Supplementary Methods

Technical details of the acoustic feature extraction

We extracted acoustic features with four sets of tools, described below, and also preprocessed them to reduce the influence of atypical observations.

Praat

We extracted intensity, pitch, and first and second formant values from the denoised recordings every 0.03125 seconds. For female participants, the pitch floor was set at 100 Hz, with a pitch ceiling at 600 Hz, and a maximum formant of 5500 Hz. For male participants, these values were 75 Hz, 300 Hz, and 5000 Hz, respectively. From these data, several summary values were calculated for each recording: mean and maximum first and second formants, mean pitch, and minimum intensity. In addition to these summary statistics, we measured the intensity and pitch rates as change in these values over time. For vowel measures, the first and second formants were used to calculate both the average vowel space used, as well as the vowel change rate (measured as change in Euclidean formant space) over time.

MIRtoolbox

All MIRtoolbox (v. 1.7.2) features were extracted with default parameters[1]. *mirattackslope* returns a list of all attack slopes detected, so final analyses were done on summary features (e.g., mean, median, etc.). Final analyses were also done on summary features for *mirroughness*, which returns time series data of roughness measures in 50ms windows. We RMS-normalized the mean of *mirroughness*, following previous work[2]. MIRtoolbox features were computed on the denoised recordings, with the exception of *mirtempo* and *mirpulseclarity*, where removing the silences between vocalizations would have altered the tempo.

Rhythmic variability

For temporal modulation spectra we followed a previous method[3], which combines discrete Fourier transforms applied to contiguous six-second excerpts. To analyze the entirety of each recording, we appended all recordings with silence to be exact multiples of six-seconds. The location of the peak (Hz) and variance of the temporal modulation spectra were extracted from their RMS values. Because intervening silence would influence temporal modulation measures, we computed them on recordings *before* they had been denoised.

Normalized pairwise variability index (nPVI)

The nPVI represents the temporal variance of data with discrete events, which makes it especially useful for comparing speech and music[4]. We used an automated syllable- and phrase-detection algorithm to extract events[5]. We computed nPVI in two ways: by averaging the nPVI of each phrase within a recording, as well as by treating the entire recording as a single phrase. Because intervening silence would influence nPVI measures, we computed them on recordings *before* they had been denoised.

Preprocessing

Automated acoustic analyses are highly sensitive at extremes (e.g., impossible values caused by non-vocal sounds, like loud wind). To correct for these issues, we Winsorized all acoustic variables. This process defines observations exceeding the lowest and highest 5 percentile ranks as outliers, recoding them as the values of those percentile boundaries. These data were used for all acoustic analyses. This approach is generally preferable to trimming extreme values, as trimming overcompensates for outliers by removing them entirely[6].

Analyses of the acoustic features using an alternate method (i.e., imputing extreme values with the mean observation for each feature within each fieldsite) yielded comparable results; readers are welcome to try alternate trimming methods with the open data and materials.

In the cases of three acoustic features (roughness, vowel travel rate, and pulse clarity), we used log-transformed data, because the raw data were highly skewed. This decision was supported by the exploratory-confirmatory approach; that is, results replicated across both exploratory and confirmatory samples in the log-transformed data.

Quantifying sensitivity with signal detection theory

To quantify sensitivity to infant-directedness in speech and song in the naïve listener experiment, and to quantify their response biases, we computed the metrics of d' and c (*criterion*) over the stimuli. These quantities were calculated with standard techniques from signal detection theory[7].

Specifically, a response on a given trial was coded as a `hit` if the trial was an infant-directed vocalization and the participant correctly responded with `baby`; a `miss` if for an infant-directed vocalization, they responded `adult`; a `false-alarm` if for an adult-directed vocalization, they responded `baby`; and a `correct-reject` if for an adult-directed vocalization, they correctly responded `adult`.

The hit rate H was then computed as the total number of hits for a given recording, divided by the total number of hits plus the misses; the false-alarm rate F was computed as the total number of misses for a given recording, divided by the total number of false-alarms plus the correct-rejects. These scores were then conservatively adjusted with the log-linear correction for extreme scores[8], and finally d' was estimated via the following equation, where the function $z(\cdot)$ represents the inverse of the normal cumulative distribution function:

$$d' = z(H) - z(F)$$

Criterion (c) was estimated as:

$$c = \frac{-(z(H) - z(F))}{2}$$

Additional naïve listener data collection via Prolific

In revising this manuscript, we discovered that a small subset of the corpus had been erroneously excluded from the naïve listener experiment. In most cases, these were recordings that had been too-conservatively edited to be too short to include in the experiment (but could reasonably be edited to include longer sections of audio); in some other cases, the original excerpting included confounding background noises that, upon additional editing, were avoidable. To ensure maximal coverage of the fieldsites studied here, we re-excerpted the audio of 103 examples and collected supplemental naïve listener data on these recordings via a Prolific experiment (N = 97, 54 male, 42 female, 1 other, mean age = 29.7 years). The Prolific experiment was identical to the citizen-science experiment, except that each participant was paid US$15/hr, rather than volunteering; and each participant rated 188 recordings instead of up to 16.

We included in the Prolific experiment the set of recordings that were erroneously excluded from the citizen-science experiment, along with 85 additional recordings randomly selected from those that *were* included in the citizen-science experiment, so as to ensure that each Prolific participant heard a balanced set of vocalization types. The two cohorts' ratings of the recordings in common across the two experiments were highly correlated ($r = 0.95$, $p < 0.0001$; two-sided test), demonstrating that they had similar intuitions concerning infant-directedness in speech and song. As such, in the main text, we report all the ratings together without disambiguating between the cohorts.

Supplementary Results

Alternate analysis of acoustic features via principal-components approach

We conducted an exploratory principal components analysis of the full 94 acoustic variables (Extended Data Fig. 2). The first three principal components accounted for 39% of total variability in acoustic features. The results provide convergent evidence that the main forms of acoustic variation partition into orthogonal clusters that most strongly distinguish speech from song overall (in PC1); most strongly distinguish infant-directedness in *song* (in PC2); and most strongly distinguish infant-directedness in *speech* (in PC3). Factor loadings are in Supplementary Table 7; these largely corroborate the findings of the LASSO and exploratory-confirmatory analyses.

One further pattern that the principal components analysis highlights is that infant-directedness makes speech more "songlike", in terms of higher pitch and reduced roughness (PC3); but speech strongly differed from song overall in terms of the variability and rate of variability of pitch, intensity, and vowels, and infant-directedness further exaggerated these differences for speech (PC1).

Robustness tests of main results in naïve listener experiment

On the suggestion of an anonymous reviewer, we repeated the main analyses of the naïve listener experiment (i.e., estimated sensitivity to infant-directedness in speech and song) with two alternate data exclusion strategies. First, the analyses and figures in the main text only study ratings of recordings that contained minimal extraneous sounds (such as a baby crying; see Methods). To ensure that the exclusion of these recordings did not account for the main findings, we repeated the analyses while including ratings of *all* recordings, including those with putatively confounding background sounds. They robustly replicated, with comparable effect sizes (speech: $d' = 1.13$, $t_{(4.78)} = 3.42$, 95% CI [0.48, 1.77], $p = 0.02$; song: $d' = 0.54$, $t_{(4.61)} = 3.35$, 95% CI [0.23, 0.86], $p = 0.023$).

A further potential confound concerns listeners' familiarity with the languages spoken or sung in the recordings. In the main text analyses, we explicitly model the expected differences in sensitivity that could result from lower or higher degrees of linguistic relatedness between the vocalizer and the listener (see, e.g., Fig. 3c). However, because the experiment was only conducted in English, many participants likely could understand at least some parts of the English-language vocalizations. To ensure that these recordings did not account for the main findings, we repeated the analyses while excluding all English-language recordings. These recordings came predominantly from the Wellington, San Diego, and Toronto fieldsites (where nearly all recordings were in English) but also appeared elsewhere, such as the Arawak fieldsite (where English Creole recordings were often comprehensible to English speakers), and in a few other sites, when a speaker happened to be bilingual and produce English-language vocalizations. The results replicated with these exclusions, although the estimated effect was weaker in song (speech: $d' = 0.79$, $t_{(4.02)} = 3.01$, 95% CI [0.28, 1.30], $p = 0.039$; song: $d' = 0.37$, $t_{(3.91)} = 3.00$, 95% CI [0.13, 0.62], $p = 0.041$).

Demographic analyses of a subsample of naïve listeners

An anonymous reviewer raised the possibility that conducting the naïve listener experiment online, as opposed to in a laboratory, reduced the diversity of the sample; if so, this could bias the results of the experiment, in principle. To test this question, we analyzed demographic information from participants living in the United States, who provided income, education level, and ethnicity data.

Descriptive statistics revealed that the subsample of United States participants was highly diverse (Supplementary Table 6), including, for example, representation from all ethnicity categories currently defined by the National Institutes of Health, and a broad range of annual household incomes. The sample was generally more representative of the United States population than are samples recruited in typical laboratory studies, which may skew towards wealthier samples with representation of fewer ethnicity categories[9,10].

Nevertheless, we proceeded by asking whether demographic factors were likely to affect people's ability to perceive infant-directedness. We ran mixed-effect regressions for each of the available demographic variables with random intercepts for the vocalist in the recording, and fixed effects for vocalization type and the demographic factor. While the main effects of income, education, or race on task performance were statistically significant ($ps < 0.0001$), in all cases, the effect sizes were tiny, explaining ~0.1% of variance in the model. These findings imply that the choice of a citizen-science approach likely did not bias the results of the experiment, at least in United States participants.

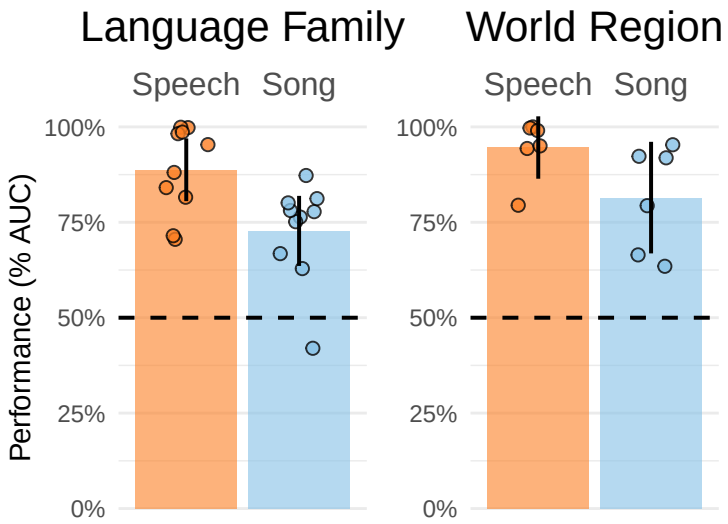
Society-level predictors for naïve listener data

Listener sensitivity within each fieldsite was correlated with a number of society-level characteristics: rank-order population size (speech: $\tau = 0.51$; song: $\tau = 0.58$), distance from fieldsite to nearest urban center (speech: r = -0.78; song: r = -0.51), and number of children per family (speech: r = -0.53; song: r = -0.72; all $ps < .001$ from two-sided tests). Each of these predictors were highly correlated with each other (all $r >$ 0.6), however, suggesting that they did not each contribute unique variance. There was no correlation with ratings of how frequently infant-directed vocalizations were used within each society ($ps > .4$). These findings suggest that at least some cross-fieldsite variability in listener sensitivity to infant-directedness is attributable to the *cultural* relatedness between vocalizers and listeners (as opposed to the *linguistic* relatedness analyzed in in the Main Text and Fig. 3c).

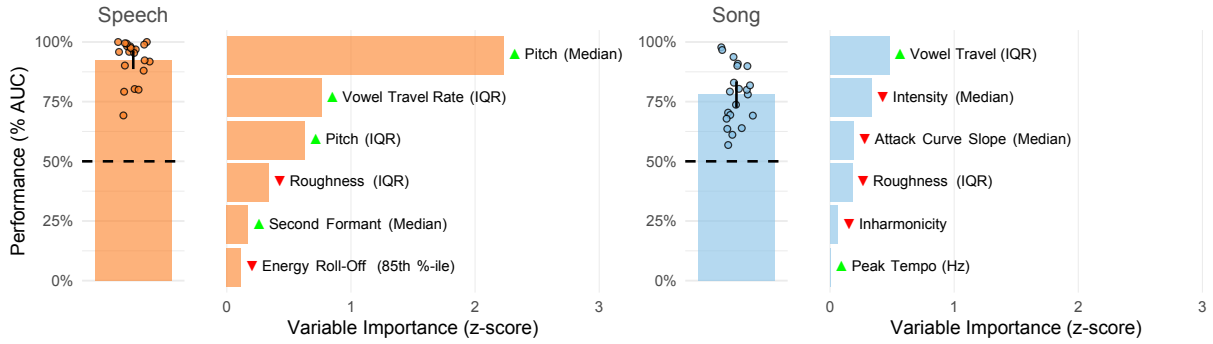
Simulated infant-directed vocalizations

Prior research has shown that simulated infant-directedness is qualitatively similar, albeit less exaggerated than when authentic, for both speech[11] and song[12]. Indeed, a model of the naïve listener results adjusting for fieldsite indeed showed a small decrease in "baby" guesses when an infant was not present (ID song: 6.4%, ID speech: 7.5%, AD song: -6.5%, AD speech: -4.2%, $ps < .0001$), but this effect was not stronger for vocalizations that were infant-directed compared to adult-directed ($\chi^2(1) = 2.93$, $p = 0.087$). Both the naive listener results and acoustic analyses were robust to whether these simulated infant-directed vocalizations were included or excluded, however, implying that the use of simulated infant-directed vocalizations did not undermine the robustness of the main effects.

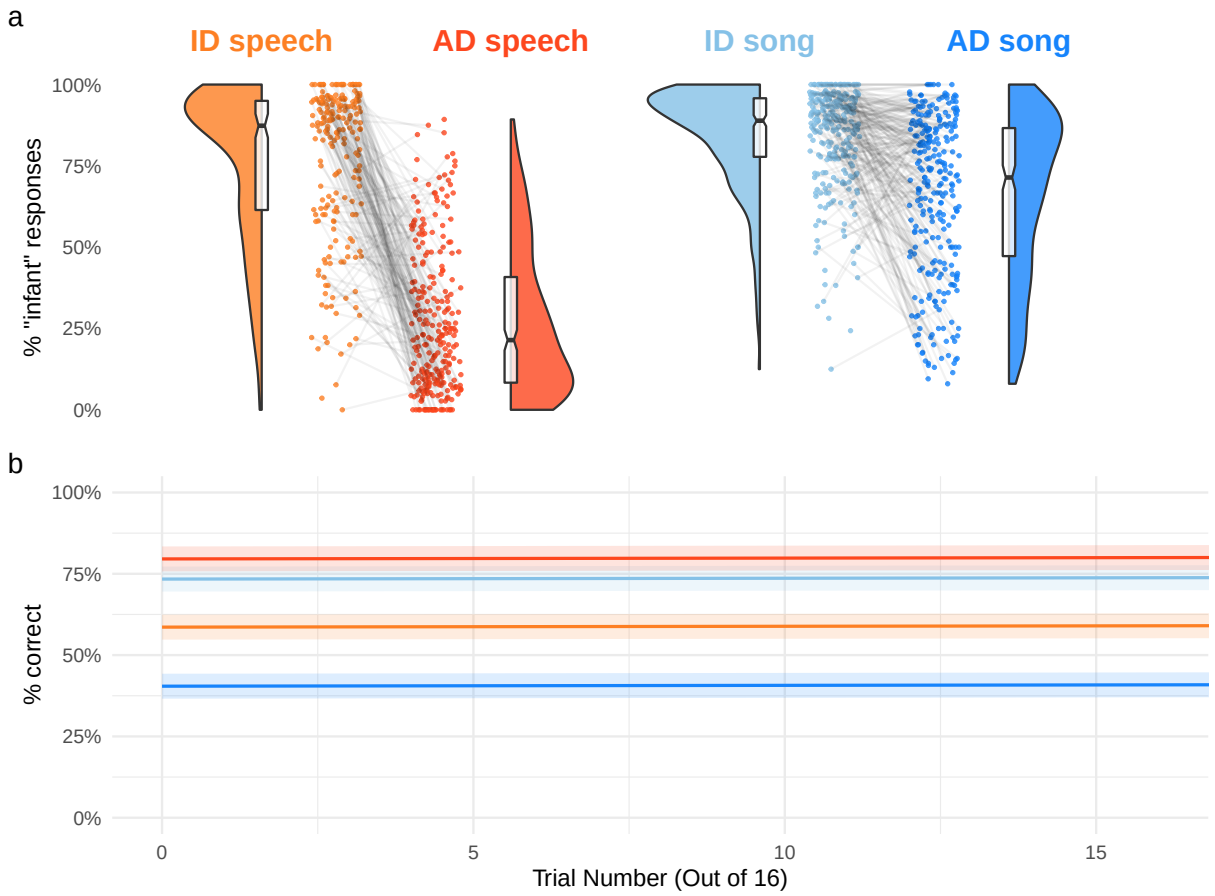
Supplementary Figures



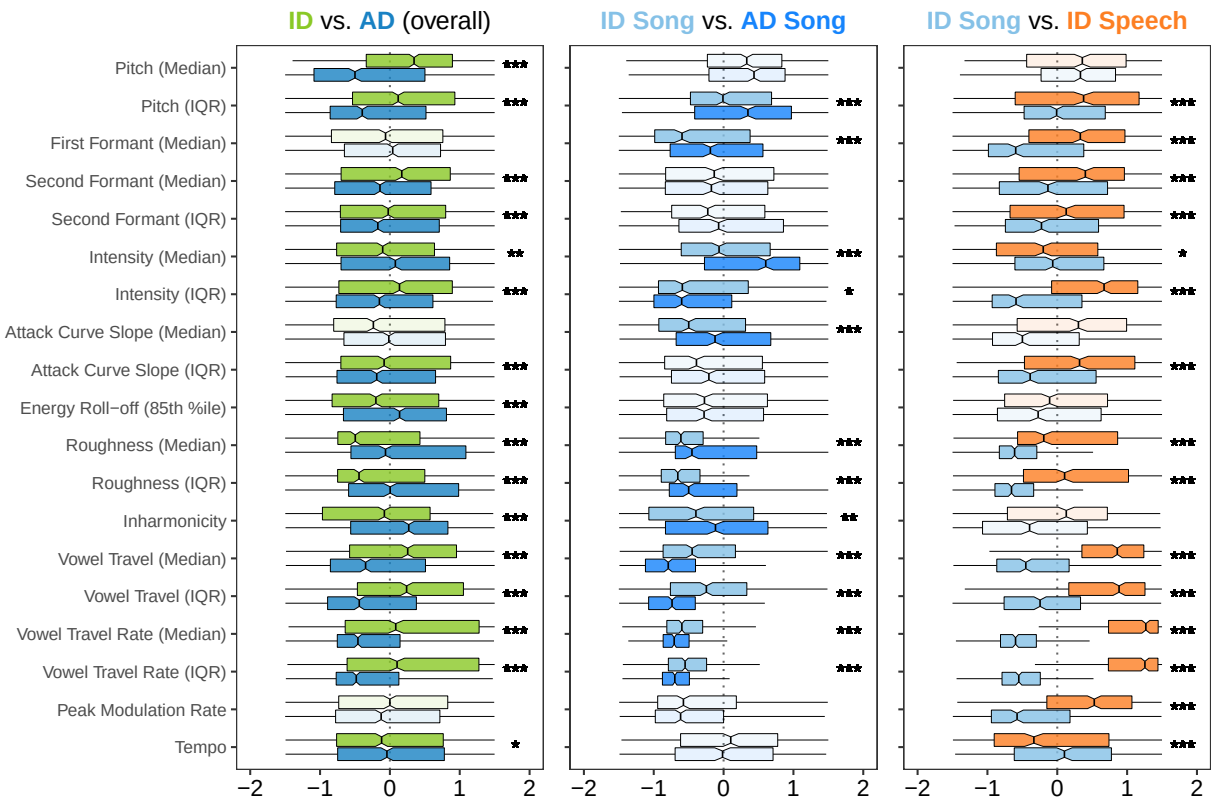
Supplementary Fig. 1 | LASSO classification of acoustic features with alternate cross-validation approaches. We repeated the main LASSO analysis (Fig. 1b) twice, but rather than conducting k-fold cross-validation across fieldsites, we did so across language families and world regions (see descriptive information about the fieldsites in Table 1). The results replicated robustly across both models, with corpus-wide classification performance significantly above chance in all cases. The vertical bars represent the mean classification performance across the cross-validation units (11 language families and 6 world-regions, respectively; quantified via receiver operating characteristic/area under the curve; AUC); the error bars represent 95% confidence intervals of the mean; the points represent the performance estimate for each language family or world region; and the horizontal dashed lines indicate chance level of 50% AUC.



Supplementary Fig. 2 | Replication of main LASSO results using unedited audio. As a test of robustness, we repeated the main LASSO analyses (Fig. 1b) with acoustic features extracted from raw, unedited audio. This approach ensures that the main results are not attributable to idiosyncrasies in the audio introduced by the editing process. The results repeated robustly, with above-chance performance in all fieldsites for both speech and song, and with the 3 most influential acoustic features selected by the model repeating across both specifications (see Fig. 1b). The vertical bars represent the overall classification performance (quantified via receiver operating characteristic/area under the curve; AUC); the error bars represent 95% confidence intervals; the points represent the average performance for each fieldsite ($n = 21$ fieldsites); and the horizontal dashed lines indicate chance level of 50% AUC. The horizontal bars show the acoustic characteristics with the largest influence in each classifier; the green and red triangles indicate the direction of the effect, e.g., with median pitch having a large, positive effect on classification of infant-directed speech. See Supplementary Methods for further details.



Supplementary Fig. 3 | The main effects in the naïve listener experiment are not attributable to learning. a, This panel repeats the raw accuracy data reported in Extended Data Fig. 4b, but using only data from responses that were participants' first trial, to avoid the possibility of any learning effects over the course of their participation (with data available from $n = 1{,}035$ recordings). The results do not change appreciably. The points indicate average ratings for each recording; the gray lines connecting the points indicate the pairs of vocalizations produced by the same voice; the half-violins are kernel density estimations; and the boxplots represent the medians, interquartile ranges, and 95% confidence intervals (indicated by the notches). **b**, Over the course of multiple trials in the experiment, which contained corrective feedback, participants' raw accuracy barely increased. The lines depict linear regressions for each of the four vocalization types and the shaded regions depict 95% confidence intervals.



Supplementary Fig. 4 | Exploratory-confirmatory selected acoustic features for pre-registered analyses. The preregistered analyses included comparisons of the acoustic features of infant-directed vocalizations, regardless of whether they included speech or song. For the reasons discussed in the Methods, and per the results reported in Fig. 2, these results should be interpreted with caution, as direct comparisons of acoustic features across modalities (language vs. music) may be spurious or may hide underlying variation within each modality. Moreover, these analyses do not include fieldsite-level random effects, so they are less conservative than those reported in Fig. 2 (i.e., they identify a larger number of acoustic features). The boxplots show the 25 acoustic features with a significant difference in at least one main comparison (e.g., infant-directed song vs. infant-directed speech, in the right panel), in both the exploratory and confirmatory analyses. All variables are normalized across participants. The boxplots represent the median and interquartile range; the whiskers indicate 1.5 × IQR; and the notches represent the 95% confidence intervals of the medians. Faded comparisons did not reach significance in exploratory analyses. Significance values are computed via linear combinations using two-sided tests (n = 1,570 recordings); *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; no adjustments made for multiple-comparisons due to the exploratory-confirmatory approach taken. Prespecified hypotheses about each comparison are posted in the project GitHub repository.

Supplementary Tables

Label	Stub	Variables	Description	Significance
Attack Curve Slope	`mir_attack`	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR, Distance	MIRtoolbox detects acoustic events in the audio; for a subset of those it can compute an attack slope from amplitude curves, which is the slope of the line from the beginning of the event to its peak.	The slope of an attack curve provides a relative measure of "alerting components," or immediately discriminable beginnings of a vocalization.
Roughness	`mir_roughness`	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR, Distance	A roughness value produced by computing the peaks of the audio spectrum and taking the average of the dissonance between all possible pairs of peaks; following Buyens et al. (2017), we reduce this to a single measure by taking the RMS-normalized mean.	Along with inharmonicity, roughness provides one measure of dissonance in a recording. Roughness similarly provides at least one measure of vocal clarity.
85th Energy Percentile	`mir_rolloff85`	Whole	An estimate of the amount of high frequency in a signal measured by the frequency such that a 85% of the total energy is contained below it.	The 85th energy percentile allows a comparison of relative measures of high-frequency acoustics in a vocalization.
Inharmonicity	`mir_inharmonicity`	Whole	An estimate of the inharmonicity in the signal produced by identifying the number of partials that are not multiples of the fundamental frequency (i.e. those outside of the ideal harmonic range).	Along with roughness, inharmonicity provides a more precise measure of dissonance in a vocalization.
Tempo	`mir_tempo`	Whole	A tempo estimate made by detecting periodicities from MIR's event detection curves. Outputs a single number.	Tempo allows assessment of the speed or pace of a vocalization.
Pule Clarity	`mir_pulseclarity`	Whole	Estimates the rhythmic clarity, or strength of the beats (Lartillot et al. 2008).	Pulse clarity provides a measure of the vocal clarity of a speaker or emphasis on individual utterances.
Rhythmic Variability	`npvi_total`	Recording	The nPVI equation measures the "average degree of durational contrast between adjacent events in a sequence" (Daniele & Patel, 2015). This makes it especially useful for comparing rhythmic units across language and music (i.e., syllables vs. notes). To automatically detect events, we used Mertens' (2004) syllable detection algorithm.	By providing a measure of durational contrast, nPVI_total is a measure of rhythmic complexity in a recording.
Rhythmic Variability	`npvi_phrase`	Phrase	In addition to detecting syllables, Mertens' algorithm detects phrases. Whereas npvi_total computes nPVI based on the whole file as a continuous phrase, this measure computes the nPVI for each detected phrase and reports the mean. In other words, it excludes the distances between the ends and beginnings of phrases.	nPVI_phrase provides a more granular measure of rhythmic complexity, within phrases, rather than between them.

(continued)

Label	Stub	Variables	Description	Significance
Temporal Modulation	`tm_peak_hz`	Whole	The temporal modulation spectrum is the frequency decomposition of the amplitude envelope of a signal. This measures how loud something is at any given moment. We then measure how fast the loudness changes. For example: if someone sings a note every second, the spectrum will have a peak at 1Hz. If someone sings a note three times a second, but with an emphasis every three seconds, there will be a large peak at 1Hz, and a smaller peak at 3Hz. The peak of the spectrum is the frequency of the amplitude spectrum which has the highest root mean square of a given recording and represents a raw value of the recording's tempo.	The peak of the temporal modulation spectrum provides a measure of how maximally modulated, or variable, the onset of notes are in a recording, providing a raw measure of metre for speech and song.
Temporal Modulation	`tm_std_hz`	StD	The temporal modulation spectrum is the frequency decomposition of the amplitude envelope of a signal. This measures how loud something is at any given moment. We then measure how fast the loudness changes. For example: if someone sings a note every second, the spectrum will have a peak at 1Hz. If someone sings a note three times a second, but with an emphasis every three seconds, there will be a large peak at 1Hz, and a smaller peak at 3Hz. The standard deviation of the spectrum is taken as a measure of how exaggerated the peak is.	The standard deviation of temporal modulation allows for an assessment of the overall variability of temporal modulations in a recording, providing a coarse measure of rhythm, with a lower standard deviation leaning towards more monorhythmic signals.
Pitch	`praat_f0`	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	The fundamental frequency (f0) in Hertz for each recording	Pitch provides a fundamental measure of the highness or lowness, in frequency, of an utterance. Likewise, the shape of the pitch curve and the overall value of pitch is a common discriminable feature in both speech and song.
Pitch Space	`praat_f0travel`	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	The distance between f0 at each .03125/sec interval to the next.	Pitch space provides a dynamic measure of pitch's range over time.
Pitch Rate	`praat_pitch_rate`	Whole, Med, IQR	The pitch rate is a measure of pitch change over time. In essence, the pitch rate provides a measure of pitch curve smoothness (a lower value corresponds to a smoother curve).	The pitch rate provides a measure of how smooth or variable pitch is over time.
Vowel Space	`praat_vowtrav`	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	The Euclidian distance travelled in vowel space. This is equivalent to distance between the two formants.	Vowel space provides a measure of how much of the possible complex vowel space is used.
Vowel Space Travel Rate	`praat_vowtrav_rate`	Whole, Med, IQR	The Euclidian distance travelled in vowel space over a rate of time. This is equivalent to distance between two formants divided by rate of time.	Vowel travel rate provides a measure of how much of the vowel space is used over time, a relative measure of acoustic "flashiness" of a signal.

(continued)

Label	Stub	Variables	Description	Significance
Amplitude	`praat_intensity`	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR, Distance	A measure of amplitude (loudness) in decibels	Amplitude provides a measure of how loud or quiet a vocalization is and can be compared between types within speakers
Amplitude Space	`praat_intensitytravel`	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	The distance between amplitude at each .03125/sec interval to the next.	Intensity space provides a dynamic measure of intensity's range over time.
Amplitude Rate	`praat_intensity_rate`	Whole, Med, IQR	A measure of decay in intensity curves in each recording measured as change in amplitude over time.	The intensity rate provides a measure of how loud or soft amplitude changes over time.
1st Formant	`praat_f1`	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	The frequency in Hertz of the 1st formant at each (.03125/sec) point	1st formants are the 1st in a harmonic series following from the fundamental frequency and is important for a number of acoustic reasons.
Second Formant	`praat_f2`	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	The frequency in Hertz of the second formant at each (.03125/sec) point	Second formants are the second in a harmonic series following from the fundamental frequency, and along with the 1st formant, is used by listeners to perceive vowels.
File duration	`meta_length`		The length of the unedited sound files	
Concatenated file duration	`meta_edit_length`		The length of the concatenated versions of the sound files	

Supplementary Table 1. Codebook for acoustic features. Variable names are stubs; in the datasets on the project GitHub repository, suffixes are added to denote summary statistics (e.g., mir_attack_mean).

Speech		Song	
Acoustic feature	Coefficient	Acoustic feature	Coefficient
Speech		**Song**	
Pitch (Median)	2.449	Vowel Travel (IQR)	0.735
Vowel Travel Rate (Median)	0.677	Intensity (Median)	-0.428
Pitch (IQR)	0.533	Attack Curve Slope (Median)	-0.419
Pulse Clarity	0.231	Roughness (Median)	-0.405
Energy Roll-Off (85th %-ile)	-0.185	Second Formant (IQR)	-0.285
Second Formant (Median)	0.170	Energy Roll-Off (85th %-ile)	-0.255
Roughness (IQR)	-0.167	Inharmonicity	-0.171
Attack Curve Slope (Median)	0.152	Attack Curve Slope (IQR)	0.159
Attack Curve Slope (IQR)	0.119	Pitch (IQR)	-0.156
Inharmonicity	-0.073	Vowel Travel Rate (IQR)	0.117
Tempo	-0.057	Second Formant (Median)	-0.105
Intensity (IQR)	0.041	Tempo	0.080
		Pulse Clarity	0.079
		Peak Tempo	0.074
		Pitch (Median)	-0.042
		Rhythmic Variability (nPVI)	-0.028

Supplementary Table 2. The predictive influence of each of the acoustical features in distinguishing infant-directed from adult-directed vocalizations, chosen via two LASSO models (performance and the top six features for each model are depicted in Fig. 1b). The coefficients can be interpreted in a similar fashion to a logistic regression, i.e., as changes in the predicted log-odds ratio (with positive values indicating a higher likelihood of infant-directedness).

Comparison	Feature	Statistic	β	SE	z	p
ID Speech vs. AD Speech						
	Intensity	Median	0.081	0.052	1.542	0.123
	Acoustic Roughness	Median	-0.220	0.100	-2.202	0.028
		IQR	-0.124	0.071	-1.740	0.082
	Vowel Travel	IQR	0.283	0.126	2.236	0.025
	Pitch (F_0)	Median	0.641	0.101	6.341	<0.001
		IQR	0.602	0.128	4.692	<0.001
	Energy Roll-off (85 %ile)	Whole	-0.261	0.063	-4.129	<0.001
	Inharmonicity	Whole	-0.274	0.072	-3.802	<0.001
	Pulse Clarity	Whole	0.213	0.069	3.092	0.002
	Vowel Travel Rate	Median	0.514	0.116	4.412	<0.001
		IQR	0.519	0.123	4.234	<0.001
ID Song vs. AD Song						
	Intensity	Median	-0.138	0.048	-2.905	0.004
	Acoustic Roughness	Median	-0.227	0.097	-2.349	0.019
		IQR	-0.190	0.083	-2.295	0.022
	Vowel Travel	IQR	0.257	0.080	3.203	0.001
	Pitch (F_0)	Median	-0.052	0.062	-0.836	0.403
		IQR	-0.191	0.079	-2.414	0.016
	Energy Roll-off (85 %ile)	Whole	-0.025	0.074	-0.330	0.742
	Inharmonicity	Whole	-0.169	0.088	-1.923	0.055
	Pulse Clarity	Whole	0.064	0.111	0.579	0.562
	Vowel Travel Rate	Median	0.179	0.094	1.896	0.058
		IQR	0.211	0.088	2.396	0.017

Supplementary Table 3. Regression results from confirmatory analyses (corresponding with the boxplots in Fig. 2). The features tested here were limited to those with significant differences in the exploratory analyses, as such no adjustments for multiple comparisons were used. Statistics are from post-hoc linear combinations using two-sided tests following multi-level mixed-effects models. Abbreviations: infant-directed (ID); adult-directed (AD).

Song type	Number of songs
Love Song	21
Caring song	3
Sad Song	3
Ballad	2
Hanging out before bed song	1
Lullaby	1
Orphan song	1
Past remembrance song	1
Religious ballad	1
Song about island home	1

Supplementary Table 4. Adult-directed songs with descriptions rated as “soothing" by two independent annotators. A mixed-effects model estimating the difference in perceived infant-directedness across these vs. other adult-directed songs, adjusting for fieldsite-wise variability, found a statistically significant difference in responses ($b = -0.027, se = 0.006, t_{42,360} = -4.107, p < .0001$), but this difference was small (an estimated average difference of ~2.7% less infant-directed) and in the opposite direction to what one might expect if soothing songs were mistaken for lullabies.

	Speech			Song		
Fieldsite	d'	95% CI	n	d'	95% CI	n
Tannese Vanuatans	0.154	[-0.487 0.796]	2	0.070	[-0.250 0.390]	10
Mentawai Islanders	0.514	[-0.269 1.297]	6	0.140	[-0.227 0.507]	13
Tsimane	0.642	[-0.003 1.288]	11	0.233	[-0.096 0.563]	12
Sápara/Achuar	0.481	[-0.151 1.113]	10	0.259	[-0.071 0.588]	11
Quechuan/Aymaran	0.958	[0.285 1.632]	3	0.355	[0.011 0.699]	6
Enga	0.910	[0.214 1.605]	2	NA	NA	0
Mbendjele	0.894	[0.216 1.572]	3	0.417	[0.066 0.768]	10
Hadza	1.142	[0.433 1.851]	10	0.440	[0.097 0.783]	9
Nyangatom	1.092	[0.394 1.789]	5	0.453	[0.108 0.799]	7
Jenu Kurubas	1.290	[0.665 1.916]	10	0.515	[0.193 0.836]	11
Toposa	1.164	[0.488 1.839]	8	0.522	[0.180 0.865]	6
Krakow	1.308	[0.483 2.134]	7	0.529	[0.110 0.949]	7
Turku	1.489	[0.812 2.167]	16	0.536	[0.198 0.874]	14
Rural Poland	1.273	[0.704 1.842]	10	0.575	[0.274 0.876]	7
Colombian mestizos	1.325	[0.680 1.969]	5	0.605	[0.268 0.943]	7
San Diego	1.407	[0.674 2.141]	13	0.612	[0.241 0.982]	17
Beijing	1.613	[1.050 2.176]	26	0.706	[0.408 1.004]	28
Arawak	1.729	[1.067 2.392]	1	0.732	[0.391 1.073]	6
Afrocolombians	1.562	[0.815 2.309]	4	0.742	[0.369 1.115]	9
Toronto	1.593	[0.807 2.379]	27	0.747	[0.375 1.119]	23
Wellington	2.417	[1.730 3.104]	20	1.066	[0.720 1.413]	26

Supplementary Table 5. Estimated fieldsite-wise d-prime values, quantifying sensitivity to infant-directedness in speech and song, independent of response bias. Values are estimated as coefficients from mixed-effects model predicting d' from vocalization type, with random effects of fieldsite for each vocalization type. n refers to the number of vocalists that had a complete pair of vocalizations in the listener experiment (e.g., where one or both of the infant- and adult-directed vocalizations were not excluded due to confounds). Due to the strict exclusion procedure (see Methods), some fieldsites have very small samples, complicating the interpretation of these results, and one fieldsite had no observations for song. These exclusions only apply to the naïve listener experiment, however, and not the acoustic analyses reported elsewhere in this paper.

Characteristic	%	N
Gender		
Female	45.6%	7352
Male	51.5%	8299
Other	2.9%	463
[participant did not report]		14
Ethnicity		
American Indian/Alaska Native	1.4%	207
Asian	23.3%	3366
Black or African-American	3.7%	536
More than one race	9.4%	1351
Native Hawaiian or other Pacific Islander	0.9%	131
White	61.2%	8836
[participant did not report]		1701
Hispanic		
No	87.6%	12712
Yes	12.4%	1804
[participant did not report]		1612
Annual household income		
Under $10,000	9.1%	912
$10,000 to $19,999	8.8%	879
$20,000 to $29,999	7.4%	747
$30,000 to $39,999	7.5%	755
$40,000 to $49,999	7.4%	747
$50,000 to $74,999	14.7%	1471
$75,000 to $99,999	12.2%	1227
$100,000 to $150,000	17.9%	1795
Over $150,000	15.0%	1503
[participant did not report]		6092

Supplementary Table 6.
Demographics of United States participants. See notes and corresponding analyses in SI Text 1.5.

Principal Component 1		Principal Component 2		Principal Component 3	
Feature	Weighting	Feature	Weighting	Feature	Weighting
Amplitude Space (Mean)	-0.200	Amplitude (Mean)	0.271	Pitch (Mean)	-0.306
Amplitude Space Travel Rate (Median)	-0.200	Amplitude (Median)	0.266	Pitch (3rd Quartile)	-0.302
Pitch Space_rate (IQR)	-0.199	Amplitude (3rd Quartile)	0.264	Pitch (Median)	-0.291
Pitch Space Travel Rate (Whole)	-0.198	Amplitude (1st Quartile)	0.243	Pitch (1st Quartile)	-0.248
Amplitude Space Travel Rate (IQR)	-0.195	Roughness (3rd Quartile)	0.213	Pitch (IQR)	-0.223
Amplitude Space (Median)	-0.188	Roughness (IQR)	0.212	Roughness (1st Quartile)	0.215
Pitch Space (3rd Quartile)	-0.187	Roughness (Standard Deviation)	0.203	Roughness (Median)	0.194
Amplitude Space Travel Rate (Whole)	-0.187	Roughness (Median)	0.188	Pitch (Standard Deviation)	-0.178
Pitch Space (IQR)	-0.187	Amplitude (Maximum)	0.188	Roughness (3rd Quartile)	0.154
Amplitude Space (1st Quartile)	-0.185	Roughness (Range)	0.174	Roughness (IQR)	0.148
Amplitude Space (3rd Quartile)	-0.185	Roughness (Maximum)	0.174	Amplitude Space (Range)	-0.144
Vowel Space Travel Rate (Median)	-0.184	Amplitude (Minumum)	0.167	Amplitude Space (Maximum)	-0.144
Pitch Space (Mean)	-0.182	Roughness (Mean)	0.155	Roughness (Mean)	0.142
Vowel Space Travel Rate (IQR)	-0.180	1st Formant (1st Quartile)	0.147	Pitch (Maximum)	-0.131
Amplitude Space (IQR)	-0.179	Amplitude Space (Maximum)	0.136	Amplitude (3rd Quartile)	-0.124
Vowel Space Travel Rate (Whole)	-0.177	Amplitude Space (Range)	0.136	Pitch Space (1st Quartile)	-0.122
Pitch Space_rate (Median)	-0.176	Roughness (1st Quartile)	0.130	Second Formant (Minumum)	0.119
Vowel Space (Mean)	-0.170	1st Formant (Standard Deviation)	-0.129	Amplitude (Mean)	-0.118
Amplitude Space (Standard Deviation)	-0.161	Vowel Space (IQR)	-0.128	Amplitude (Median)	-0.116
Vowel Space (Median)	-0.161	Vowel Space (3rd Quartile)	-0.124	85th Energy Percentile	0.116
Vowel Space (Standard Deviation)	-0.159	Second Formant (Mean)	-0.123	Pitch Space (Maximum)	-0.114
Vowel Space (1st Quartile)	-0.158	1st Formant (Range)	-0.121	Pitch Space (Range)	-0.114
Vowel Space (3rd Quartile)	-0.152	1st Formant (Minumum)	0.121	Second Formant (IQR)	-0.111
Pitch Space (Standard Deviation)	-0.152	Second Formant (3rd Quartile)	-0.120	Amplitude (Maximum)	-0.110
Pitch Space (Median)	-0.150	Second Formant (Maximum)	-0.118	Amplitude (Range)	-0.110
Vowel Space (IQR)	-0.143	Second Formant (Median)	-0.117	Pitch (Range)	-0.107
Amplitude (IQR)	-0.127	Second Formant (Range)	-0.116	Second Formant (Standard Deviation)	-0.106
Temporal Modulation (Peak)	-0.107	1st Formant (Median)	0.114	Inharmonicity	0.104
nPVI Recording	0.100	Vowel Space (Mean)	-0.109	Amplitude (1st Quartile)	-0.103
Amplitude (Standard Deviation)	-0.099	1st Formant (Maximum)	-0.107	1st Formant (Mean)	0.101

Supplementary Table 7. Factor loadings for the top three principal components reported in Extended Data Fig. 2.

Supplementary references

1. Lartillot, O., Toiviainen, P. & Eerola, T. A Matlab toolbox for music information retrieval. in *Data analysis, machine learning and applications* (eds. Preisach, C., Burkhardt, H., Schmidt-Thieme, L. & Decker, R.) 261–268 (Springer Berlin Heidelberg, 2008).
2. Buyens, W., Moonen, M., Wouters, J. & van Dijk, B. A model for music complexity applied to music preprocessing for cochlear implants. in *2017 25th European Signal Processing Conference (EUSIPCO)* 971–975 (IEEE, 2017).
3. Ding, N. *et al.* Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews* **81**, (2017).
4. Patel, A. D. Musical rhythm, linguistic rhythm, and human evolution. *Music Perception* **24**, 99–104 (2006).
5. Mertens, P. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. in *Speech Prosody 2004, International Conference* (2004).
6. Yale, C. & Forsythe, A. B. Winsorized regression. *Technometrics* **18**, 291–300 (1976).
7. Hautus, M. J., Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide*. (Routledge, 2022).
8. Snodgrass, J. G. & Corwin, J. Pragmatics of Measuring Recognition Memory: Applications to Dementia and Amnesia. *Journal of Experiment Psychology: General* **117**, 34–50 (1988).
9. Sheskin, M. *et al.* Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences* (2020). doi:10.1016/j.tics.2020.06.004
10. Hartshorne, J. K., de Leeuw, J., Goodman, N., Jennings, M. & O'Donnell, T. J. A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods* **51**, 1782–1803 (2019).
11. Fernald, A. & Simon, T. Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* **20**, 104–113 (1984).
12. Trehub, S. E. *et al.* Mothers' and fathers' singing to infants. *Developmental Psychology* **33**, 500–507 (1997).