# Supplementary Materials for "Universal interpretations of vocal music"

Lidya Yurdum[1,2,*], Manvir Singh[3], Luke Glowacki[4], Tom Vardy[5], Quentin D. Atkinson[5], Courtney B. Hilton[5], Disa Sauter[2], Max M. Krasnow[6], & Samuel A. Mehr[1,5,*]

[1]Child Study Center, Yale University, New Haven, CT 06520, USA.
[2]Department of Psychology, University of Amsterdam, Amsterdam 1018WT, Netherlands.
[3]Department of Anthropology, University of California, Davis, Davis CA 95616.
[4]Department of Anthropology, Boston University, Boston, MA 02215, USA.
[5]School of Psychology, University of Auckland, Auckland 1010, New Zealand.
[6]Division of Continuing Education, Harvard University, Cambridge, MA 02138, USA.

*Corresponding authors. E-mail: lidya.yurdum@yale.edu, sam@auckland.ac.nz

## 1.1 Deviations from the preregistration

We preregistered the study in November 2017 at https://osf.io/msvwz. The data collected and analyses reported deviate from the preregistration in five ways.

1. In the industrialised societies, we planned to collect data from 100 participants in each of 45 countries. Recruitment difficulties in some countries led us to increase the sampling range to include nearby countries where the same targeted language was also an official language. For instance, while we initially intended to recruit native speakers of English in Zambia, our sample from this region also included native speakers of English in nearby Namibia. This approach primarily affected African countries where internet access was limited relative to, e.g., the East Asian countries where we collected data via the same method.
2. In the smaller-scale societies, we planned to collect data in six communities, but due to the COVID-19 pandemic, we were only able to collect data in three.
3. For all participants, we planned the listening task to include 36 songs. This proved to be too long; in industrialised societies we shortened it to 24 songs, and in smaller-scale societies, to 18 songs.
4. To further shorten the task in industrialised societies, we reduced the number of dimensions on which participants rated each song; we planned to use four distractor dimensions but included only two in the full sample. The two we omitted ("...to tell a story" and "...to mourn the dead") had previously been studied in ref. (1). Participants in the smaller-scale societies completed all four distractor dimensions for each song, however.
5. We planned to collect data concerning listeners' intuitions surrounding two forms of songs: the original, naturalistic recordings from the *Natural History of Song Discography* as well as artificially produced (i.e., synthesised) versions of the songs, created using transcriptions of them reported in ref. (2). Due to limitations on the amount of data we could collect, we obtained far less data on listeners' responses to the synthesised songs than the naturalistic recordings. As such, we leave those data for a future paper. Note that this decision limited the number of participants in smaller-scale societies reported here, as roughly half of the participants studied in those societies heard synthesised songs rather than the naturalistic recordings.

## 1.2 Replication of confirmatory analyses using mixed-effects models

We replicated the main analyses of the industrialised society data using mixed-effects models with random intercepts for participant, song, and language. The results of these models conceptually replicated the simpler confirmatory analyses, but with slightly attenuated effect sizes.

Taking into account the variance accounted for by participant, stimulus and language, dance songs were rated significantly above the base rate on the "...for dancing" dimension ($\beta$ = 0.49, $SE$ = 0.08, $t(128.27)$ = 5.93, $p <$ .0001), and lullabies were rated significantly below the base rate ($\beta$ = -0.46, $SE$ = 0.08, $t(128.27)$ = -5.52, $p <$ .0001). On the "...to soothe a baby" dimension, lullabies were rated highest ($\beta$ = 0.52, $SE$ = 0.07, $t(131.92)$ = 7.57, $p <$ .0001), and dance songs lowest ($\beta$ = -0.29, $SE$ = 0.07, $t(131.92)$ = -4.15, $p <$ .0001); as in the song-level analyses, healing songs were rated below the base rate ($\beta$ = -0.17, $SE$ = 0.07, $t(131.06)$ = -2.36, $p$ = 0.02). On the "...to heal

illness" dimension, healing songs were rated higher than the base rate ($\beta$ = 0.12, $SE$ = 0.05, $t(132.26)$ = 2.30, $p$ = 0.02), whereas dance songs were rated below it ($\beta$ = -0.18, $SE$ = 0.05, $t(133.01)$ = -3.54, $p < .001$). Last, as in the song-level analyses, love songs were not reliably identified as "... to express love for another person" ($p > 0.05$).

## 1.3 Principal Components Analysis

To test whether the four behavioural context dimensions we used have distinct latent underpinnings, we conducted a principal components analysis using song-wise averages on the four dimensions, for both cohorts separately. Since the analysis is summarising only four dimensions, we present the loadings and eigenvalues for all four resulting components in Table S2 (n.b., as sign direction in principal components analysis is arbitrary, we reversed the loadings in the smaller-scale cohort for easier comparison to the industrialised society results).

In both cohorts, the first component loaded highly on "... for dancing" and low on "... to soothe a baby", capturing a latent behavioural context we could interpret as "inducing arousal". Plotting the four song types in principal components space shows that dance songs and lullabies are indeed most clearly distinguished along this first component, while healing songs are distinguished by the second component (Figure S2).

We then tested whether a given component differentiated the four song types. To do this, we calculated each of the four components' loadings, for each song, for the two cohorts separately. We then regressed each of the four components onto song type as a categorical predictor, with the intercept fixed at zero. This is equivalent to testing which of the four song types differ significantly from zero on a given component.

In both cohorts, the loadings for the first three components differed significantly based on song type. Lullabies' and dance songs' loadings both differed from zero on component 1: dance songs loaded positively on this dimension ($\beta$ = 1.19, $SE$ = 0.207, $p < .0001$), while lullabies loaded negatively ($\beta$ = -1.11, $SE$ = 0.207, $p < .0001$). Component 2 characterised healing songs and lullabies: healing songs loaded positively on this component ($\beta$ = 0.71, $SE$ = 0.221, $p$ = 0.002), while lullabies loaded negatively ($\beta$ = -0.79, $SE$ = 0.214, $p < .001$). Component 3 differentiated lullabies and love songs, with love songs loading positively on this component ($\beta$ = 0.25, $SE$ = 0.093, $p$ = 0.009), and lullabies loading negatively ($\beta$ = -0.21, $SE$ = 0.093, $p$ = 0.03).

Similarly, in the smaller-scale society cohort, component 1 also differentiated lullabies and dance songs: dance songs loaded positively on this dimension ($\beta$ = 0.72, $SE$ = 0.207, $p < .001$), while lullabies loaded negatively ($\beta$ = -0.97, $SE$ = 0.207, $p < .0001$). Once again, component 2 differentiated healing songs ($\beta$ = 0.63, $SE$ = 0.180, $p < .001$) and lullabies ($\beta$ = -0.40, $SE$ = 0.173, $p$ = 0.02). In contrast to the industrialised societies, the third component in the smaller-scale cohort did not differentiate any of the song types.

## 1.4 Sampling variation in cross-cohort correlations

In smaller-scale societies, each song excerpt was rated by fewer listeners than in industrialised societies. This sampling noise limits the explainable variance in our data and is likely to bias the cross-cohort correlations reported in Figure 3a downwards (3, 4). In line with an anonymous reviewer's suggestion, we compensated for this bias with a noise ceiling metric, as described in (5), estimated for each cohort's ratings on each dimensions using the equation below:

$$\sqrt{1 - \frac{mean(\text{squared SE per song on a given rating dimension})}{var(\text{mean score per song on a given rating dimension})}}$$

The correlation noise ceilings for the industrialised society cohort were close to 1, which is to be expected given the large number of ratings per song: 0.998 for the "... to soothe a baby" dimension, 0.999 for "... for dancing", 0.994 for "... to heal illness", and 0.994 for "... to express love for another person".

The noise ceilings for the smaller-scale societies were lower, reflecting the fact that we had fewer observations per song in the smaller-scale societies and that the ratings were on a 3-point scale rather than 4-point. The estimated noise ceilings for the four behavioural context dimensions were 0.645 for soothing a baby, 0.824 for dancing, 0.254 to heal illness, and 0.340 to express love.

We then normalised the cross-cultural correlations by the noise ceiling estimates, accounting for the estimated maximum explainable variance in each cohort's ratings (6). Note that the higher the noise ceiling, the smaller the adjustment to the observed correlation will be, and conversely, smaller noise ceilings will result in more dramatic adjustments. Because the noise ceiling estimates were especially low for the healing and expressing love contexts in

the smaller-scale society cohorts, we opted to make a more conservative adjustment by using the average of the two cohorts' noise ceilings, instead of normalising for each ceiling independently. (Using the average of two noise ceilings will always result in a more conservative adjustment than correcting for both ceilings independently.)

To adjust for this bias, we divided the correlations for each of the four behavioural contexts by the average noise ceiling in the two cohorts on that dimension, resulting in $r_{dancing} = 0.919$, $r_{soothing} = 0.713$, $r_{healing} = 0.541$, and $r_{love} = 0.362$ (corresponding to the unadjusted correlations reported in Figure 3a). While less biased, these scores still likely underestimate the true correlations, given the noisiness of the small-scale cohort stimuli-level ratings. (n.b., The larger sampling variability in the smaller-scale cohorts is an issue specific to the stimuli-level estimates of the ratings on each behavioural context dimension; the main comparisons reported at the level of the four behavioural contexts are still adequately powered.)

## 1.5 Acoustic Correlates of Musical Function Inferences

In previous work using the same corpus we demonstrated the consistency and distinctiveness with which musical features characterise dance songs, healing songs and lullabies worldwide (1, 2), as well as listeners' subjective ratings of their behavioural contexts (7). Here, we analysed which musical features were predictive of listeners' inferences for each song. This analysis differs from our prior work in refs. (1) and (2) where we focused on the musical features associated with the actual songs, rather than listener inferences. It also complements our work in ref. (7), which tested the influence of musical features on listener inferences in an English-speaking, Internet-connected sample using a citizen science approach and forced-choice task. Although we ran this analysis for the industrialised and smaller-scale society cohorts separately, we note that our sample size was much smaller in the smaller-scale cohort, and that this cohort also provided ratings on a 3-point scale (rather than 4), further reducing the sensitivity of their data. As such, we focused mostly on the results from the industrialised cohort.

Following the same procedure reported in ref. (7), we started with an initial set of 36 musical features, consisting of annotations from expert musicians and variables derived from the transcriptions of the songs. We then z-scored these variables. For each cohort separately, we used Least Absolute Shrinkage and Selection Operator (LASSO) regularization (8) to select a smaller subset of the full 36 features that were most relevant to each behavioural context dimension (dancing, soothing a baby, healing illness; given the overall "fuzziness" of love songs and our listeners' difficulty identifying them, we did not include ratings on this dimension in the analysis). Each model predicted the average listener ratings on a given dimension for each of the 118 songs and was trained with 10-fold cross-validation. Finally, we discarded resulting model coefficients below a threshold of .02, leaving us with a simpler, more conservative set of features that were most predictive of listeners' behavioural context ratings (see Table S3 for the list of selected musical features and their definitions). We then combined the smaller subset of features selected for the two cohorts into one pooled set for each behavioural context. Perhaps unsurprisingly, given the noisier data, the LASSO model did not pick out any features that predicted ratings on the "... to heal illness" dimension for the smaller-scale society cohort. Last, we regressed listeners' ratings on each dimension onto this group of features. The results are reported in Table S4.

A variety of musical features were predictive of listeners' ratings on the behavioural context dimensions, as in previous work (7), and many of these features were the same ones that universally characterise dance songs, lullabies and healing songs. For example, dance songs and lullabies are universally differentiated by the prevalence of accents and the songs' tempo (2) and both of these features guided listeners' ratings on the "... for dancing" and "... to soothe a baby" dimensions in the present data (Table S4). Similarly, a consistent metric structure is associated with dance songs globally (2) and this feature predicted listener ratings on the "... for dancing" dimension. We also found that tempo and consistency in macrometer, which characterised healing songs, guided our participants' inferences on the "... to heal illness" dimension.

In sum, we find evidence that the acoustic forms of a song reliably predicted listeners' interpretations of that song, and the specific musical features at work tended to be those that universally characterised the songs found globally in each behavioural context.

**Figure S1 | Testing setup in smaller-scale societies.** The photo depicts author M.S. testing a Mentawai participant in Indonesia. In each of the smaller-scale societies, participants sat across from the experimenter, listened on headphones only, and entered their responses on a button box. The experimenter was unaware of the song being played on each trial and the participant could not see the laptop's screen.

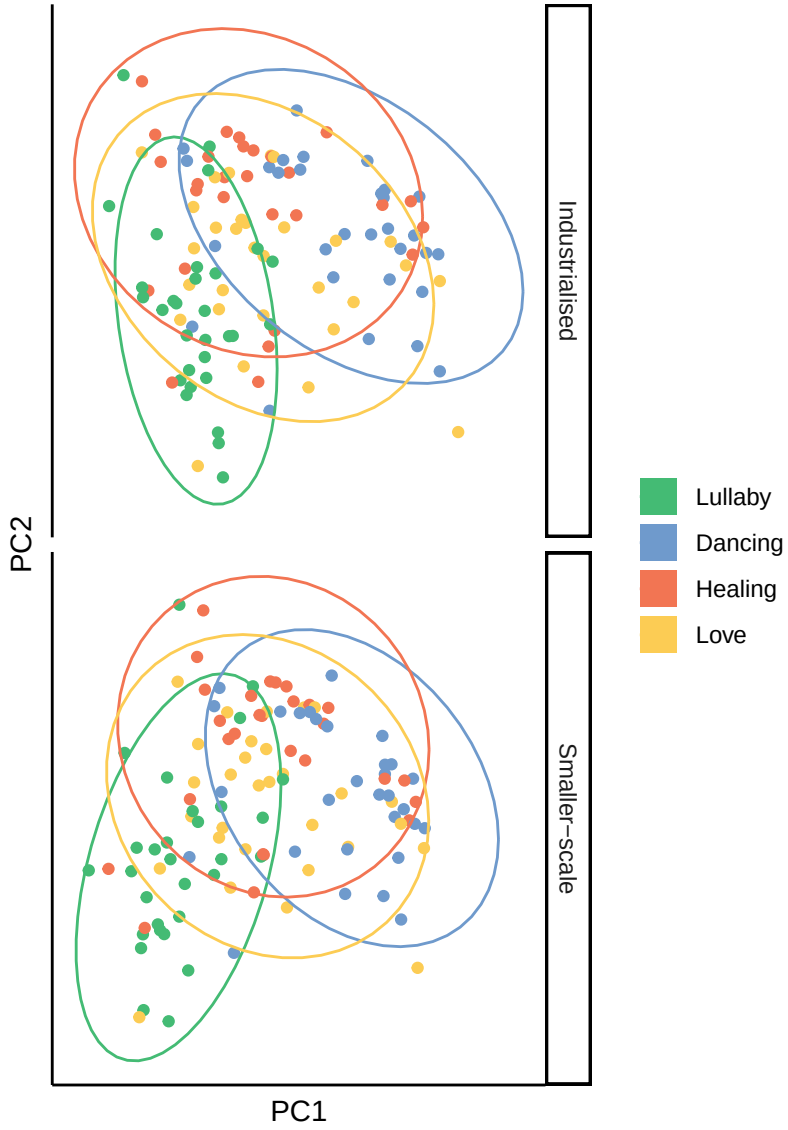


**Figure S2 | The principal components space of listener interpretations.** We ran two separate principal components analyses to summarise the four behavioural context dimensions with a smaller number of latent dimensions. In both cohorts, dance songs and lullabies were clearly differentiated along the first component, which loaded positively on "...for dancing'' and negatively on "...to soothe a baby''. In both cohorts, healing songs loaded positively on the second component. The principal component scores for each song are depicted by the dots; colours indicate the song type. The circles denote the 95% confidence interval.

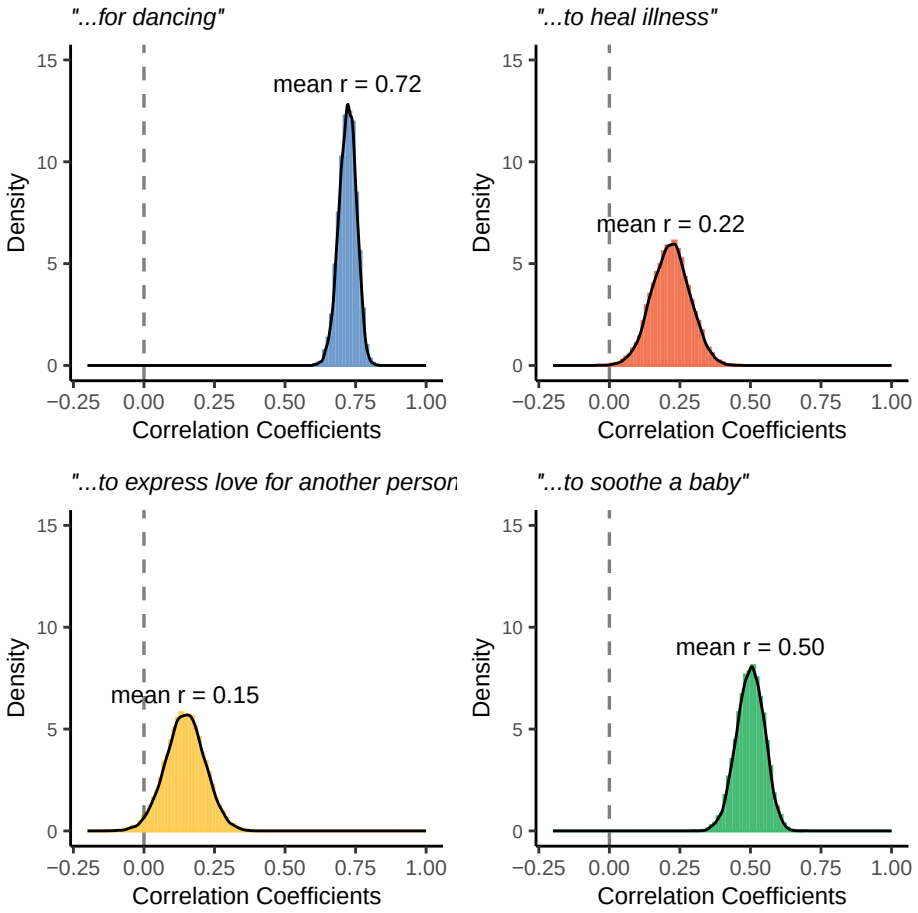


**Figure S3 | Bootstrapped correlations between song-wise ratings from the industrialised societies and the smaller-scale societies.** As an alternative to correlations across cohorts reported in the main text, we computed distributions of correlations via stratified bootstrapping. This approach helps to account for large differences in sample sizes between the cohorts and provides a principled estimate of the variability in each correlation coefficient. We sampled 30 observations per song from each cohort, generated new song-wise averages, and correlated these averages across both cohorts. This procedure was repeated 10,000 times. The plots show the four distributions correlations; in all four cases, the correlations were significantly larger than 0, but they varied in magnitude across behavioural contexts.

| | Estimate | Std. Error | df | t value | p value |
|---|---|---|---|---|---|
| **Dance songs** | | | | | |
| Intercept | 2.91 | 0.10 | 29.24 | 28.51 | 0.00 |
| Shared Language | 0.04 | 0.02 | 32320.46 | 1.99 | 0.05 |
| Shared Sub-Region | 0.17 | 0.04 | 31136.96 | 4.08 | 0.00 |
| Interaction | -0.04 | 0.08 | 31516.69 | -0.54 | 0.59 |
| **Lullabies** | | | | | |
| Intercept | 2.64 | 0.09 | 29.35 | 30.78 | 0.00 |
| Shared Language | 0.05 | 0.03 | 31529.05 | 1.72 | 0.09 |
| Shared Sub-Region | 0.10 | 0.04 | 31285.53 | 2.17 | 0.03 |
| Interaction | 0.15 | 0.09 | 31664.08 | 1.66 | 0.10 |
| **Healing Songs** | | | | | |
| Intercept | 2.63 | 0.06 | 27.65 | 45.19 | 0.00 |
| Shared Language | 0.02 | 0.03 | 27283.82 | 0.55 | 0.58 |
| Shared Sub-Region | 0.05 | 0.05 | 28487.64 | 1.00 | 0.32 |
| Interaction | -0.03 | 0.10 | 28663.04 | -0.31 | 0.76 |
| **Love Songs** | | | | | |
| Intercept | 2.55 | 0.05 | 30.10 | 48.06 | 0.00 |
| Shared Language | 0.05 | 0.02 | 28053.63 | 1.90 | 0.06 |
| Shared Sub-Region | 0.07 | 0.04 | 29923.48 | 1.84 | 0.07 |
| Interaction | -0.03 | 0.08 | 30229.86 | -0.42 | 0.67 |

**Table S1** To test for a super-additive effect of linguistic and geographic proximity, we regressed the target behavioural context ratings (on their relevant dimension) onto two binary variables: language family (shared vs. different) and geographic subregion (shared vs. different), with random intercepts for participant and song. After including both language family and geographic subregion in the regression, sharing a language predicted higher ratings for dance songs only. Geographic proximity was associated with higher ratings on the appropriate dimensions for lullabies and dance songs. Super-additivity would be indicated by a significant interaction between the effect of linguistic and geographic proximity, such that the effect of sharing a geographic region depends on whether the listener is also more familiar with the language of the song. However, the interaction between the two variables was not significant for any of the four behavioural contexts.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| **Industrialised Societies** | | | | |
| Eigenvalues | 1.94 | 1.65 | 0.29 | 0.12 |
| Explained Variance | 49% | 41% | 7% | 3% |
| **Loadings** | | | | |
| ... for dancing | 0.65 | 0.25 | -0.11 | 0.70 |
| ... to heal illness | -0.55 | 0.43 | 0.56 | 0.45 |
| ... to soothe a baby | -0.44 | -0.57 | -0.43 | 0.54 |
| ... to express love for another person | 0.27 | -0.66 | 0.70 | 0.09 |
| **Smaller-Scale Societies** | | | | |
| Eigenvalues | 1.64 | 1.02 | 1.00 | 0.34 |
| Explained Variance | 41% | 26% | 25% | 9% |
| **Loadings** | | | | |
| ... for dancing | 0.71 | -0.01 | -0.02 | 0.70 |
| ... to heal illness | -0.13 | 0.61 | 0.77 | 0.16 |
| ... to soothe a baby | -0.53 | -0.60 | 0.27 | 0.53 |
| ... to express love for another person | 0.44 | -0.52 | 0.58 | -0.44 |

**Table S2.**
We ran two principal component analyses, one for each cohort, using the mean ratings for each song on the four behavioural context dimensions. We reversed the signs of the loadings in the smaller-scale society cohort to match those of the industrialised cohort. In the industrialised cohort, the first two components explained 90 percent of the observed variance. The first component loads highly on "...for dancing" and low on "...to soothe a baby", and captures something comparable to arousal. In the smaller-scale society cohort, the first three components explained over 90 percent of the variance. Again, the first component loaded positively on "...for dancing" and negatively on "...to soothe a baby".

| Source | Feature | Definition | Variable Name | Behavioural Context |
|---|---|---|---|---|
| **Expert Annotation** | | | | |
| | Macrometer consistency (ordinal) | The perceived clarity of the metrical rhythmic structure (groupings of perceived strong and weak beats). Consistency of macrometer converted to ordinal scale, from "No macrometer" (1) to "Totally clear macrometer" (6). | macrometer_ord | for dancing, to soothe a baby, to heal illness |
| | Tempo (adjusted) | The rate of salient rhythmic pulses (measured by having annotators tap the beat), adjusted to correspond with "quarter note" equivalent; measured in beats per minute; the perceived speed of the music. A fast song will have a high value. | tempo_adj | for dancing, to soothe a baby, to heal illness |
| | Accent | The differentiation of musical pulses, usually by volume or emphasis of articulation. A fluid, gentle song will have few accents and a correspondingly low value. | accent | for dancing, to soothe a baby |
| | Triple rhythm | The presence of triple subdivisions of the beat. | micrometer_triple | for dancing, to soothe a baby |
| | Duple rhythm | The presence of duple subdivisions of the beat. | micrometer_duple | for dancing |
| | Ornamentation | Complex melodic variation or "decoration" of a perceived underlying musical structure. A song perceived as having ornamentation is annotated with a value of 1. | ornament | for dancing, to soothe a baby |
| | Syncopation | The degree of syncopation in the song (roughly equateable with "rhythmic complexity"). Syncopation is when rhythmic accents align with metrically weak positions of a metrical beat structure (e.g., having accents half way between when people tap their foot rather than directly aligned with the foot taps). | syncopate | for dancing, to heal illness |
| | Duple macrometer present | Duple macrometer present. | macrometer_duple | for dancing |
| | Triple macrometer present | Triple macrometer present. | macrometer_triple | to heal illness |
| | Tension | The degree to which the passage is perceived to build and release tension via changes in melodic contour, harmonic progression, rhythm, motivic development, accent, or instrumentation. If so, the song is annotated with a value of 1. | tension | to heal illness |
| | Vibrato present | Presence of vibrato in the singing. Vibrato is a rapid, slight oscillation in pitch that is typically perceived as adding intensity. | vibrato | to heal illness |
| **Transcription** | | | | |
| | Variety in pitch | A pitch class is the group of pitches that sound equivalent at different octaves, such as all the Cs, not just middle C. This variable, an indicator of melodic variety, counts the number of pitch classes that appear at least once in the song. | pitch_class_variety | for dancing |
| | Melodic Thirds | Prevalence of 3 or 4 semitone intervals. | melodic_thirds | for dancing, to heal illness |
| | Average Note Duration | Average duration of a note, in seconds. | average_note_duration | for dancing, to soothe a baby |

*(continued)*

| Source | Feature | Definition | Variable Name | Behavioural Context |
|---|---|---|---|---|
| | Overall direction of motion | Number of rising melodic intervals divided by number of intervals that are either rising or falling—that is, fraction of moving intervals that are rising (unisons are ignored). If a piece has no moving intervals, this field is 0. This feature considers intervals across rests as contributing to the direction of motion. | direction_of_motion | for dancing, to soothe a baby, to heal illness |
| | Pitch range | The difference between the highest and lowest pitches, in semitones. | range | for dancing, to soothe a baby |
| | Estimated simplified mode of the transcription | Quality or mode of the transcription (major or minor) based on the Krumhansl-Schmuckler key-finding algorithm. This is done by finding the most likely key and then returning the mode of that key – rather than weighting the likelihood of all major and minor keys. 0 = Major, 1 = Minor. | quality | to soothe a baby |
| | Prevalence of modal pitch class | Variety versus monotony of the melody, measured by the ratio of the proportion of occurrences of the second most common pitch (collapsing across octaves) to the proportion of occurrences of the most common pitch; monotonous melodies will have low values. | modal_pitchcls_prev | to soothe a baby, to heal illness |
| | Prevalence of stepwise motion | Fraction of melodic intervals one or two semitones in size. | stepwise_motion | to soothe a baby, to heal illness |
| | Relative strength of most-common intervals | Fraction of melodic intervals that belong to the second most common interval divided by the fraction of melodic intervals belonging to the most common interval. This field is 0 if there are not two distinct most common melodic intervals. | rel_strength_modal_intervals | to soothe a baby |
| | Note density | Average number of notes per second, using durations from NHSDiscography_metadata. | note_density | to soothe a baby |
| | Amount of arpeggiation | Fraction of horizontal intervals that are repeated notes, minor thirds, major thirds, perfect fifths, minor sevenths, major sevenths, octaves, minor tenths or major tenths. | amount_of_arpeggiation | to heal illness |
| | Interval between modal pitch classes | Absolute value of the difference between the two most common pitch classes, in semitones. This field is 0 if there are not two distinct most common pitches. | interval_btwn_strongest_pitchcls | to heal illness |

**Table S3.** Musical features selected by the LASSO procedure. We derived these features from Transcription and Expert Annotation datasets from Mehr et al. (2019; denoted by the "Source" column). The original names of the variables used in the analyses are noted in the "Variable Name" column. The "Behavioural context" column indicates which of the three behavioural context ratings the feature was selected for.

| Behavioural Context | Cohort | Feature | Estimate | Standard Error | Statistic | p value |
|---|---|---|---|---|---|---|
| ...for dancing | Industrialised | (Intercept) | -5.28 | 0.61 | -8.71 | $< .0001$ |
| | | Macrometer consistency | 0.53 | 0.08 | 6.49 | $< .0001$ |
| | | Tempo (adjusted) | 0.01 | 0.00 | 6.85 | $< .0001$ |
| | | Accent | 1.99 | 0.35 | 5.73 | $< .0001$ |
| | | Ornamentation | 0.43 | 0.19 | 2.26 | $= 0.026$ |
| | | Average note duration | 0.55 | 0.23 | 2.33 | $= 0.021$ |
| | Smaller-scale | (Intercept) | -3.56 | 0.78 | -4.55 | $< .0001$ |
| | | Macrometer consistency | 0.39 | 0.10 | 3.71 | $< .001$ |
| | | Tempo (adjusted) | 0.01 | 0.00 | 5.77 | $< .0001$ |
| | | Accent | 1.04 | 0.45 | 2.32 | $= 0.022$ |
| | | Triple rhythm | -1.15 | 0.54 | -2.11 | $= 0.037$ |
| | | Ornamentation | 0.52 | 0.24 | 2.14 | $= 0.034$ |
| | | Melodic thirds | 1.30 | 0.57 | 2.28 | $= 0.024$ |
| ...to soothe a baby | Industrialised | (Intercept) | 3.99 | 0.65 | 6.13 | $< .0001$ |
| | | Accent | -2.36 | 0.43 | -5.51 | $< .0001$ |
| | | Tempo (adjusted) | -0.01 | 0.00 | -2.83 | $= 0.006$ |
| | | Ornamentation | -1.03 | 0.25 | -4.12 | $< .0001$ |
| | | Triple rhythm | 0.63 | 0.23 | 2.71 | $= 0.008$ |
| | | Mode of transcription | 0.29 | 0.13 | 2.22 | $= 0.029$ |
| | | Average note duration | -0.85 | 0.31 | -2.71 | $= 0.008$ |
| | | Prevalence of modal pitch class | -1.34 | 0.55 | -2.44 | $= 0.016$ |
| | | Relative strength of most-common intervals | -0.68 | 0.26 | -2.57 | $= 0.012$ |
| | | Macrometer consistency | -0.15 | 0.07 | -2.20 | $= 0.03$ |
| | Smaller-scale | (Intercept) | 3.50 | 0.85 | 4.10 | $< .0001$ |
| | | Accent | -1.16 | 0.56 | -2.06 | $= 0.042$ |
| | | Mode of transcription | 0.46 | 0.17 | 2.69 | $= 0.008$ |
| | | Prevalence of modal pitch class | -1.75 | 0.72 | -2.42 | $= 0.017$ |
| | | Macrometer consistency | -0.25 | 0.09 | -2.71 | $= 0.008$ |
| ...to heal illness | Industrialised | (Intercept) | 4.00 | 1.02 | 3.93 | $< .001$ |
| | | Macrometer consistency | -0.39 | 0.07 | -5.94 | $< .0001$ |
| | | Tension | -0.64 | 0.17 | -3.69 | $< .001$ |
| | | Tempo (adjusted) | -0.01 | 0.00 | -2.64 | $= 0.01$ |
| | | Vibrato present | 0.71 | 0.23 | 3.08 | $= 0.003$ |
| | | Melodic thirds | -1.28 | 0.61 | -2.09 | $= 0.039$ |
| | Smaller-scale | Amount of arpeggiation | 3.52 | 1.36 | 2.59 | $= 0.011$ |
| | | Melodic thirds | -1.83 | 0.89 | -2.06 | $= 0.041$ |

**Table S4** The musical features that influence ratings on each of the three behavioural context dimensions. We first selected a smaller subset of all musical features using LASSO for each cohort, and then regressed the ratings for each dimension onto the (pooled) selected features. For brevity, we report only the features that were predictive of listener ratings, and their directions.

# References


1. S. A. Mehr, M. Singh, H. York, L. Glowacki, M. M. Krasnow, Form and function in human song. *Current Biology* **28**, 356–368 (2018).
2. S. A. Mehr, *et al.*, Universality and diversity in human song. *Science* **366**, 957–970 (2019).
3. G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, D. A. Drummond, Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS genetics* **11**, e1005206 (2015).
4. C. Spearman, The proof and measurement of association between two things. *The American journal of psychology* **100**, 441–471 (1987).
5. A. S. Cowen, D. Keltner, Universal facial expressions uncovered in art of the ancient Americas: A computational approach. *Science advances* **6**, eabb1005 (2020).
6. A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, D. Keltner, The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour* **3**, 369–382 (2019).
7. C. B. Hilton, L. Crowley-de Thierry, R. Yan, A. Martin, S. A. Mehr, Children infer the behavioral contexts of unfamiliar foreign songs. *Journal of Experimental Psychology: General* (2022) https:/doi.org/https://doi.org/10.1037/xge0001289.
8. J. Friedman, T. Hastie, R. Tibshirani, Lasso and elastic-net regularized generalized linear models. Rpackage version 2.0-5. (2016).