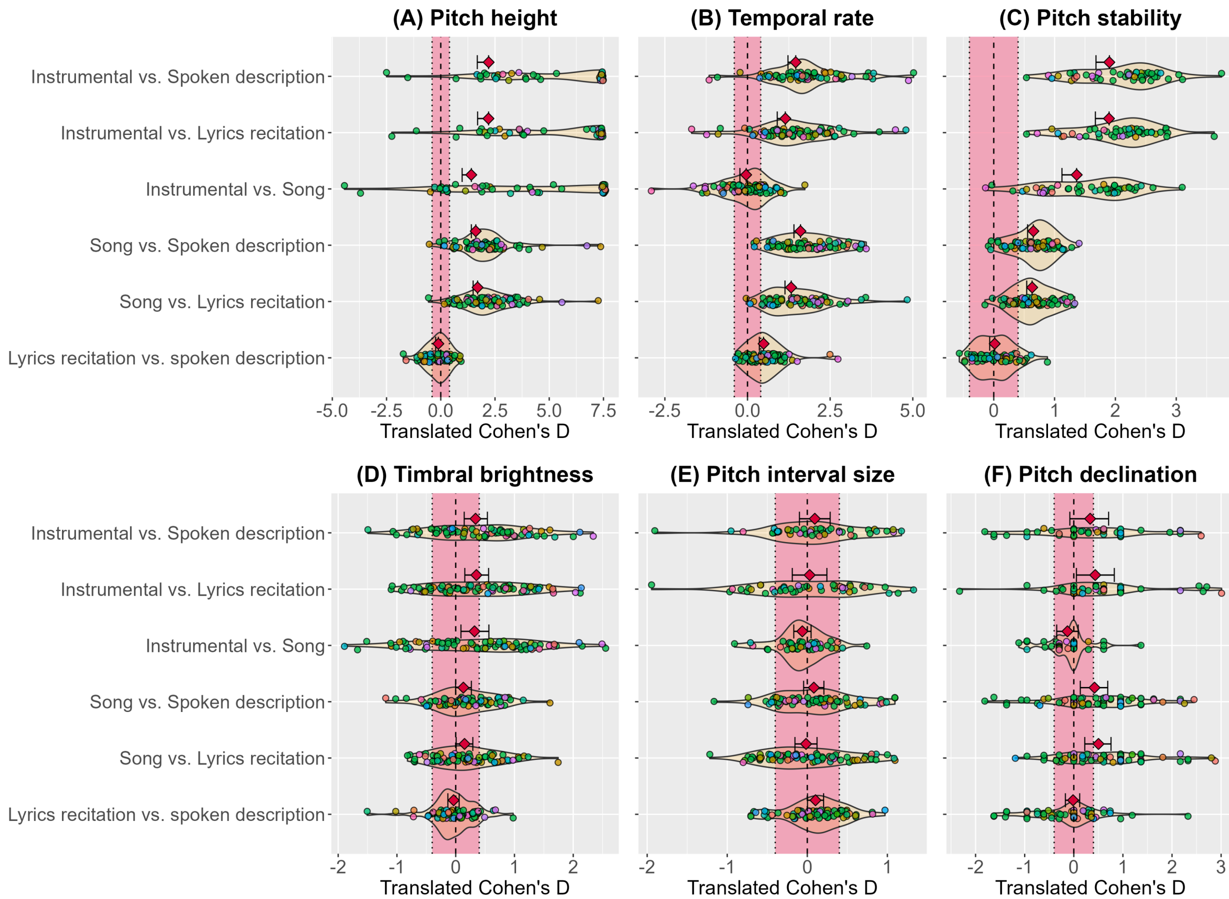



**Fig. S13**. **Effect sizes of each feature using the same data as in Fig. 4 but with exploratory comparisons with recitation and instrumental recording types.** Refer to Fig. 4 for the explanation of the figure description.

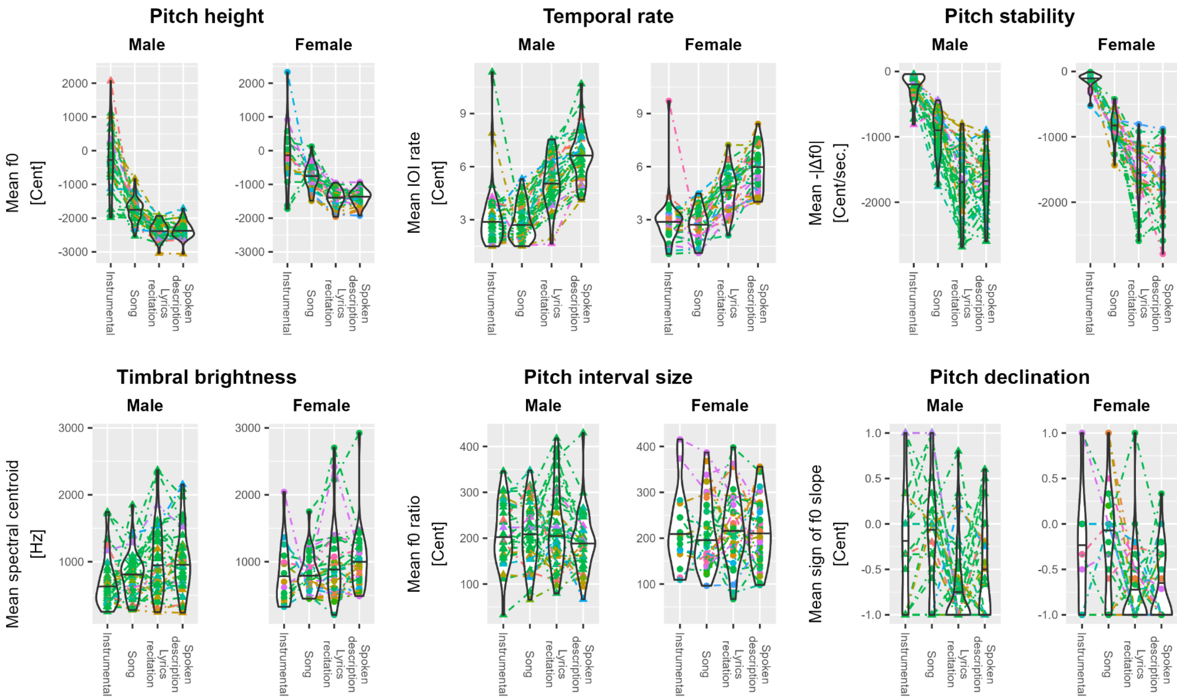



**Fig. S14. Alternative visualization of Fig. 5 showing mean values of each feature by biological sex and focusing on the features subject to the main confirmatory analysis.** Note that the colors of data points indicate language families, which are coded the same as in Fig. 5.

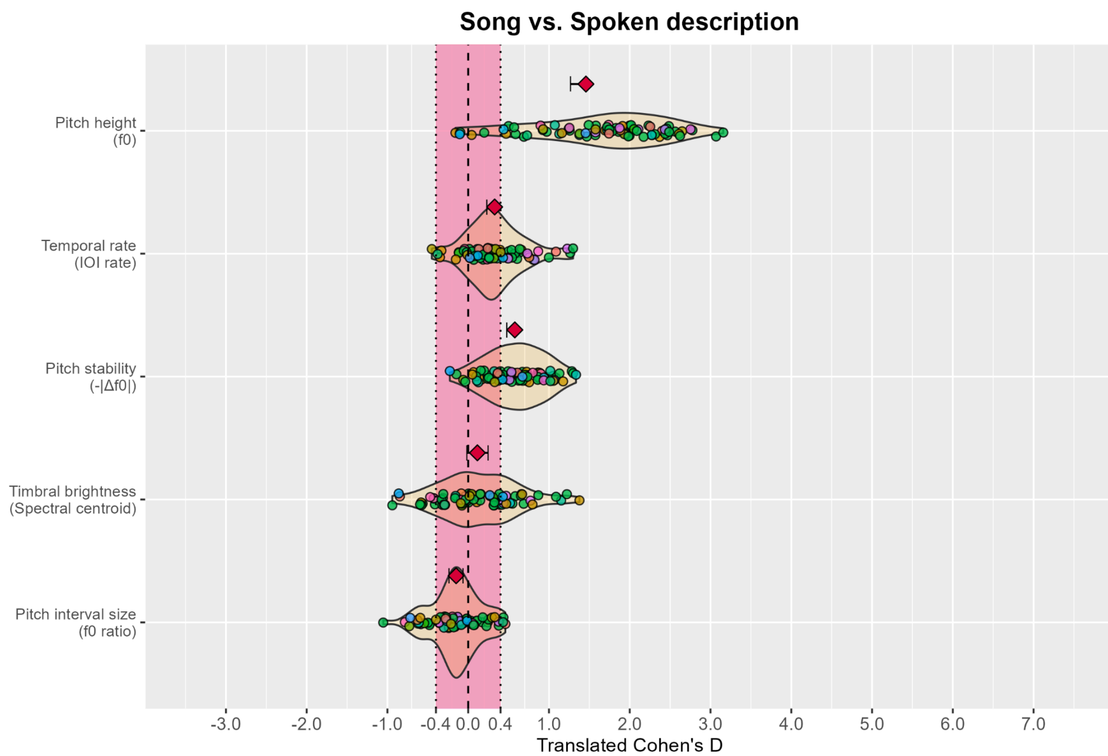



**Fig. S15. Re-running of the analysis on our full data with automated feature extraction.** pYIN (*60*) was used for f0 extraction and de (*59*) Praat script was used for onset timing extraction. Break annotation was not automated so pitch declination was not measured.


**Fig. S16. Color mapping of Fig. 6.** The colors of data points in Fig. 6 correspond to the language families as depicted in this figure.

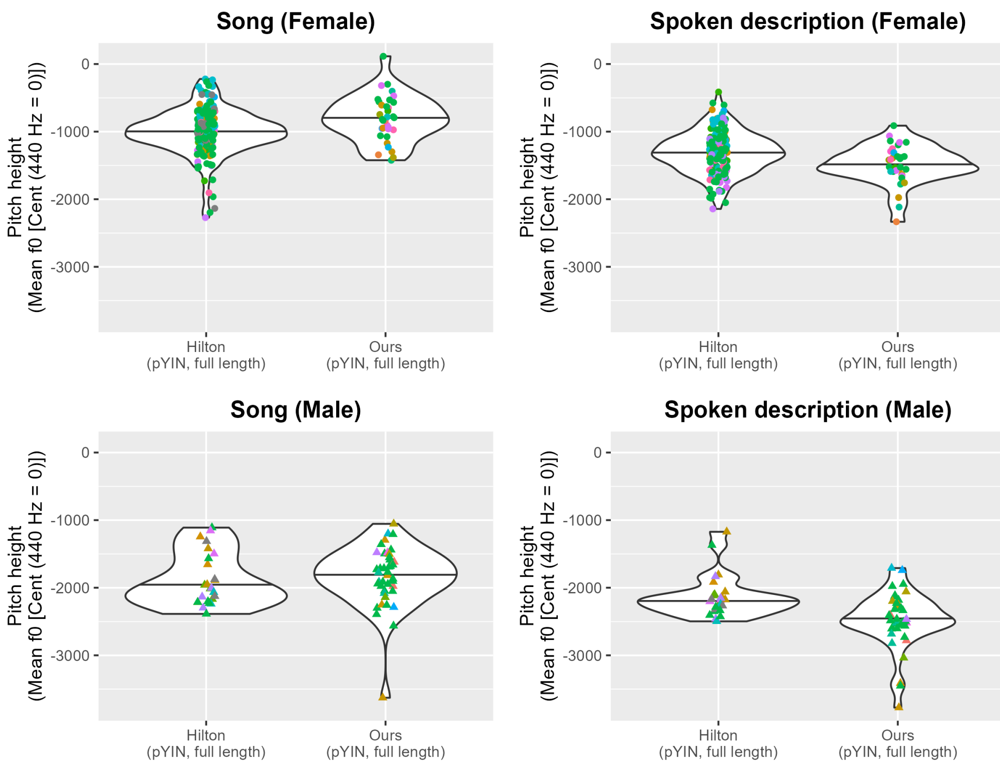



**Fig. S17. Supplementary information for Fig. 6.** Mean values of pitch height of each recording are displayed. $f_0$s were extracted by pYIN (*60*). The horizontal lines in the violin plots are medians.

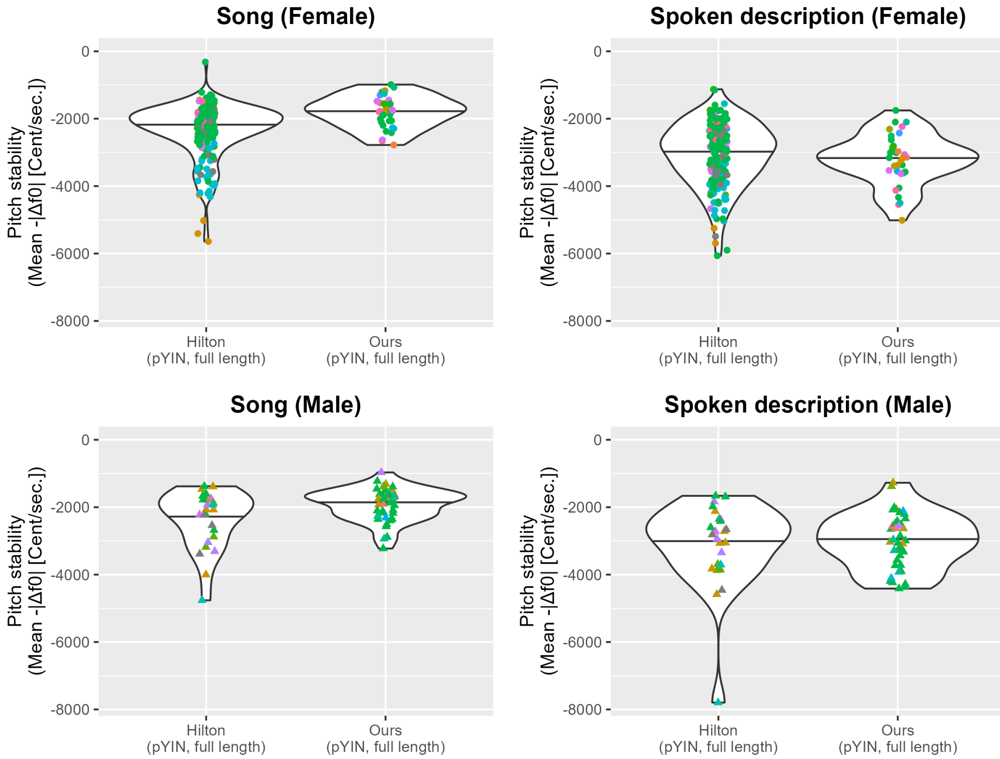



**Fig. S18. Supplementary information for Fig. 6.** Mean values of pitch stability of each recording are displayed. $f_0$s were extracted by pYIN (*60*). The horizontal lines in the violin plots are medians.

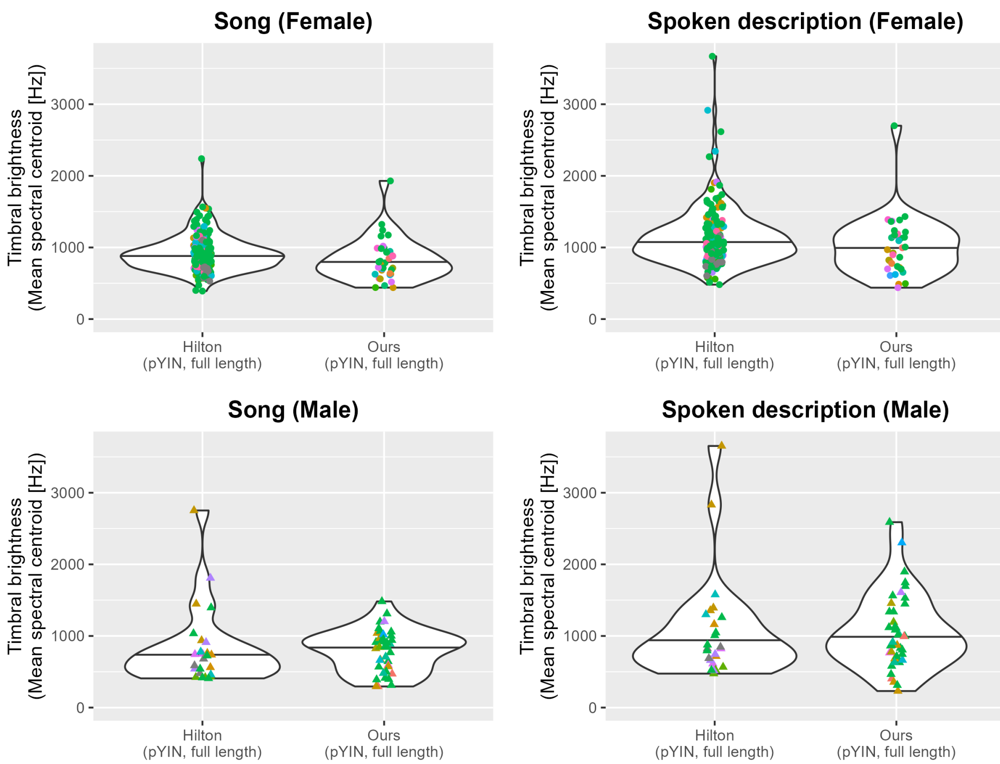



**Fig. S19. Supplementary information for Fig. 6.** Mean values of timbral brightness of each recording are displayed. $f_0$s were extracted by pYIN (*60*). The horizontal lines in the violin plots are medians.

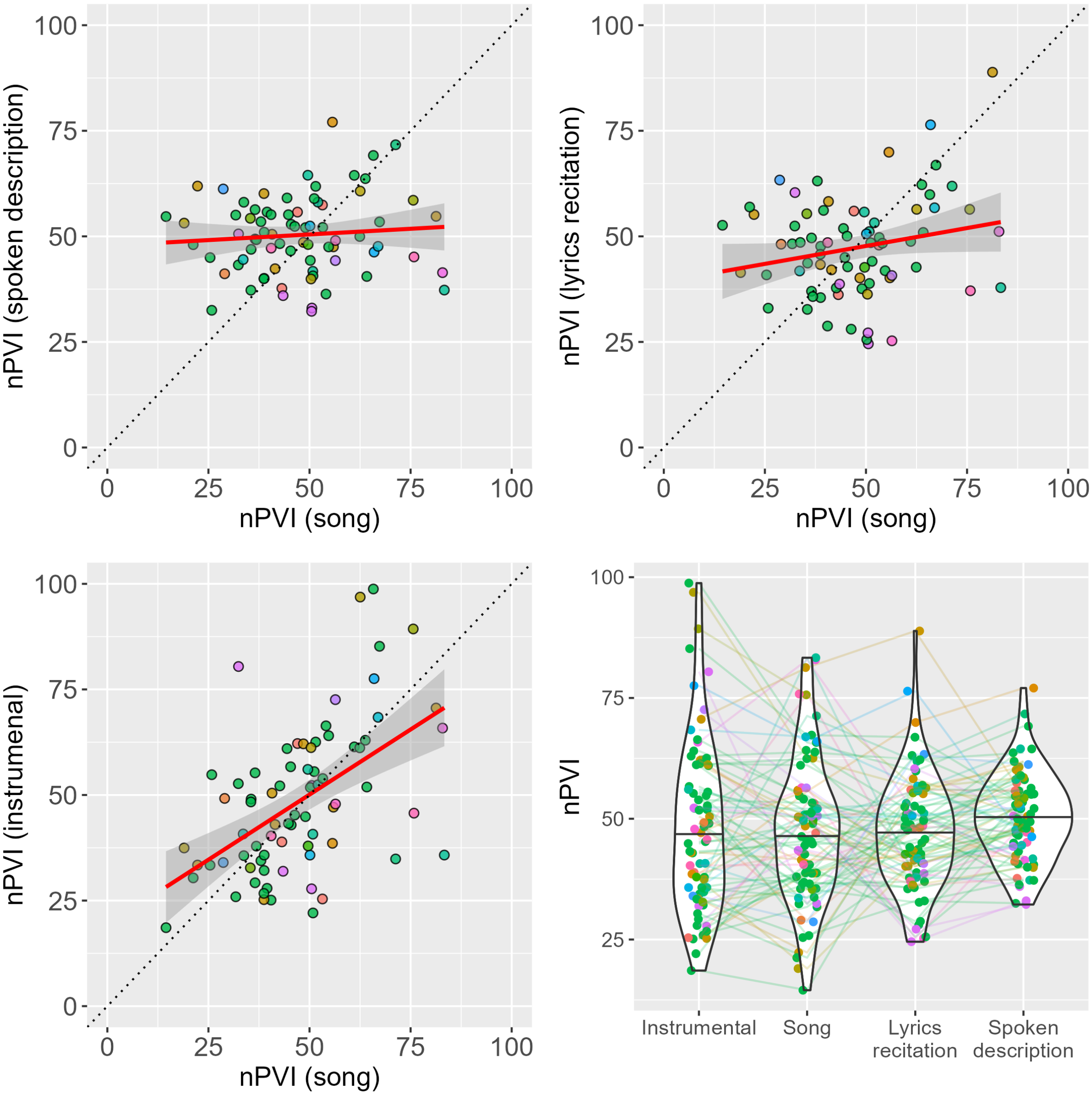



**Fig. S20. Mapping data by nPVIs of song and spoken description, song and lyrics recitation, and song and instrumental by each collaborator, and the density plot of nPVIs of each.** The red lines are linear fitting of nPVIs, and the dotted line is y = x which can be used to grasp if the nPVI of the particular form is larger than that of another and vice versa.

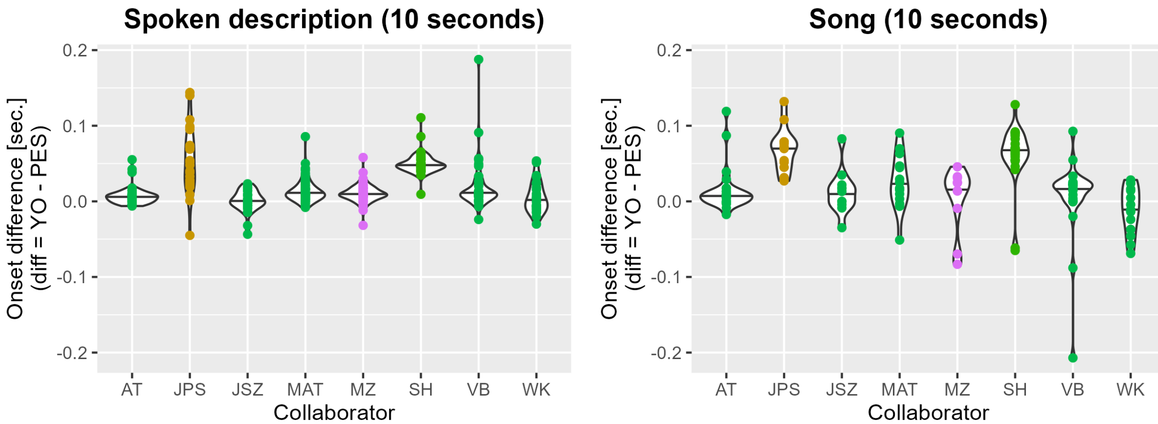



**Fig. S21. Difference between onset times annotated by Ozaki (YO) and onset times annotated by Savage (PES) per recording for the 8 codings re-annotated by Savage to assess inter-rater reliability.** The horizontal lines in the violin plots indicate the median. The abbreviation on the x-axis indicates the initials of randomly chosen collaborators for inter-rater reliability assessment using their recordings. Color is coded as the same in Fig. 4.

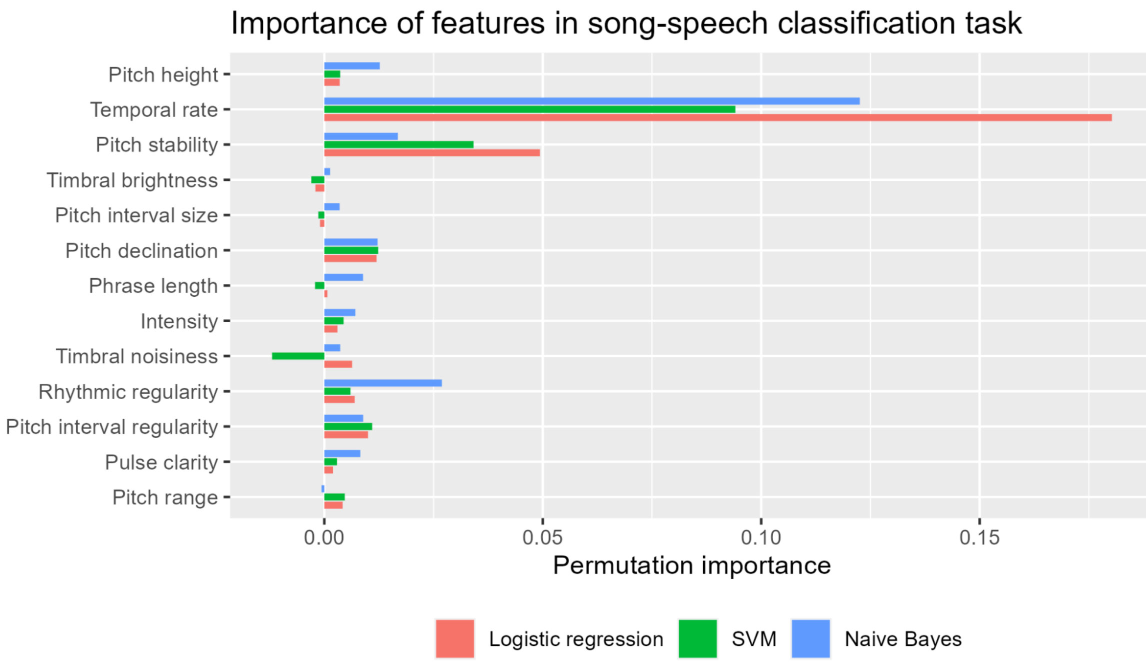



**Fig. S22. Permutation importance of the features in three binary classifiers.** A large permutation importance score indicates a strong influence on classification performance.

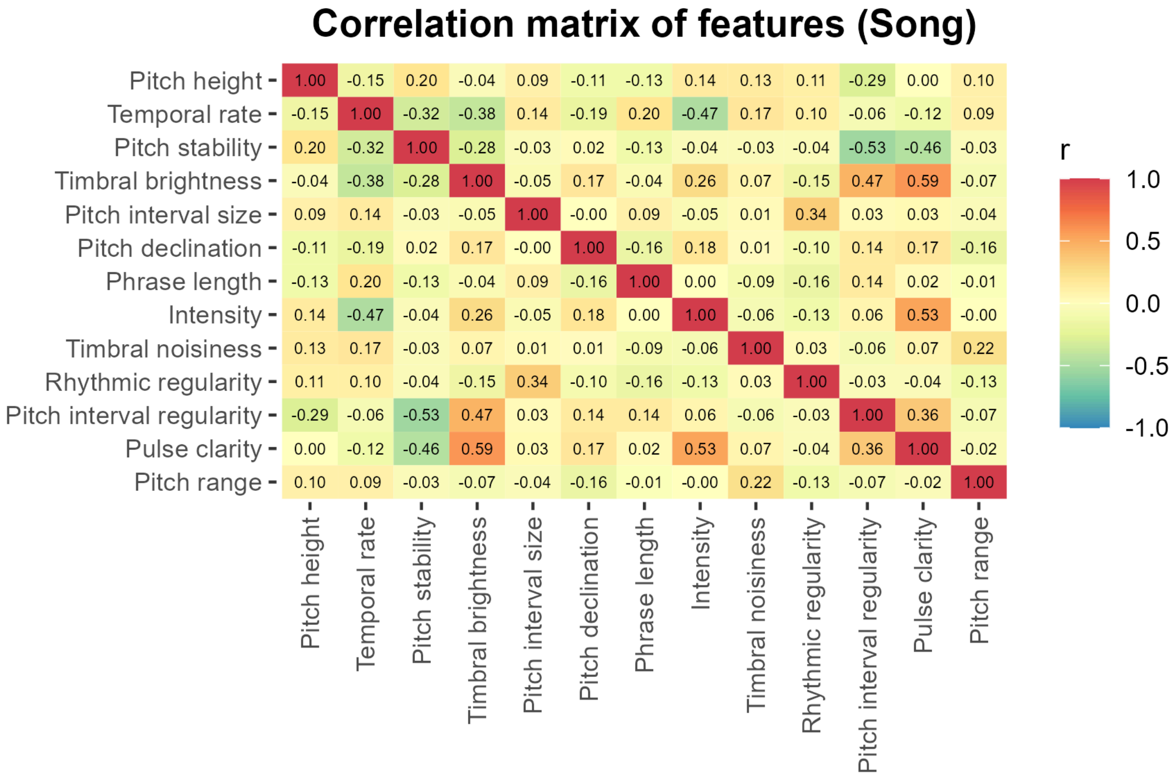



**Fig. S23. Correlation matrix of the features within song recordings.** The data are the mean values of the features, which are plotted in Fig. 5.

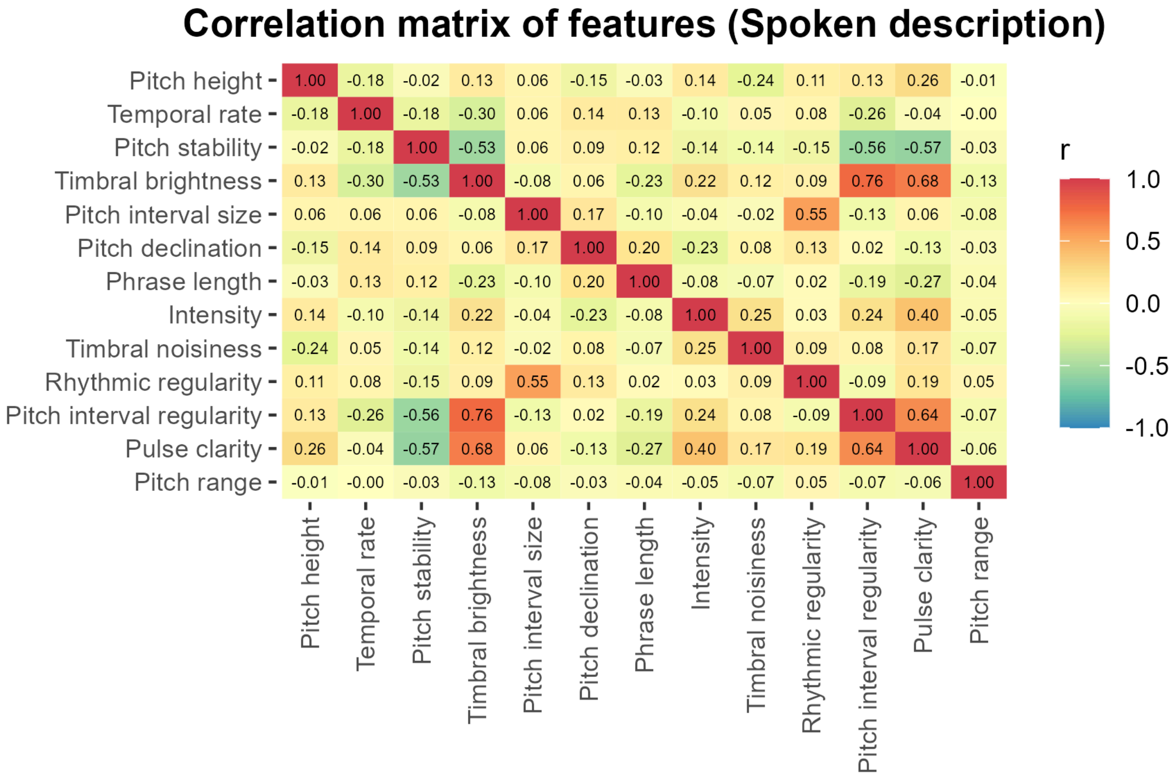



**Fig. S24. Correlation matrix of the features within spoken description recordings.** The data are the mean values of the features, which are plotted in Fig. 5.

SUPPLEMENTARY TABLES

| **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 | $-\|\Delta f_0\|$ | | 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 | 0.57 |

**Table S1**. **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.

| Hypothesis | Feature | Test | Combined ES | CI ($\alpha$ = 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 | ***< $1.0\text{x}10^{-8}$** |
| 2) Song is slower than speech | IOI rate | | 1.64 | 1.40, n/a | ***< $1.0\text{x}10^{-8}$** |
| 3) Song uses more stable pitches than speech | $-\lvert\Delta f_0\rvert$ | | 0.64 | 0.51, n/a | ***< $1.0\text{x}10^{-8}$** |
| 4) Song and speech use similar timbral brightness | Spectral centroid | Equivalence test for the combined effect size | 0.14 | -0.028, 0.31 | ***$3.3\text{x}10^{-4}$** |
| 5) Song and speech use similar sized pitch intervals | $f_0$ ratio | | 0.10 | -0.067, 0.27 | ***$3.5\text{x}10^{-5}$** |
| 6) Song and speech use similar pitch contours | Sign of $f_0$ slope | | 0.23 | -0.11, 0.60 | 0.12 |

**Table S2**. **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).** Refer to Table S1 for the explanation of the table description.

| **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 | $< 1.0\times10^{-8}$ |
| 2) Song is slower than speech | **-111.64** | -109.73 | 57.82 | 57.86 | $1.86\times10^{-3}$ |
| 3) Song uses more stable pitches than speech | **-153.53** | -151.53 | 78.76 | 78.76 | $< 1.0\times10^{-8}$ |
| 4) Song and speech use similar timbral brightness | **-86.32** | -84.90 | 45.16 | 45.45 | $2.07\times10^{-3}$ |
| 5) Song and speech use similar sized pitch intervals | **-95.90** | -93.90 | 49.95 | 49.95 | $< 1.0\times10^{-8}$ |
| 6) Song and speech use similar pitch contours | **-7.24** | -5.48 | 5.62 | 5.74 | $2.29\times10^{-3}$ |

**Table S3**. **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 adequate to model the data.

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

**Table S4. 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 (as in a between-participants design), but our data are collected in a within-participants 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 participants. The p-values were calculated by a Monte Carlo permutation procedure.

| | | 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 |

**Table S5. Average over performance metrics measured by randomly splitting recording sets into training and test sets 1024 times.** Each cell represents the classification accuracy of song and spoken description, and the precision, recall, and F1 score of each classifier.

| 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 |

**Table S6. 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 the differences were non-existent ("equivalence") or negligible (as small as our chosen SESOI [Smallest Effect Size Of Interest]).** Audio files to listen to the effect of manipulation are available at https://osf.io/8mcev.

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 xFormat 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.

COLLABORATION AGREEMENT FORM

**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.**

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 [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: ____________________________________________________________

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

LIST OF SONGS, INSTRUMENTS, AND LANGUAGES

*NB: Heritage speakers are indicated using italics, all others speak the language as their 1st language.*

| # | Name | Song title (Romanization) | Language *(italics = heritage speaker)* | 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 Bosch der Nederlanden | Sleep Now Rest Now | English [Michigan] | Cello |
| 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<br>(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 |

## REFERENCES AND NOTES


1. C. Chambers, "Strong evidence for cross-cultural regularities in music and speech," *Peer Community Registered Report* **1**, (100469) (2023); https://rr.peercommunityin.org/articles/rec?id=469.

2. C. Chambers, "Exploring cross-cultural variation in speech and song," *Peer Community Registered Report* (2023); https://rr.peercommunityin.org/articles/rec?id=316.

3. N. Evans, S. C. Levinson, The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).

4. P. E. Savage, Universals, in *The SAGE International Encyclopedia of Music and Culture* (SAGE Publications Inc., 2019), pp. 2283–2285; https://sk.sagepub.com/reference/the-sage internationalencyclopedia-of-music-and-culture/i21528.xml.

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

6. N. Jacoby, E. H. Margulis, M. Clayton, E. Hannon, H. Honing, J. Iversen, T. R. Klein, S. A. Mehr, L. Pearson, I. Peretz, M. Perlman, R. Polak, A. Ravignani, P. E. Savage, G. Steingo, C. J. Stevens, L. Trainor, S. Trehub, M. Veal, M. Wald-Fuhrmann, Cross-cultural work in music cognition: Challenges, insights, and recommendations. *Music Percept.* **37**, 185–195 (2020).

7. Y. Ozaki, M. de Heer Kloots, A. Ravignani, P. E. Savage, Cultural evolution of music and language. PsyArXiv 10.31234/osf.io/s7apx [Preprint] (2024). https://doi.org/10.31234/osf.io/s7apx.

8. C. Darwin, *The Descent of Man* (Watts & Co., 1871).

9. A. D. Patel, *Music, Language, and the Brain* (Oxford Univ. Press, 2008); https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195123753.001.0001/acprof-9780195123753.

10. J. V. Valentova, P. Tureček, M. A. C. Varella, P. Šebesta, F. D. C. Mendes, K. J. Pereira, L. Kubicová, P. Stolařová, J. Havlíček, Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Front. Psychol.* **10**, 2029 (2019).

11. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. Mithen, W. T. Fitch, Music as a coevolved system for social bonding. Behav. Brain Sci. 44(e59), 1–22 (2021).

12. S. A. Mehr, M. M. Krasnow, G. A. Bryant, E. H. Hagen, Origins of music in credible signaling. *Behav. Brain Sci.* **44**, e60 (2021).

13. F. Haiduk, W. T. Fitch, Understanding design features of music and language: The choric/dialogic distinction. *Front. Psychol.* **13**, 786899 (2022).

14. I. Peretz, Music, language and modularity framed in action. *Psychol. Belg.* **49**, 157–175 (2009).

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

16. C. Rogalsky, F. Rong, K. Saberi, G. Hickok, Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *J. Neurosci.* **31**, 3843–3852 (2011).

17. T. H. Morrill, J. D. McAuley, L. C. Dilley, D. Z. Hambrick, Individual differences in the perception of melodic contours and pitch-accent timing in speech: Support for domain-generality of pitch processing. *J. Exp. Psychol. Gen.* **144**, 730–736 (2015).

18. K. B. Doelling, M. F. Assaneo, D. Bevilacqua, B. Pesaran, D. Poeppel, An oscillator model better predicts cortical entrainment to music. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10113–10121 (2019).

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

20. A. D. Patel, J. R. Iversen, J. C. Rosenberg, Comparing the rhythm and melody of speech and music: The case of British English and French. *J. Acoust. Soc. Am.* **119**, 3034–3047 (2006).

21. N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, D. Poeppel, Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).

22. D. E. Brown, *Human Universals* (McGraw-Hill, 1991).

23. B. Bickel, Absolute and statistical universals, in *The Cambridge Encyclopedia of the Language Sciences*, P. C. Hogan, Ed. (Cambridge Univ. Press, 2011), pp. 77–79.

24. S. Brown, J. Jordania, Universals in the world's musics. *Psychol. Music* **41**, 229–248 (2013).

25. P. E. Savage, S. Brown, E. Sakai, T. E. Currie, Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8987–8992 (2015).

26. C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, A. Galbarczyk, G. Jasienska, C. T. Ross, M. B. Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, A. Schachner, T. A. Vardy, Q. D. Atkinson, A. Salenius, J. Andelin, J. Antfolk, P. Madhivanan, A. Siddaiah, C. D. Placek, G. D. Salali, S. Keestra, M. Singh, S. A. Collins, J. Q. Patton, C. Scaff, J. Stieglitz, S. C. Cutipa, C. Moya, R. R.

Sagar, M. Anyawire, A. Mabulla, B. M. Wood, M. M. Krasnow, S. A. Mehr, Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **6**, 1545–1556 (2022).

27. Y. Ozaki, S. Sato, J. Mcbride, P. Q. Pfordresher, A. T. Tierney, J. Six, S. Fujii, P. E. Savage, Automatic acoustic analyses quantify pitch discreteness within and between human music, speech, and birdsong, in *Proceedings of the 10th International Workshop on Folk Music Analysis* (2022).

28. A. T. Tierney, F. A. Russo, A. D. Patel, The motor origins of human and avian song structure. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15510–15515 (2011).

29. G. List, On the non-universality of musical perspectives. *Ethnomusicology* **15**, 399–402 (1971).

30. A. Lomax, V. Grauer, The Cantometric coding book, in *Folk Song Style and Culture*, A. Lomax, Ed. (American Association for the Advancement of Science, 1968), pp. 34–74.

31. J. Blacking, *How Musical Is Man?* (University of Washington Press, 1973).

32. J. H. L. Hansen, M. Bokshi, S. Khorram, Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing. *J. Acoust. Soc. Am.* **148**, 829–844 (2020).

33. J. Merrill, P. Larrouy-Maestri, Vocal features of song and speech: Insights from Schoenberg’s Pierrot lunaire. *Front. Psychol.* **8**, 1108 (2017).

34. B. Sharma, X. Gao, K. Vijayan, X. Tian, H. Li, NHSS: A speech and singing parallel database. *Speech Commun.* **133**, 9–22 (2021).

35. C. M. Vanden Bosch der Nederlanden, X. Qi, S. Sequeira, P. Seth, J. A. Grahn, M. F. Joanisse, E. E. Hannon, Developmental changes in the categorization of speech and song. *Dev. Sci.* **26**, e13346 (2023).

36. D. E. Blasi, J. Henrich, E. Adamou, D. Kemmerer, A. Majid, Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **26.** 1153–1170 (2022).

37. Y. Ozaki, J. Kuroyanagi, G. Chiba, J. McBride, P. Proutskova, A. Tierney, E. Benetos, F. Liu, P. E. Savage, Similarities and differences in a cross-linguistic sample of song and speech recordings, in *Proceedings of the 2022 Joint Conference on Language Evolution* (Joint Conference on Language Evolution (JCoLE) Max Planck Institute for Psycholinguistics, 2022), pp. 569–572.

38. C. Durojaye, L. Fink, T. Roeske, M. Wald-Fuhrmann, P. Larrouy-Maestri, Perception of Nigerian dùndún talking drum performances as speech-like vs. music-like: The role of familiarity and acoustic cues. *Front. Psychol.* **12**, 652673 (2021).

39. T. C. Roeske, O. Tchernichovski, D. Poeppel, N. Jacoby, Categorical rhythms are shared between songbirds and humans. *Curr. Biol.* **30**, 3544–3555.e6 (2020).

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

41. S. Brown, The musilanguage model of music evolution, in *The Origins of Music*, S. Brown, B. Merker, C. Wallin, Eds. (The MIT Press, 2000), pp. 271–300; https://direct.mit.edu/books/book/2109/chapter/56564/The-Musilanguage-Model-of-Music-Evolution.

42. J. D. Leongómez, J. Havlíček, S. C. Roberts, Musicality in human vocal communication: An evolutionary perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200391 (2022).

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

44. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. Mithen, W. T. Fitch, Authors' response: Toward inclusive theories of the evolution of musicality. Behav. Brain Sci. 44 (e121), 132–140 (2021).

45. A. D. Patel, Music as a transformative technology of the mind: An update, in *The Origins of Musicality*, H. Honing, Ed. (The MIT Press, 2018); pp. 113–126; https://direct.mit.edu/books/book/4115/ chapter/170183/Music-as-a-Transformative-Technology-of-the-Mind.

46. M. Hoeschele, W. T. Fitch, Cultural evolution: Conserved patterns of melodic evolution across musical cultures. *Curr. Biol.* **32**, R265–R267 (2022).

47. H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, glottolog/glottolog: Glottolog database 4.7 (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2022); https://doi.org/10.5281/zenodo.7398962.

48. M. S. Dryer, M. Haspelmath, *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, 2013); http://wals.info.

49. S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, M. Müller, Erkomaishvili dataset: A curated corpus of traditional georgian vocal music for computational musicology. *Trans. Int. Soc. Music Inf. Retr.* **3**, 31–41 (2020).

50. M. Müller, S. Rosenzweig, J. Driedger, F. Scherbaum, *Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research* (Audio Engineering Society, 2017); www.aes.org/elib/ browse.cfm?elib=18777.

51. N. Bannan, R. I. M. Dunbar, J. S. Bamford, The evolution of gender dimorphism in the human voice: The role of octave equivalence. PsyArXiv 10.31234/osf.io/f4j6b [Preprint] (2024). https://doi.org/10.31234/osf.io/f4j6b.

52. S. Chen, C. Han, S. Wang, X. Liu, B. Wang, R. Wei, X. Lei, Hearing the physical condition: The relationship between sexually dimorphic vocal traits and underlying physiology. *Front. Psychol.* **13**, 983688 (2022).

53. D. R. Feinberg, B. C. Jones, M. M. Armstrong, Sensory exploitation, sexual dimorphism, and human voice pitch. *Trends Ecol. Evol.* **33**, 901–903 (2018).

54. D. A. Puts, S. J. C. Gaulin, K. Verdolini, Dominance and the evolution of sexual dimorphism in human voice pitch. *Evol. Hum. Behav.* **27**, 283–296 (2006).

55. D. A. Puts, A. K. Hill, D. H. Bailey, R. S. Walker, D. Rendall, J. R. Wheatley, L. L. M. Welling, K. Dawood, R. Cárdenas, R. P. Burriss, N. G. Jablonski, M. D. Shriver, D. Weiss, A. R. Lameira, C. L. Apicella, M. J. Owren, C. Barelli, M. E. Glenn, G. Ramos-Fernandez, Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proc. R. Soc. B Biol. Sci.* **283**, 20152830 (2016).

56. A. D. Patel, J. R. Daniele, An empirical comparison of rhythm in language and music. *Cognition* **87**, B35–B45 (2003).

57. L. E. Ling, E. Grabe, F. Nolan, Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech* **43**, 377–401 (2000).

58. E. Grabe, E. L. Low, Durational variability in speech and the Rhythm Class Hypothesis, in *Laboratory Phonology 7*, C. Gussenhoven, N. Warner, Eds. (De Gruyter Mouton, 2002), pp. 515–546; www.degruyter.com/document/doi/10.1515/9783110197105.2.515/html?lang=en.

59. N. H. de Jong, T. Wempe, Praat script to detect syllable nuclei and measure speech rate automatically. *Behav. Res. Methods* **41**, 385–390 (2009).

60. M. Mauch, S. Dixon, PYIN: A fundamental frequency estimator using probabilistic threshold distributions, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2014), pp. 659–663.

61. F. Weber, G. Knapp, K. Ickstadt, G. Kundt, Ä. Glass, Zero-cell corrections in random-effects metaanalyses. *Res. Synth. Methods* **11**, 913–919 (2020).

62. C. R. Adams, Melodic contour typology. *Ethnomusicology* **20**, 179–215 (1976).

63. D. Huron, The melodic arch in western folksongs. *Comput. Musicol.* **10**, 3–23 (1996).

64. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).

65. "Permutation importance," ELI5; https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html.

66. J. P. B. Pereira, E. S. G. Stroes, A. H. Zwinderman, E. Levin, Covered information disentanglement: Model transparency via unbiased permutation importance. *Proc. AAAI Conf. Artif. Intell.* **36**, 7984–7992 (2022).

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

68. P. E. Savage, A. T. Tierney, A. D. Patel, Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Percept.* **34**, 327–334 (2017).

69. Nature addresses helicopter research and ethics dumping. *Nature* **606**, 7 (2022).

70. M. Urassa, D. W. Lawson, J. Wamoyi, E. Gurmu, M. A. Gibson, P. Madhivanan, C. Placek, Crosscultural research must prioritize equitable collaboration. *Nat. Hum. Behav.* **5**, 668–671 (2021).

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

72. J. Troy, L. Barwick, Claiming the 'Song of the Women of the Menero Tribe'. *Musicol. Aust.* **42**, 85– 107 (2020).

73. P. E. Savage, H. Matsumae, H. Oota, M. Stoneking, T. E. Currie, A. Tajima, M. Gillan, S. Brown, How 'circumpolar' is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicol. Forum* **24**, 443–467 (2015).

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

75. D. Temperley, Music and language. *Annu. Rev. Linguist.* **8**, 153–170 (2022).

76. F. Ramus, Acoustic correlates of linguistic rhythm: Perspectives. *Proc. Speech Prosody* **2002**, 115–120 (2002).

77. M. Berg, M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, T. Berger, The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors. *J. Voice* **31**, 257.e13–257.e24 (2017).

78. K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. M. O'Connor, S. Röder, P. W. Andrews, B. Fink, L. M. DeBruine, B. C. Jones, D. R. Feinberg, Vocal indicators of body size in men and women: A metaanalysis. *Anim. Behav.* **95**, 89–99 (2014).

79. B. Barsties, Einfluss verschiedener Methoden zur Bestimmung der mittleren Sprechstimmlage. *HNO* **61**, 609–616 (2013).

80. F. Pellegrino, C. Coupé, E. Marsico, Across-Language perspective on speech information rate. *Language* **87**, 539–558 (2011).

81. D. Poeppel, M. F. Assaneo, Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* **21**, 322– 334 (2020).

82. A. Anikin, V. Canessa-Pollard, K. Pisanski, M. Massenet, D. Reby, Beyond speech: Exploring diversity in the human voice. *iScience* **26**, 108204 (2023).

83. W. T. Fitch, The biology and evolution of music: A comparative perspective. *Cognition* **100**, 173–215 (2006).

84. W. Ma, A. Fiveash, W. F. Thompson, Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage. *Semiotica* **2019**, 1–23 (2019).

85. A. Lomax, *Folk Song Style and Culture* (American Association for the Advancement, 1968).

86. F. Alipour, R. C. Scherer, On pressure-frequency relations in the excised larynx. *J. Acoust. Soc. Am.* **122**, 2296–2305 (2007).

87. Y. Suzuki, H. Takeshima, Equal-loudness-level contours for pure tones. *J. Acoust. Soc. Am.* **116**, 918– 933 (2004).

88. D. J. Levitin, Knowledge songs as an evolutionary adaptation to facilitate information transmission through music. *Behav. Brain Sci.* **44**, e105 (2021).

89. M. A. C. Varella, Nocturnal selective pressures on the evolution of human musicality as a missing piece of the adaptationist puzzle. *Front. Psychol.* **14**, 1215481 (2023).

90. S. E. Trehub, Cross-cultural convergence of musical features. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8809– 8810 (2015).

91. M. Singh, S. A. Mehr, Universality, domain-specificity and development of psychological responses to music. *Nat. Rev. Psychol.* **2**, 333–346 (2023).

92. D. Lakens, Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).

93. M. Brysbaert, How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* **2**, 16 (2019).

94. E. Brunner, A. C. Bathke, F. Konietschke, *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS* (Springer, 2018); https://ci.nii.ac.jp/ncid/BB28708839.

95. A. Vargha, H. D. Delaney, The Kruskal-Wallis test and stochastic homogeneity. *J. Educ. Behav. Stat.* **23**, 170–192 (1998).

96. J. Ruscio, A probability-based measure of effect size: Robustness to base rates and other factors. *Psychol. Methods* **13**, 19–30 (2008).

97. S. E. Brockwell, I. R. Gordon, A comparison of statistical methods for meta-analysis. *Stat. Med.* **20**, 825–840 (2001).

98. S. Liu, L. Tian, S. Lee, M.-g. Xie, Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostat. Epidemiol.* **2**, 1–22 (2018).

99. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

100. Y. Wang, L. Tian, An efficient numerical algorithm for exact inference in meta analysis. *J. Stat. Comput. Simul.* **88**, 646–656 (2018).

101. R. DerSimonian, N. Laird, Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).

102. J. P. Romano, Optimal testing of equivalence hypotheses. *Ann. Stat.* **33**, 1036–1047 (2005).
103. S. B. Tan, M. Ostashewski, Eds., *DIALOGUES: Towards Decolonizing Music and Dance Studies* (International Council for Traditional Music, 2022); https://ictmdialogues.org/.
104. P. E. Savage, N. Jacoby, E. H. Margulis, H. Daikoku, M. Anglada-Tort, S. E.-S. Castelo-Branco, F. E. Nweke, S. Fujii, S. Hegde, H. Chuan-Peng, J. Jabbour, C. Lew-Williams, D. Mangalagiu, R. McNamara, D. Müllensiefen, P. Opondo, A. D. Patel, H. Schippers, Building sustainable global collaborative networks: Recommendations from music studies and the social sciences, in *The Science- Music Borderlands: Reckoning with the Past, Imagining the Future*, E. H. Margulis, L. Loughridge, P. Loui, Eds. (The MIT Press, 2023), pp. 347–365; https://direct.mit.edu/books/oa-editedvolume/ 5578/chapter/4162120/Building-Sustainable-Global-Collaborative-Networks.
105. N. Novitski, M. Tervaniemi, M. Huotilainen, R. Näätänen, Frequency discrimination at different frequency levels as indexed by electrophysiological and behavioral measures. *Cogn. Brain Res.* **20**, 26– 36 (2004).
106. A. Anikin, The link between auditory salience and emotion intensity. *Cogn. Emot.* **34**, 1246–1259 (2020).
107. C. Cox, C. Bergmann, E. Fowler, T. Keren-Portnoy, A. Roepstorff, G. Bryant, R. Fusaroli, A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nat. Hum. Behav.* **7**, 114–133 (2023).
108. T. Verhoef, A. Ravignani, Melodic universals emerge or are sustained through cultural Evolution. *Front. Psychol.* **12**, 668300 (2021).
109. P. Q. Pfordresher, S. Brown, K. M. Meier, M. Belyk, M. Liotti, Imprecise singing is widespread. *J. Acoust. Soc. Am.* **128**, 2182–2190 (2010).

110. U. Natke, T. M. Donath, K. Th. Kalveram, Control of voice fundamental frequency in speaking versus singing. *J. Acoust. Soc. Am.* **113**, 1587–1593 (2003).

111. B. Raposo de Medeiros, J. P. Cabral, A. R. Meireles, A. A. Baceti, A comparative study of fundamental frequency stability between speech and singing. *Speech Commun.* **128**, 15–23 (2021).

112. E. L. Stegemöller, E. Skoe, T. Nicol, C. M. Warrier, N. Kraus, Music training and vocal production of speech and song. *Music Percept.* **25**, 419–428 (2008).

113. B. Thompson, Discrimination between singing and speech in real-world audio, in *2014 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2014), pp. 407–412.

114. G. A. Bryant, The evolution of human vocal emotion. *Emot. Rev.* **13**, 25–33 (2021).

115. S. E. Trehub, A. M. Unyk, S. B. Kamenetsky, D. S. Hill, L. J. Trainor, J. L. Henderson, M. Saraza, Mothers' and fathers' singing to infants. *Dev. Psychol.* **33**, 500–507 (1997).

116. A. Nikolsky, E. Alekseyev, I. Alekseev, V. Dyakonova, The overlooked tradition of "personal music" and its place in the evolution of music. *Front. Psychol.* **10**, 3051 (2020).

117. A. D. Patel, C. von Rueden, Where they sing solo: Accounting for cross-cultural variation in collective music-making in theories of music evolution. *Behav. Brain Sci.* **44**, e85 (2021).

118. D. Ross, J. Choi, D. Purves, Musical intervals in speech. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9852–9857 (2007).

119. D. A. Schwartz, C. Q. Howe, D. Purves, The statistical structure of human speech sounds predicts musical universals. *J. Neurosci.* **23**, 7160–7168 (2003).

120. S. Han, J. Sundararajan, D. L. Bowling, J. Lake, D. Purves, Co-variation of tonality in the music and speech of different cultures. *PLOS ONE* **6**, e20160 (2011).

121. J. P. Robledo, E. Hurtado, F. Prado, D. Román, C. Cornejo, Music intervals in speech: Psychological disposition modulates ratio precision among interlocutors' nonlocal f0 production in real-time dyadic conversation. *Psychol. Music* **44**, 1404–1418 (2016).

122. R. E. Stone Jr., T. F. Cleveland, J. Sundberg, Formant frequencies in country singers' speech and singing. *J. Voice* **13**, 161–167 (1999).

123. B. Lindblom, J. Sundberg, The human voice in speech and singing, in *Springer Handbook of Acoustics*, T. D. Rossing, Ed. (Springer, 2007), pp. 669–712; https://doi.org/10.1007/978-0-387-30425-0_16.

124. J. J. Barnes, P. Davis, J. Oates, J. Chapman, The relationship between professional operatic soprano voice and high range spectral energy. *J. Acoust. Soc. Am.* **116**, 530–538 (2004).

125. J. Sundberg, Level and center frequency of the singer's formant. *J. Voice* **15**, 176–186 (2001).

126. D. R. Ladd, Declination: A review and some hypotheses. *Phonol. Yearb.* **1**, 53–74 (1984).

127. J. Slifka, Respiratory system pressures at the start of an utterance, in *Dynamics of Speech Production and Perception*, P. Divenyi, S. Greenbarg, G. Meyer, Eds. (IOS Press, 2006), pp. 45–57; https://ebooks.iospress.nl/volumearticle/392.

128. H. Bârzan, V. V. Moca, A.-M. Ichim, R. C. Muresan, Fractional superlets, in *2020 28th European Signal Processing Conference (EUSIPCO)* (IEEE, 2021), pp. 2220–2224.

129. V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, R. C. Mureșan, Time-frequency super-resolution with superlets. *Nat. Commun.* **12**, 337 (2021).

130. I. Djurović, L. Stanković, An algorithm for the Wigner distribution based instantaneous frequency estimation in a high noise environment. *Signal Process.* **84**, 631–643 (2004).

131. A. Danielsen, K. Nymoen, E. Anderson, G. S. Câmara, M. T. Langerød, M. R. Thompson, J. London, Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds. *J. Exp. Psychol. Hum. Percept. Perform.* **45**, 402–418 (2019).

132. S. K. Scott, The point of P-centres. *Psychol. Res.* **61**, 4–11 (1998).

133. J. Vos, R. Rasch, The perceptual onset of musical tones. *Percept. Psychophys.* **29**, 323–335 (1981).

134. P. Howell, Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Percept. Psychophys.* **43**, 90–93 (1988).

135. J. Morton, S. Marcus, C. Frankish, Perceptual centers (P-centers). *Psychol. Rev.* **83**, 405–408 (1976).

136. B. Pompino-Marschall, On the psychoacoustic nature of the P-center phenomenon. *J. Phon.* **17**, 175– 192 (1989).

137. X. Shao, C. Ma, A general approach to derivative calculation using wavelet transform. *Chemom. Intel. Lab. Syst.* **69**, 157–165 (2003).

138. Z.-H. Tan, A. k. Sarkar, N. Dehak, rVAD: An unsupervised segment-based robust voice activity detection method. *Comput. Speech Lang.* **59**, 1–21 (2020).

139. J. E. Chacón, The modal age of statistics. *Int. Stat. Rev.* **88**, 122–141 (2020).

140. P. Chaudhuri, J. S. Marron, SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* **94**, 807– 823 (1999).

141. F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, L. Wasserman, Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18**, 1–40 (2018).

142. Y.-C. Chen, C. R. Genovese, L. Wasserman, A comprehensive approach to mode clustering. *Electron. J. Stat.* **10**, 210–241 (2016).

143. D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 603–619 (2002).

144. B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, Confidence sets for persistence diagrams. *Ann. Stat.* **42**, 2301–2339 (2014).

145. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **78**, 99–126 (2016).

146. M. Sommerfeld, G. Heo, P. Kim, S. T. Rush, J. S. Marron, Bump hunting by topological data analysis. *Stat* **6**, 462–471 (2017).

147. R. Zhang, R. Ghanem, Normal-bundle bootstrap. *SIAM J. Math. Data Sci.* **3**, 573–592 (2021).

148. F. T. Pokorny, C. H. Ek, H. Kjellström, D. Kragic, "Topological constraints and kernel-based density estimation," *Advances in Neural Information Processing Systems 25, Workshop on Algebraic Topology and Machine Learning*, Nevada, USA, 8 December 2012.

149. G. Carlsson, Topology and data. *Bull. Am. Math. Soc.* **46**, 255–308 (2009).

150. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, 1986).

151. P. Hall, S. J. Sheather, M. C. Jones, J. S. Marron, On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–269 (1991).

152. Y.-C. Chen, C. R. Genovese, S. Ho, L. Wasserman, Optimal ridge detection using coverage risk, in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2015), vol.

28; pp. 1–9; https://papers.nips.cc/paper/2015/hash/0aa1883c6411f7873cb83dacb17b0afc-Abstract.html.

153. O. Lartillot, T. Eerola, P. Toiviainen, J. Fornari, Multi-feature modeling of pulse clarity: Design, validation and optimization, in *Proceedings of the 9th International Conference on Music Information Retrieval* (International Society for Music Information Retrieval, Philadelphia, PA, USA, 2008), pp. 521–526.

154. G. Peeters, "A large set of audio features for sound description (similarity and classification) in the Cuidado Project" (Technical Report, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 2004).

155. J. D. Johnston, Transform coding of audio signals using perceptual noise criteria. *IEEE J Sel Areas Commun* **6**, 314–323 (1988).

156. R. Villing, "Hearing the moment: Measures and models of the perceptual centre," thesis, National University of Ireland Maynooth (2010).

157. P. A. Barbosa, P. Arantes, A. R. Meireles, J. M. Vieira, Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors, in *Interspeech 2005* (ISCA, 2005), pp. 1441–1444; www.isca-speech.org/archive/interspeech_2005/barbosa05_interspeech.html.

158. I. Chow, M. Belyk, V. Tran, S. Brown, Syllable synchronization and the P-center in Cantonese. *J. Phon.* **49**, 55–66 (2015).

159. A. M. Cooper, D. H. Whalen, C. A. Fowler, P-centers are unaffected by phonetic categorization. *Percept. Psychophys.* **39**, 187–196 (1986).

160. C. Cannam, C. Landone, M. Sandler, Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files, in *Proceedings of the 18th ACM*

*International Conference on Multimedia* (Association for Computing Machinery, 2010), pp. 1467–1468; https://doi.org/10.1145/1873951.1874248.

161. M. Dunn, S. J. Greenhill, S. C. Levinson, R. D. Gray, Evolved structure of language shows lineagespecific trends in word-order universals. *Nature* **473**, 79–82 (2011).

162. S. Brown, P. E. Savage, A. M.-S. Ko, M. Stoneking, Y.-C. Ko, J.-H. Loo, J. A. Trejaut, Correlations in the population structure of music, genes and language. *Proc. R. Soc. B Biol. Sci.* **281**, 20132072 (2014).

163. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci. Adv.* **7**, eabd9223 (2021).

164. S. Passmore, A. Wood, C. Barbieri, D. Shilton, H. Daikoku, Q. Atkinson, P. E. Savage, Global musical diversity is largely independent of linguistic and genetic histories. PsyArXiv 10.31234/osf.io/pty34 [Preprint] (2023). https://doi.org/10.31234/osf.io/pty34.

165. F. Sera, B. Armstrong, M. Blangiardo, A. Gasparrini, An extended mixed-effects framework for metaanalysis. *Stat. Med.* **38**, 5429–5444 (2019).

166. H. Bozdogan, Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987).

167. S. Watanabe, *Mathematical Theory of Bayesian Statistics* (Chapman and Hall/CRC, 2018).

168. G. S. Dell, M. F. Schwartz, N. Martin, E. M. Saffran, D. A. Gagnon, The role of computational models in neuropsychological investigations of language: Reply to Ruml and Caramazza (2000). *Psychol. Rev.* **107**, 635–645 (2000).

169. D. Fraser, Interpolation by the FFT revisited-an experimental investigation. *IEEE Trans. Acoust. Speech Signal Process.* **37**, 665–675 (1989).

170. R. W. Schafer, L. R. Rabiner, A digital signal processing approach to interpolation. *Proc. IEEE* **61**, 692–702 (1973).

171. M. Cychosz, A. Cristia, E. Bergelson, M. Casillas, G. Baudet, A. S. Warlaumont, C. Scaff, L. Yankowitz, A. Seidl, Vocal development in a large-scale crosslinguistic corpus. *Dev. Sci.* **24**, e13090 (2021).

172. F. Anvari, D. Lakens, Using anchor-based methods to determine the smallest effect size of interest. *J. Exp. Soc. Psychol.* **96**, 104159 (2021).

173. S.-H. Jung, Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104 (2005).

174. S. Pounds, C. Cheng, Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263– 4271 (2005).

175. M. Horn, C. W. Dunnett, Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Lect. Notes-Monogr. Ser.* **47**, 48–65 (2004).

176. L. V. Hedges, T. D. Pigott, The power of statistical tests in meta-analysis. *Psychol. Methods* **6**, 203– 217 (2001).

177. D. Jackson, R. Turner, Power analysis for random-effects meta-analysis. *Res. Synth. Methods* **8**, 290– 302 (2017).