bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acoustic regularities in infant-directed speech and song across cultures

Courtney B. Hilton[1,^], Cody J. Moser[1,2,^], Mila Bertolo[1], Harry Lee-Rubin[1], Dorsa Amir[3], Constance M. Bainbridge[1,4], Jan Simson[1,5], Dean Knox[6], Luke Glowacki[7], Andrzej Galbarczyk[8], Grazyna Jasienska[8], Cody T. Ross[9], Mary Beth Neff[10,11], Alia Martin[10], Laura K. Cirelli[12,13], Sandra E. Trehub[13], Jinqi Song[14], Minju Kim[15], Adena Schachner[15], Tom A. Vardy[16], Quentin D. Atkinson[16,17], Jan Antfolk[18], Purnima Madhivanan[19,20,21,22], Anand Siddaiah[22], Caitlyn D. Placek[23], Gul Deniz Salali[24], Sarai Keestra[24], Manvir Singh[25,26], Scott A. Collins[27], John Q. Patton[28], Camila Scaff[29], Jonathan Stieglitz[26,30], Cristina Moya[31,32], Rohan R. Sagar[33], Brian M. Wood[34], Max M. Krasnow[1,35] & Samuel A. Mehr[1,10,36,*]

[1]Department of Psychology, Harvard University, Cambridge, MA 02138, USA.
[2]Department of Cognitive and Information Sciences, University of California Merced, Merced, CA 95343, USA.
[3]Boston College Department of Psychology, Chestnut Hill, MA 02467, USA.
[4]Department of Communication, University of California Los Angeles, Los Angeles, CA 90095, USA.
[5]Department of Psychology, University of Amsterdam, 1012 WX Amsterdam, The Netherlands.
[6]Department of Politics, Princeton University, Princeton, NJ 08544, USA.
[7]Department of Anthropology, Boston University, Boston, MA 02215, USA.
[8]Department of Environmental Health, Faculty of Health Sciences, Jagiellonian University Medical College, 31-066 Krakow, Poland.
[9]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany.
[10]School of Psychology, Victoria University of Wellington, Wellington 6012, New Zealand.
[11]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo 0315, Norway.
[12]Department of Psychology, University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada.
[13]Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada.
[14]Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095, USA.
[15]Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA.
[16]School of Psychology, University of Auckland, Auckland 1010, New Zealand.
[17]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, D-07745 Jena, Germany.
[18]Department of Psychology, Åbo Akademi, 20500 Turku, Finland.
[19]Department of Health Promotion Sciences, College of Public Health, University of Arizona, Tucson, AZ 85724, USA.
[20]Department of Medicine, Division of Infectious Diseases, College of Medicine, University of Arizona, Tucson, AZ 85724, USA.
[21]Department of Family & Community Medicine, College of Medicine, University of Arizona, Tucson, AZ 85724, USA.
[22]Public Health Research Institute of India, Mysuru 570020, India.
[23]Department of Anthropology, Ball State University, Muncie, IN 47306, USA.
[24]Department of Anthropology, University College London, WC1H 0BW London, UK.
[25]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
[26]Institute for Advanced Study in Toulouse, 31080 Toulouse Cedex 6, France.
[27]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85281, USA.
[28]Division of Anthropology, California State University, Fullerton, CA 92831, USA.
[29]Institute of Evolutionary Medicine, University of Zurich, 8006 Zürich, Switzerland.
[30]Université Toulouse 1 Capitole, 31080 Toulouse Cedex 6, France.
[31]Department of Anthropology, University of California, Davis, Davis, CA 95616, USA.
[32]Centre for Culture & Evolution, Brunel University London, UB8 3PH Uxbridge, UK.
[33]Future Generations University, Circle Ville, WV 26807, USA.
[34]Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[35]Division of Continuing Education, Harvard University, Cambridge, MA 02138, USA.
[36]Data Science Initiative, Harvard University, Cambridge, MA 02138, USA.

[^]These authors contributed equally and are listed alphabetically.

*Corresponding author. E-mail: sam@wjh.harvard.edu

Abstract

Across taxa, the forms of vocal signals are shaped by their functions[1–15]. In humans, a salient context of vocal signaling is infant care, as human infants are altricial[16,17]. Humans often produce "parent-ese", speech and song for infants that differ acoustically from ordinary speech and song[18–35], in fashions that are thought to support parent-infant communication and infant language learning[36–39]; modulate infant affect[33,40–45]; or credibly signal information to infants[46]. These theories predict a universal form-function link in infant-directed vocalizations, with consistent differentiation between infant-directed and adult-directed vocalizations across cultures. Some evidence supports this prediction[23,27,28,32,47–50], but the limited generalizability of individual ethnographic reports and laboratory experiments[51] and small stimulus sets[52], along with intriguing reports of counterexamples[53–60], leave the question open. Here, we show that infant-directed speech and song are robustly differentiable from their adult-directed counterparts, within voices and across cultures. We built a corpus of 1615 recordings of infant- and adult-directed singing and speech produced by 410 people living in 21 urban, rural, and small-scale societies and played the recordings to 45,745 people recruited online from many countries. We asked them to guess whether or not each vocalization was, in fact, infant-directed. The patterns of inferences of these naïve listeners, supported by acoustic analyses and predictive modelling, demonstrate acoustic cues to infant-directedness that are cross-culturally robust. The cues to infant-directedness differ across language and music, however, informing hypotheses of the psychological functions and evolution of both.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main

The forms of many animal signals are shaped by their functions, a link arising from production- and reception-related rules that help to maintain reliable signal detection within and across species[1–6]. Form-function links are widespread in vocal signals across taxa, from meerkats to fish[3,7–10], causing acoustic regularities that allow cross-species intelligibility[11–13,15]. This facilitates the ability of some species to eavesdrop on the vocalizations of other species, for example, as in superb fairywrens (*Malurus cyaneus*), who learn to flee predatory birds in response to alarm calls that they themselves do not produce[14].

In humans, an important context for the effective transmission of vocal signals is between parents and infants, as human infants are particularly helpless[16]. To elicit care, infants use a distinctive alarm signal: they cry[17]. In response, adults produce infant-directed language and music (sometimes referred to as "parent-ese") in forms of speech and song with putatively stereotyped acoustics[18–35].

These stereotyped acoustics are thought to be functional: supporting language acquisition[36–39], modulating infant affect and temperament[33,40,41], and signalling information to infants[46]. These theories all share a key prediction: like the vocal signals of other species, the forms of infant-directed vocalizations should be universally shaped by their functions, instantiated with clear regularities across cultures. Evidence for a universal form-function link is mixed, however, given the limited generalizability of individual ethnographic reports and laboratory studies[51]; small stimulus sets[52]; and a variety of counterexamples[53,54,56–60].

In language, infant-directed speech is primarily characterized by higher and more variable pitch[61] and more exaggerated and variable vowels[23,62,63], in modern industrialized societies[23,28,47,48,50,64,65] and a few small-scale societies[49,66]. Infants are themselves sensitive to these features, preferring them, even if spoken in unfamiliar languages[67–69]. But these acoustic features are *less* exaggerated in some cultures[58,64,70] and apparently vary in relation to the age and sex of the infant[64,71,72].

In music, infant-directed songs also have stereotyped acoustic features. Lullabies, for example, tend toward slower tempos, reduced accentuation, and simple repetitive melodic patterns[31,32,35,73], supporting functional roles associated with infant care[33,41,46] in industrialized[34,74–76] and small-scale societies[77,78]. Infants are soothed by these acoustic features, whether produced in familiar[44,45] or unfamiliar songs[79], and both adults and children reliably associate the same features with a soothing function[31,32,73]. But cross-cultural studies of infant-directed song have primarily relied upon archival recordings from disparate sources[29,31,32]; an approach that poorly controls for differences in voices, behavioral contexts, recording equipment, and historical conventions.

The degree to which infant-directed vocalizations are acoustically stereotyped across cultures is therefore unclear. To address this, we created a corpus of infant-directed song, infant-directed speech, adult-directed song, and adult-directed speech from a diverse set of 21 human societies, totaling 1615 field recordings of 410 individual voices (Fig. 1a, Table 1, and Methods; the corpus is open-access at https://doi.org/10.5281/zenodo.5525161). Participants were asked to provide all four vocalization types, enabling within-voice analyses.

Here, we report analyses of the corpus, using computational methods and a citizen-science experiment, to study three questions: (i) Is infant-directedness mutually intelligible across cultures? (ii) Are the acoustic cues to infant-directedness cross-culturally robust? (iii) Are human inferences about infant-directedness aligned to such acoustic cues?

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

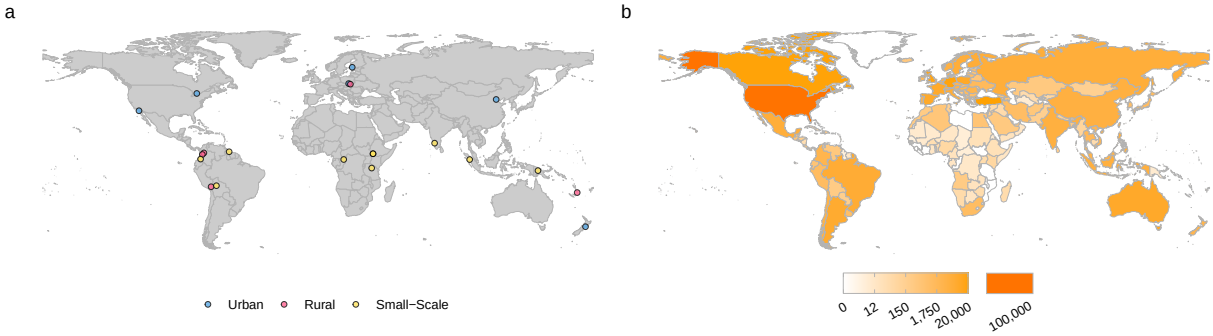


Fig. 1 | Where vocalizations were recorded and where citizen scientists were recruited. a, We recorded examples of speech and song from 21 urban, rural, or small-scale societies, whose locations are represented by coloured circles. **b**, Participants in the citizen-science experiment, who listened to the vocalizations and guessed whether each was directed toward an adult or to an infant, hailed from many countries; the gradients indicate the total number of vocalization ratings gathered from each country.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Region	Sub-Region	Society	Language	Language family	Subsistence type	Population	Distance to city (km)	Children per family	Recordings
Africa	Central Africa	Mbendjele BaYaka	Mbendjele	Niger-Congo	Hunter-Gatherer	61-152	120	7	60
	Eastern Africa	Hadza	Hadza	Hadza	Hunter-Gatherer	35	80	6	38
		Nyangatom	Nyangatom	Nilotic	Pastoralist	155	180	5.6	56
		Toposa	Toposa	Nilotic	Pastoralist	250	180	5.2	60
Asia	East Asia	Beijing	Mandarin	Sino-Tibetan	Urban	21.5M	0	1	124
	South Asia	Jenu Kurubas	Kannada	Dravidian	Other	2000	15	1	80
	Southeast Asia	Mentawai Islanders	Mentawai	Austronesian	Horticulturalist	260	120	.	60
Europe	Eastern Europe	Krakow	Polish	Indo-European	Urban	771,069	0	1.54	44
		Rural Poland	Polish	Indo-European	Agriculturalists	6,720	70	1.83	55
	Scandinavia	Turku	Finnish & Swedish	Uralic and Indo-European	Urban	186,000	0	1.41	80
North America	North America	San Diego	English (USA)	Indo-European	Urban	3.3M	0	1.7	116
		Toronto	English (Canadian)	Indo-European	Urban	5.9M	0	1.5	198
Oceania	Melanesia	Ni-Vanuatu	Bislama	Indo-European Creole	Horticulturalist	6,000	224	3.78	90
		Enga	Enga	Trans-New Guinea	Horticulturalist	500	120	6	22
	Polynesia	Wellington	English (New Zealand)	Indo-European	Urban	210,400	0	1.45	228
South America	Amazonia	Arawak	English Creole	Indo-European	Other	350	32	3	48
		Tsimane	Tsimane	Moseten-Tsimane	Horticulturalist	150	234	9	51
		Sapara & Achuar	Quechua & Achuar	Quechuan & Jivaroan	Horticulturalist	200	205	9	59
	Central Andes	Quechua/Aymara	Spanish	Indo-European	Agro-Pastoralist	200	8	4	49
	Northwestern South America	Afrocolombians	Spanish	Indo-European	Horticulturalist	300-1,000	100	6.6	53
		Colombian Mestizos	Spanish	Indo-European	Commercial Economy	470,000	0	3.5	43

Table 1. Societies from which recordings were gathered.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Naïve listeners distinguish infant-directed from adult-directed vocalizations

We played excerpts from the vocalization corpus to 45,745 people in the "Who's Listening?" game on https://themusiclab.org (after exclusions; see Methods). The participants resided in 184 countries and reported speaking 164 native languages. We asked them to judge, quickly, whether each vocalization was directed to a baby or to an adult (see Methods and Extended Data Fig. 1). We only included recordings that lacked confounding contextual/background cues (e.g., an audible infant; see Methods). Unless noted otherwise, all estimates reported here are generated by mixed-effects linear regression, adjusting for fieldsite (as a random effect), and with p-values generated via linear combination tests.

Corpus-wide, infant-directed speech was far more likely to be rated as infant-directed than was adult-directed speech from the same voice (Fig. 2a; ID speech = 51%, AD speech = 22%; $\chi^2(1) = 25.3$, p $< .0001$); and infant-directed song was far more likely to be rated as infant-directed than was adult-directed song from the same voice (Fig. 2a; ID song = 72%, AD song = 57%; $\chi^2(1) = 13.58$, p $< .001$). These results were robust to learning effects, as they repeated when only analyzing each participant's first exposure to a vocalization in the experiment and listener accuracy increased by only 0.06% after each trial (Extended Data Fig. 2). They were also robust to post-hoc data trimming decisions, such as excluding recordings with confounding background noise and/or trials where the listener could likely understand the words in the vocalization (Extended Data Fig. 3).

There was, however, an overall bias toward "baby" responses for songs (67% of all responses were "baby", but only 51% of songs were infant-directed) and toward "adult" responses for speech (64% "adult" responses vs. 56% actually adult-directed), however, which led adult-directed songs to be reliably *mis*-identified as infant-directed. To quantify sensitivity to infant-directedness independently from this bias, we ran a d-prime analysis at the level of each vocalist, i.e., analyzing participants' ability to identify infant-directedness within each voice after correcting for response bias. Sensitivity was significantly higher than the chance level of 0 (speech: $d' = 1.05$, 95% CI [0.64, 1.46]; song: $d' = 0.42$, 95% CI [0.22, 0.62]; $ps < .0001$) implying that the naïve listeners reliably differentiated between infant- and adult-directed vocalizations across both speech and song, and with ~2.5 times higher sensitivity in speech.

We also analyzed performance in the task within the subset of recordings drawn from each fieldsite. Cross-site variability was evident, especially in the size of effects (but less so in their direction); we caution that some fieldsites had small samples, making it impossible to know whether such effects represent true cross-cultural variability, sampling variability, or both. In 20 of 21 fieldsites, mean "baby" ratings were higher for infant-directed speech than adult-directed speech (Fig. 2b) and in 17 of 21 fieldsites, mean "baby" ratings were higher for infant-directed song than adult-directed song (Fig. 2b). In all fieldsites that failed to replicate the overall pattern in song, however, the mean "baby" rating for infant-directedness was nonetheless above the chance level of 50%. Fieldsite-wise d' scores are reported in Extended Data Table 1.

Listener sensitivity within each fieldsite was also correlated with a number of society-level characteristics: rank-order population size (speech: $\tau = 0.53$; song: $\tau = 0.6$), distance from fieldsite to nearest urban center (speech: r = -0.75; song: r = -0.49), and number of children per family (speech: r = -0.57; song: r = -0.8; all $ps < .001$). 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 ($ps > .4$).

Tests of cross-cultural variability among *listeners* also revealed strong similarity in the perception of infant-directedness. On trials where the vocalization being judged was in a closely related language to the native language of the listener (e.g., when the vocalization was in Spanish and the listener's native language was English, which are both Indo-European languages), performance increased only modestly relative to trials where the language family did not match (e.g., when the vocalization was in Mentawai, an Austronesian language, and the listener's native language was Mandarin, a Sino-Tibetan language); the effect was statistically significant but small (difference in $d' = 0.18$, $p = 0.01$; Extended Data Fig. 4). Linguistic relatedness therefore only accounted for a small amount of variability in naïve listeners' intuitions of infant-directedness. More generally, random effects of listener country, gender, and age on sensitivity were all small (each varying by $< 1\%$), implying cross-demographic consistency in listener intuitions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

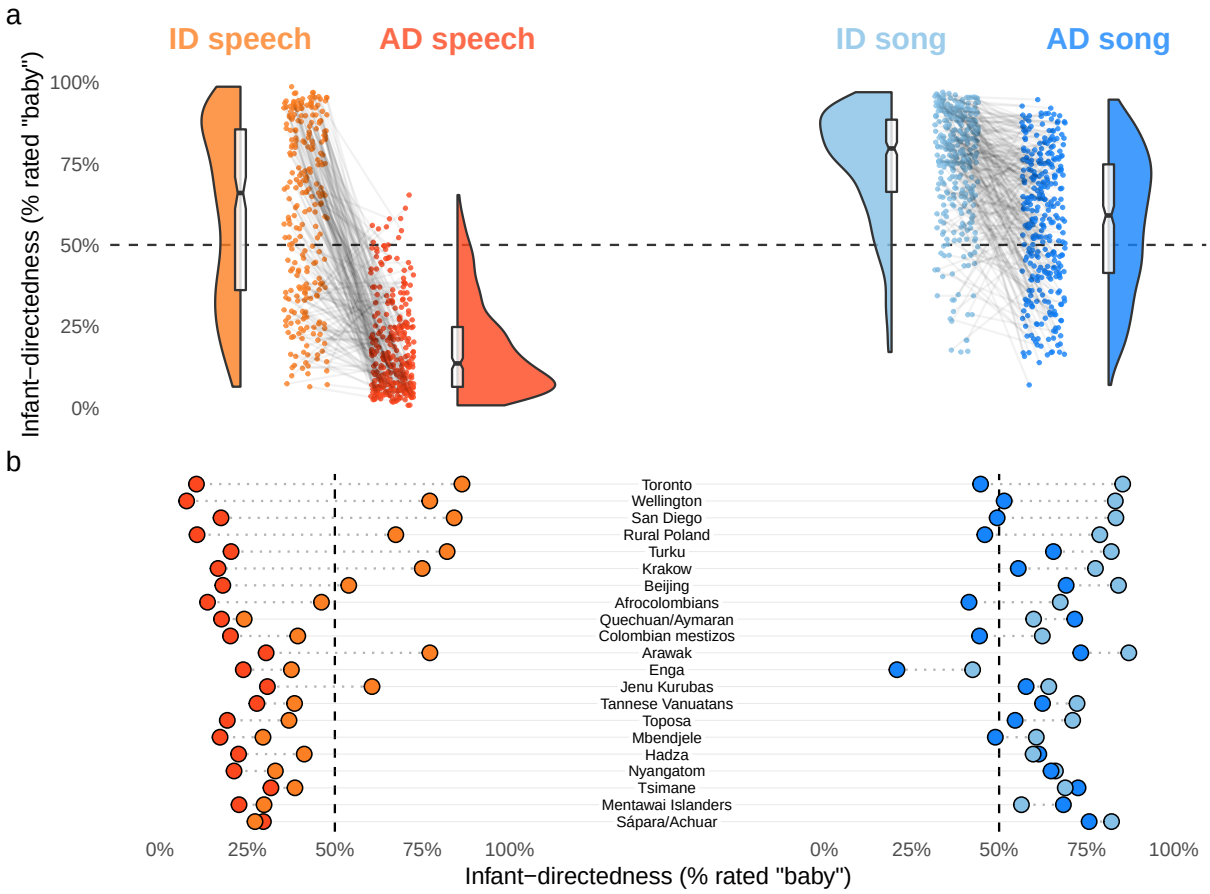


Fig. 2 | Naïve listeners distinguish infant-directed vocalizations from adult-directed vocalizations across cultures. **a**, Participants listened to vocalizations drawn at random from the corpus and responded to the prompt "Someone is speaking or singing. Who do you think they are singing or speaking to?" with either "adult" or "baby". In almost all cases, infant-directed (ID) vocalizations were more frequently rated as "baby" than were adult-directed (AD) vocalizations produced by the same voice. This was true of both speech and song. The points indicate averages for each recording (from a median of 447 ratings per recording; IQR = 152-497; minimum = 96); the gray lines connecting the points indicate the pairs of vocalizations produced by the same voice; the half-violins are kernel density estimations; the boxplots represent the medians, interquartile ranges, and 95% confidence intervals (indicated by the notches); and the horizontal dotted line represents the expected accuracy under random guesses, 50%. Surprisingly, there was an overall bias toward inferring infant-directedness in music, wherein *both* types of song were more likely than chance to be rated as "infant-directed". Trials where the listener's native language is the same as the language of the vocalization are excluded, among others; this figure also includes a small amount of supplementary data via a follow-up experiment (see Methods). **b**, The results replicated in most fieldsites, although with varying effect sizes; this may result from sampling variability, true variability in the differences between infant- and adult-directed vocalizations across fieldsites, or both. The circles depict each fieldsite's mean % infant-directed rating across each of the four vocalization types; with the same color-coding as **a**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acoustic correlates of infant-directedness across cultures

What enables such a diverse group of people to arrive at such similar conclusions about unfamiliar, foreign vocalizations, in languages that they do not understand? One possibility is that there exists a universal set of acoustic features driving listeners' inferences concerning the intended targets of speech and song, which are reliably instantiated within and across societies, as suggested by functional accounts of infant-directed vocalization[33,36–43,46].

To test this possibility, we studied 15 types of acoustic features in each recording (e.g., pitch, rhythm, timbre) via multiple variables (e.g., median, interquartile range); these were treated to reduce the influence of atypical observations (e.g., extreme values caused by loud wind, rain, and other background noises), and standardized within-voices to eliminate between-voice variability. This yielded a total of 99 variables (see Methods; a codebook is in Extended Data Table 2).

Following a preregistered exploratory-confirmatory design, we fitted a multi-level mixed-effects regression predicting each acoustic variable from the vocalization types, after adjusting for voice and fieldsite as random effects, and using linear combinations to test for infant-directedness differences in song and speech separately. To reduce the risk of Type I error, we performed this analysis on a randomly selected half of the corpus (exploratory, weighting by fieldsite) and only report results that successfully replicated in the other half (confirmatory). We did not correct for multiple tests because the exploratory-confirmatory design restricts the tests to those with a directional prediction.*

This procedure identified 16 acoustic features that distinguished infant-directedness in song, speech, or both (Fig. 3; statistics are in Extended Data Table 3), in the context of producing infant-directed vocalizations "when baby is fussy". For example, across cultures and within voices, infant-directed speech had considerably higher pitch, greater pitch range, and more contrasting vowels than adult-directed speech. These results repeated consistently in each fieldsite: pitch, energy-rolloff, and inharmonicity showed the same direction of difference in all 21 fieldsites; and other features, such as vowel contrasts and attack curve slopes, were consistent in the majority of them (see the doughnut plots in Fig. 3a). These patterns align with prior claims of pitch and vowel-contrast being robust features of infant-directed speech[23,65], and substantiate them across many cultures.

The distinguishing features of infant-directed song were more subtle, however, but nevertheless corroborate its purported soothing functions[33,41,46]: reduced loudness, intensity, and acoustic attack; reduced pitch range; and purer-sounding vocal qualities (reduced roughness and inharmonicity), which were mostly consistent across sites. The smaller effects in song, relative to speech, may result from the fact that while solo-voice speaking is fairly natural and representative of most adult-directed speech (i.e., people rarely speak at the same time), much of the world's song occurs in social groups where there are multiple singers and accompanying instruments[32,46,80]. Asking participants to produce solo adult-directed song may have biased participants toward choosing more soothing and intimate songs (e.g., ballads, love songs; see Extended Data Table 4), or less naturalistic renditions of songs that would normally be sung in less constrained social contexts. Further, the adult-directed songs were produced in the presence of an infant, which can in principle alter participants' singing style[35] (although this may comparably alter the adult-directed speech examples; see Methods for one test of this question). Thus the distinctiveness of infant-directed song (relative to adult-directed song) may be underestimated in these data.

Some acoustic correlates of infant-directedness had very different trends across language and music. For example, whereas median pitch strongly differentiated infant-directed speech from adult-directed speech, it had no such effect in music; pitch variability had the *opposite* effect across language and music; and similar patterns were evident in first and second formants. Loudness-related features showed a similar

*We note one important deviation from the preregistration: we originally planned post-hoc linear combinations to test hypothesized differences between (1) infant-directed and adult-directed vocalizations overall; (2) infant-directed song and adult-directed song; and (3) infant-directed song and infant-directed speech. We retain the second comparison in the main text, but no longer focus on (1) or (3) as the analysis approach is confounded by the fact that acoustic differences between speech and song overall far outstrip the acoustic correlates of infant-directedness. Instead, we adopted the simpler and more informative approach of post-hoc comparisons that are only within speech and within song. We also retained the exploratory-confirmatory design, as it mitigates the potential for inflated Type I errors. For transparency, we still report the preregistered post-hoc tests in Extended Data Fig. 5, but suggest that these comparisons be interpreted with caution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pattern, where intensity and attack slope were increased in infant-directed speech and decreased in infant-directed song, on average, and relative to their adult-directed counterparts. That some basic acoustic features operate differently across infant-directed speech and song supports the possibility of differentiated functional roles[18,33,34,45,46,79,81].

But some acoustic features were nevertheless common to both language and music; in particular, overall, infant-directedness was characterized by reduced roughness and inharmonicity, which may facilitate parent-infant signalling[5,41] through better contrast with the sounds of screaming and crying[17,82]; and increased vowel contrasts, potentially to aid language acquisition[36,37,39] or as a byproduct of socio-emotional signalling[1,63].

Last, we conducted an exploratory principal components analysis of the full 99 features (Fig. 3b; the analysis accounts for ~40% of total variability in acoustic features). The results provide convergent evidence that the main forms of acoustic variation partition into orthogonal clusters distinguishing (PC1) speech from song overall; (PC2) infant-directedness in song; and (PC1 and PC3) infant-directedness in speech. Factor loadings are in Extended Data Table 5; they largely replicate the findings of the 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).

Human intuitions of infant-directedness are modulated by vocalization acoustics

Last, we assessed whether these acoustic features alone are sufficient to replicate human performance in classifying infant-directedness. To do this, we trained two least absolute shrinkage and selection operator (LASSO) classifiers[83] with fieldsite-wise leave-one-out cross-validation, separately for speech and song recordings. This approach[32] gives a strong test of the cross-cultural consistency of acoustical correlates of infant-directedness, as the model's classification accuracy is evaluated on held-out data from a fieldsite that it has not been trained on.

Both models performed significantly above the 50% chance level (Fig. 4a; speech: 77% correct, 95% CI [71%, 83%]; song: 65% correct, 95% CI [59%, 71%]). When accounting for response bias, model performance was highly similar to the aggregate guessing patterns of human listeners, as evaluated via a receiver operating characteristic analysis (Extended Data Fig. 6), for both speech (human AUC: 90.77, 95% CI [88.41, 93.14]; model AUC: 92.13, 95% CI [90.33, 93.93]) and song (human AUC: 75.52, 95% CI [71.7, 79.33]; model AUC: 77.37, 95% CI [74.14, 80.6]). Using this same bias-free metric, both models also performed similarly to humans at the level of each individual fieldsite (speech: $r = 0.38$, $p = 0.04$; song: $r = 0.56$, $p = 0.004$; see Fig. 4a and Extended Data Fig. 7). These results demonstrate that the measured acoustic correlates of infant-directedness operate reliably across the 21 societies studied, at least with sufficient consistency to replicate the overall level of human classification performance.

We then examined the precise relations between acoustic features and the experiment-wide *proportions* of infant-directedness ratings for each vocalization, in a similar approach to prior research[73]. The proportions are a more strenuous target to predict than a binary classification (as in the first two LASSO models) in that they form a continuous measure of infant-directedness per the ears of the naïve listeners. We trained two further LASSO models to predict the proportions, using the same cross-validation procedure. Both models explained considerable variation in human listeners' intuitions (Fig. 4b; speech $R^2 = 0.56$; song $R^2 = 0.21$, ps $< .0001$), albeit more so in speech than in song.

We also measured the relations between the influence of each acoustic cue on human intuitions and the effect sizes of each variable in the corpus-wide acoustical analyses. If human inferences are attuned to some universal profile of acoustic correlates of infant-directedness, one might expect a close relationship between the strength of *actual* acoustic differences between vocalizations on a given feature and the relative influence of that feature on human intuitions. We compared the variable importance scores from the LASSO model predicting human inferences (visualized in the bar plots in Fig. 4a) to a measure of how acoustically salient each feature was (estimated as mean differences in the corpus; Fig. 3). We found a significant positive relationship for speech ($r = 0.82$, $p < .001$) but not for song ($r = 0.32$, $p = 0.14$), implying that human

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

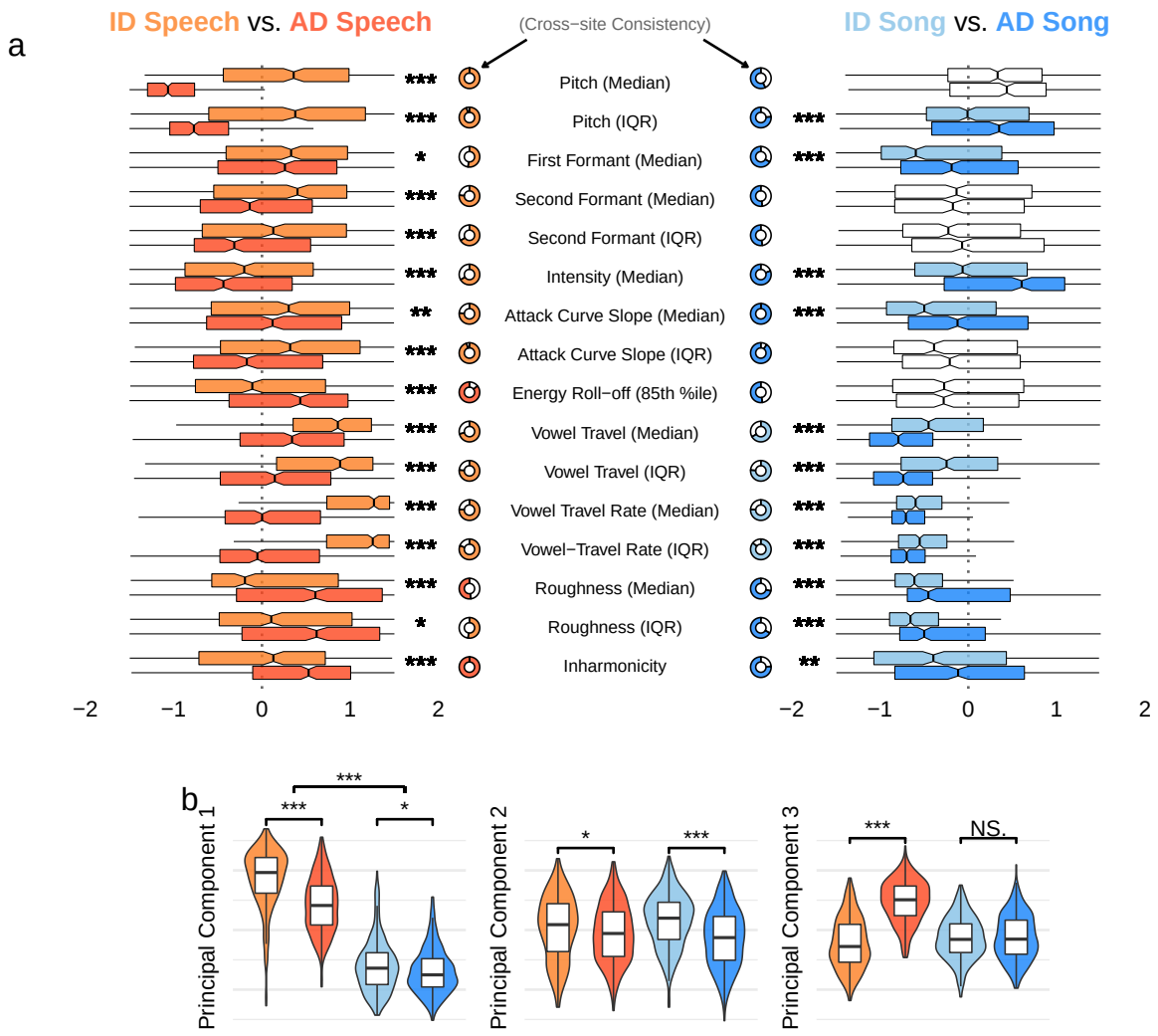


Fig. 3 | Acoustic correlates of infant-directedness in speech and song. **a**, Acoustic analyses revealed 16 features with a statistically significant difference between infant-directed and adult-directed vocalizations in speech, song, or both. These features operated differently across speech and song. For example, median pitch was far higher in infant-directed speech than in adult-directed speech in all 21 fieldsites, whereas median pitch was comparable across both forms of song. The boxplots, which are ordered approximately from largest differences between speech and song to smallest, represent each acoustic feature's median (vertical black lines) and interquartile range (boxes); the whiskers indicate 1.5 × IQR; the notches represent the 95% confidence intervals of the medians; and the doughnut plots represent the proportion of fieldsites where the main effect repeated, based on estimates of fieldsite-wise random effects. All acoustic features were normalized within-voices; only comparisons that survived an exploratory-confirmatory analysis procedure are plotted; faded comparisons did not reach significance in confirmatory analyses. Significance values are computed via linear combinations, following multi-level mixed-effects models; $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Statistics are in Extended Data Table 3. **b**, A principal components analysis on the full 99 acoustic variables independently supports the idea that the acoustic features operate differently in language and music: the first principal component most strongly distinguishes speech from song, overall; the second distinguishes infant-directed from adult-directed *song*; and the third distinguishes infant-directed from adult-directed *speech*. The violins indicate kernel density estimations and the boxplots represent the medians and interquartile ranges. Feature loadings are in Extended Data Table 4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

intuitions concerning infant-directed song were likely driven by more subjective features of the recordings, higher-level acoustic features that we did not measure, or both; this contrasts with intuitions concerning infant-directed speech, which were largely explicable from simple, objective acoustic features.

Discussion

We provide convergent evidence for cross-cultural regularities in the acoustic design of infant-directed speech and song. Naïve listeners reliably identified infant-directed vocalizations as infant-directed, despite the fact that the vocalizations were of largely unfamiliar cultural, geographic, and linguistic origin; acoustic analyses showed cross-culturally reliable acoustic differentiation of infant-directed and adult-directed vocalizations, in both speech and song; and these acoustic distinctions explained substantial variability in human intuitions concerning infant-directedness.

Thus, despite evident variability in language, music, and infant care practices worldwide, when people speak or sing to fussy infants, they modify the acoustic features of their vocalizations in similar and mutually intelligible ways across cultures. This implies that the forms of infant-directed vocalizations are shaped by their functions, in a fashion similar to the vocal signals of many non-human species.

By analyzing both speech and song recorded from the same voices, we discerned precise differences in the ways infant-directedness is instantiated in language and music. In response to the same prompt of addressing a "fussy infant", infant-directedness in speech and song was instantiated with opposite trends in acoustic modification (relative to adult-directed speech and song): infant-directed speech was more intense and contrasting (e.g., more pitch variability, higher intensity) while infant-directed song was more subdued and soothing (e.g., less pitch variability, lower intensity). These acoustic dissociations suggest functional dissociations, with speech being more attention-grabbing, the better to distract from baby's fussiness[37,38]; and song being more soothing, the better to lower baby's arousal[32,33,41–43,45,79]. Speech and song are both capable of playful or soothing roles[60] but each here tended toward one acoustic profile over the other, despite both types of vocalization being elicited here in the *same* context: vocalizations used "when the baby is fussy".

Many of the reported acoustic differences are consistent with the bioacoustics of vocal signalling in non-human animals[1–15]. For example, in both speech and song, infant-directedness was robustly associated with purer and less harsh vocal timbres, and greater formant-frequency dispersion. In non-human animals, these features have convergently evolved across taxa in the functional context of signalling friendliness or approachability in close contact calls[1,3,63,84], in contrast to alarm calls or signals of aggression, which are associated with rough sounds that have less formant dispersal[4,85–87]. The use of these features in infant care may originate from signalling approachability to baby, but may have later acquired further functions more specific to the human context. For example, greater formant-frequency dispersion accentuates vowel contrasts, which could facilitate language acquisition[36,63,88–90]; and purer vocal timbre may facilitate communication by contrasting conspicuously with the acoustic context of infant cries[5] (for readers unfamiliar with infants, this context is acoustically harsh[17,82]).

Higher pitch is also routinely a cue for animal vocal signalling of approachability and friendliness; accordingly, one of the largest and most robust results in our study was that infant-directedness raised the vocal pitch (f0) of speech to a songlike level. But infant-directedness had no effect on pitch within song. This curious asymmetry is consistent with the idea that pitched aspects of music may originate from elaborations to generic infant-directed vocalizations, where both use less harsh but more variable pitch patterns and more temporally variable and expansive vowel spaces to provide infants with ostensible "flashy" signals of attention and pro-social friendliness[41,46,61,91,92]. This does not mean that pitch alterations are *absent* from infant-directed song (indeed, in one study, mothers sang a song at higher pitch when producing a more playful rendition, and a lower pitch when producing a more soothing rendition[44]), but on average, both infant- and adult-directed song, along with infant-directed speech, tend to be higher in pitch than adult-directed speech.

We leave open at least two further questions. First, the results are suggestive of universality, because the corpus covers a swath of geographic locations (21 societies on 6 continents), languages (12 language families), and different subsistence regimes (8 types) (see Table 1). But these do not constitute a representative sample

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

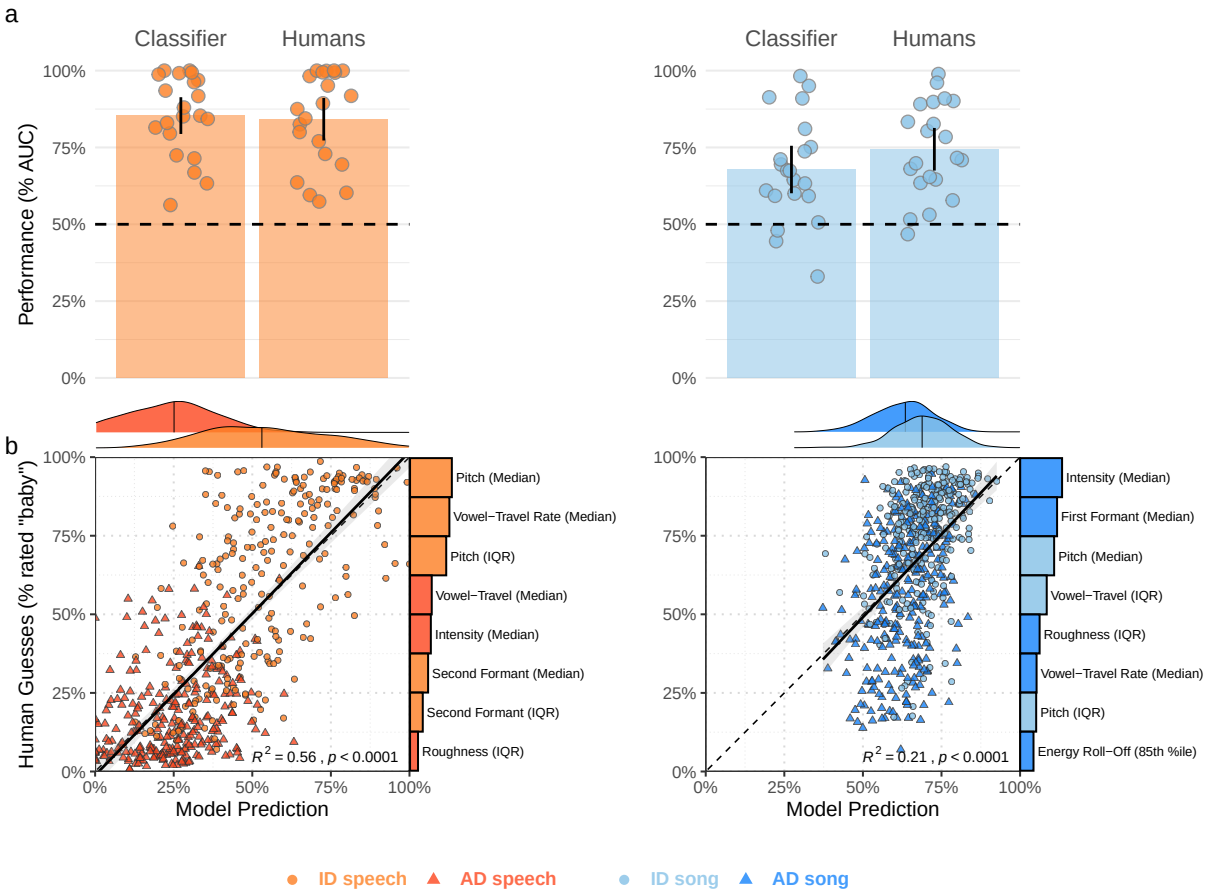


Fig. 4 | Human inferences about infant-directedness are predictable from acoustic features of vocalizations. **a**, We trained two Least absolute shrinkage and selection operator (LASSO) models, one for speech and one for song, to classify whether recordings were infant-directed or adult-directed on the basis of the 16 acoustic features identified by our exploratory-confirmatory analysis. These predictors were then regularized using cross-validation across fieldsites. The bars represent the overall classification performance averaged across fieldsites (quantified via receiver operating characteristic/area under the curve; AUC) for both the classifier and for the humans in the naïve-listener experiment; the error bars represent 95% confidence intervals; and the points represent the average performance for each fieldsite. **b**, Two further LASSO models were trained using the same procedure, predicting the *percentage* of "baby" responses for each recording from the human listeners. Each point represents a recorded vocalization, plotted in terms of the model's estimated infant-directedness of the model and the average "infant-directed" rating from the naïve listeners; the barplots depict the relative explanatory power of the top 8 acoustical features in each LASSO model, showing which features were most strongly associated with human inferences (the colors indicate whether each feature was higher in value for ID or AD); the dotted diagonal lines represent a hypothetical perfect match between model predictions and human guesses; the solid black lines depict linear regressions; the grey ribbons represent the standard errors of the mean, from the regressions; and the shaded regions represent kernel density estimations of the distribution of model estimates for the vocalization types plotted in each panel (with vertical black lines depicting the medians).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of humans, so strong claims of universality are not justified; indeed, we found both cross-cultural consistency and variability (e.g., with the fieldsite in Wellington, New Zealand demonstrating main effects an order of magnitude larger than some other fieldsites). In addition to studying more representative samples of infant-directed vocalizations; other future approaches may (i) use phylogenetic methods to examine whether people in societies that are distantly related nonetheless produce similar infant-directed vocalizations; (ii) test perceived infant-directedness in more diverse samples of listeners, to more accurately characterize cross-cultural variability in the *perception* of infant-directedness; and (iii) test listener intuitions among groups with reduced exposure to a given set of infant-directed vocalizations, such as very young infants or people from isolated, distantly related societies, as in related efforts[27,67,93]. Such research would benefit in particular from a focus on societies previously reported to have unusual vocalization practices, infant care practices, or both[53,56–58]; and would also clarify the extent to which convergent practices across cultures are due to cultural borrowing (in the many cases where societies are not fully isolated from the influence of global media).

Second, speech and song are used in a multiple contexts with infants, of which "addressing a fussy infant" (the type of vocalization we elicited from participants) is just one[18,34]. One curious finding may bear on this question: naïve listeners displayed a bias toward "adult" guesses for speech and "baby" guesses for song, regardless of their actual targets. This suggests that listeners treated "adult" and "baby" as the default reference levels for speech and song, respectively, against which acoustic evidence was compared, a pattern consistent with theories that posit song as having a special connection to infant care in human psychology[33,46].

Methods

Vocalization corpus

We built a corpus of 1,615 recordings of infant-directed song, infant-directed speech, adult-directed song, and adult-directed speech (all audio is available at https://doi.org/10.5281/zenodo.5525161). Participants (N = 411) living in 21 societies (Fig. 1a and Table 1) produced each of these vocalizations, respectively, with a median of 15 participants per society (range: 6-57). From those participants for whom information was available, most were female (86%) and nearly all were parents or grandparents of the focal infant (95%).

Recordings were collected by principal investigators and/or staff at their field sites, all using the same data collection protocol. They translated instructions to the native language of the participants, following the standard research practices at each site. There was no procedure for screening out participants, but we encouraged our collaborators to collect data from parents rather than non-parents. Fieldsites were selected partly by convenience (i.e., via recruiting principal investigators at fieldsites with access to infants and caregivers) and partly to maximize cross-fieldsite diversity (see Table 1).

For infant-directed song and infant-directed speech, participants were asked to sing and speak to their infant as if they were fussy, where "fussy" could refer to anything from frowning or mild whimpering to a full tantrum. At no fieldsites were difficulties reported in the translation of the English word "fussy", suggesting that participants understood it. For adult-directed speech, participants spoke to the researcher about a topic of their choice (e.g., they described their daily routine). For adult-directed song, participants sang a song that was not intended for infants; they also stated what that song was intended for (e.g., "a celebration song"). The record collection protocol is posted at https://github.com/themusiclab/infant-speech-song.

For most participants (90%) an infant was physically present during the recording (the infants were 48% female; age in months: M = 11.4; SD = 7.61; range 0.5-48). When an infant was not present, participants were asked to imagine that they were vocalizing to their own infant or grandchild, and simulated their infant-directed vocalizations. Prior research has shown that simulated infant-directedness is qualitatively similar, albeit less exaggerated than when authentic, for both speech[94] and song[35]. Indeed, a model of the naïve listener results adjusting for fieldsite indeed showed a small decrease in "baby" guesses when an infant

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was not present (ID song: 7.1%, ID speech: 8.4%, AD song: 6.5%, AD speech: 4.3%, $ps < .0001$), and this effect was stronger for vocalizations that were infant-directed than adult-directed ($\chi^2(1) = 5.67$, p = 0.02). Both the naive listener results and acoustic analyses were robust to whether these simulated infant-directed vocalizations were included or excluded, however.

In all cases, participants were free to determine the content of their vocalizations. This was intentional: imposing a specific content category on their vocalizations (e.g., "sing a *lullaby*") would likely alter the acoustic features of their vocalizations, which are known to be influenced by experimental contexts[95]. Some participants produced adult-directed songs that shared features with the intended soothing nature of the infant-directed songs; data on the intended behavioral context of each adult-directed song are in Extended Data Table 4.

All recordings were made with Zoom H2n digital field recorders, using foam windscreens (where available). To ensure that participants were audible along with researchers (who stated information about the participant and environment before and after the vocalizations), recordings were made with a 360° dual x-y microphone pattern. This produced two uncompressed stereo audio files (WAV) per participant at 44.1 kHz; we only analyzed audio from the two-channel file on which the participant was loudest.

The principal investigator at each fieldsite also provided standardized background data on the behavior and cultural practices of the society (e.g., whether there was access to mobile-phones/TV/radio, and how commonly people used ID speech or song in their daily lives). Most items were based on variables included in the D-PLACE cross-cultural corpus[96]. Complete data are posted on the project GitHub repository.

The 21 societies varied widely in their characteristics, from cities with millions of residents (Beijing) to small-scale hunter-gatherer groups of as few as 35 people (Hadza). All of the small-scale societies studied had limited access to TV, radio, and the internet, mitigating against the influence of exposure to the music and/or infant care practices of other societies. Four of the small-scale societies (Nyangatom, Toposa, Sápara/Achuar, and Mbendjele) were completely without access to these communication technologies.

The societies also varied in the prevalence of infant-directed speech and song in day-to-day life. The only site reported to lack infant-directed song in contemporary practice was the Quechuan/Aymaran site, although it was also noted that people from this site know infant-directed songs in Spanish and use other vocalizations to calm infants. Conversely, the Mbendjele BaYaka were noted to use infant-directed song, but rarely used infant-directed speech. In most sites, the frequency of infant-directed song and speech varied. For example, among the Tsimane, song is reportedly infrequent in the context of infant care; when it appears, however, it is specifically used to soothe and encourage infants to sleep.

Naïve listener experiment

We analyzed all data available at the time of writing this manuscript from the "Who's Listening?" game at https://themusiclab.org/quizzes/ids, a continuously running jsPsych[97] experiment distributed via Pushkin[98]. A total of 63,481 participants began the experiment, the first in January 2019 and the last in October 2021.

We played participants vocalizations from a subset of the corpus, excluding those that were less than 10 seconds in duration (n = 113) and those with confounding sounds that were not produced by the target voice in the first 5 seconds of the recording (e.g., a crying baby or laughing adult in the background; n = 364), as determined by two independent annotators who remained unaware of vocalization type and fieldsite (with disagreements resolved by discussion). We also excluded trials where the native language of the listener matched the language of the vocalization (N = 85,968 of 709,628 trials, or 12.1%), as this could enable listeners to infer whether a vocalization was infant-directed independently of the vocalization's acoustic characteristics. Robustness checks confirmed that the data trimming decisions did not substantially alter the results (Extended Data Fig. 3). Irrespective of the recordings each participant was assigned, we also excluded participants who reported having previously participated in the same experiment (n = 3,514); participants who reported being younger than 12 years old (n = 1,340); and those who reported having a hearing impairment (n = 1,201).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This yielded a sample of 45,745 participants (gender: 20,664 female, 24,126 male, 922 other, 33 did not disclose; age: median 22 years, interquartile range 18-29). Participants self-reported living in 184 different countries (Fig. 1b) and speaking 164 different native languages; roughly half the participants were native English speakers from the United States.

Participants listened to at least 1 and at most 16 vocalizations drawn from the subset of the corpus (as they were free to leave the experiment before completing it) for a total of 388,985 ratings (Fig. 1b; infant-directed song: n = 109,994; infant-directed speech: n = 77,317; adult-directed song: n = 104,023; adult-directed speech: n = 97,651). The vocalizations were selected with weighted randomization, such that a set of 16 trials included 4 vocalizations in English and 12 in other languages; roughly half the corpus was English-language vocalizations, so this method ensured that participants heard a substantial number of vocalizations in other languages. This yielded over 46 ratings per vocalization (median = 447; interquartile range 151-496.75) and thousands of ratings for each society (median = 18,631; interquartile range: 12,100-21,393).

We asked participants to classify each vocalization as either directed toward a baby or an adult (Extended Data Fig. 1), as quickly as possible, either by pressing a key corresponding to a drawing of an infant or adult face (when the participant used a desktop computer) or by tapping one of the faces (when the participant used a tablet or smartphone). The locations of the faces (left vs. right on a desktop; top vs. bottom on a tablet or smartphone) were randomized participant-wise. As soon as they made a choice, playback stopped. After each trial, we told participants whether or not they had answered correctly and how long, in seconds, they took to respond; at the end of the experiment, we gave participants a total score and percentile rank (relative to other participants).

In revising this manuscript, we discovered that a small subset of the corpus had been erroneously excluded from the main 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 (at US$15/hr) rather than volunteering; and each participant rated 188 instead of up to 16. In addition to the erroneously excluded recordings, we included in the Prolific experiment 85 additional recordings randomly selected from those that *were* included in the citizen-science experiment, ensuring that each Prolific participant heard an exactly balanced set of vocalization types. The two cohorts' ratings of the recordings in common across the two experiments were highly correlated ($r = 0.95$, $p < .0001$), 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.

Acoustic feature extraction

We manually extracted the longest continuous and uninterrupted section of audio from each of the four samples per participant (i.e., isolating vocalizations by the participant from interruptions from other speakers, the infant, and so on), using Adobe Audition. We then used the silence detection tool in Praat[99], with minimum sounding intervals at 0.1 seconds and minimum silent intervals at 0.3 seconds, to remove all portions of the audio where the participant was not speaking (i.e., the silence between vocalization phrases). These were manually concatenated in Python, producing denoised recordings, which were subsequently checked manually to ensure minimal loss of content.

We extracted and subsequently analyzed acoustic features using Praat[99], MIRtoolbox[100], temporal modularity using discrete Fourier transforms for rhythmic variability[101], and normalized pairwise variability indices[102]. These features consisted of measurements of pitch (e.g., F_0, the fundamental frequency), timbre (e.g., roughness), and rhythm (e.g., tempo); all summarized over time: producing 99 variables in total. We standardized feature values within-voices, eliminating between-voice variability. In the main acoustic analyses (Fig. 3a), we restricted the variable set to 26 summary statistics of median and interquartile range,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as these correlated highly with other summary statistics (e.g., maximum, range) but were less sensitive to extreme observations. The principal components analysis (Fig. 3b) used the full variable set of 99 variables.

Praat

We extracted intensity, pitch, and first and second formant values from the denoised recordings every 0.03125 seconds. For male participants, the pitch floor was set at 75 Hz, with a pitch ceiling at 300 Hz, and a maximum formant of 5000 Hz. For female participants, these values were 100 Hz, 600 Hz, and 5500 Hz, respectively. From these data, several summary values were calculated per 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[100]. *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[103]. 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 Ding's[104] method, 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.

Normalized pairwise variability index

The nPVI represents the temporal variance of data with discrete events, which makes it especially useful for comparing speech and music[101]. We used an automated syllable- and phrase-detection algorithm to extract events[102]. 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 both temporal modulation and nPVI measures, we used recordings before they had been denoised.

Outlier preprocessing

Because automated acoustic analyses are highly sensitive at extremes (e.g., impossible values caused by non-vocal sounds, like loud wind), we Winsorized all variables. This process arbitrarily defines outliers as being those exceeding the lowest and highest 5 percentile ranks, recoding them as precisely the values of those percentile boundaries. These data were used for all acoustic analyses. This decision had no impact on the interpretation of results, but is preferable to trimming extreme values[105]; pilot analyses using an alternate method, imputing extreme values with the mean observation for each feature within each fieldsite, yielded comparable results.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

End notes

Data, code, and materials availability

A fully reproducible manuscript; data, analysis code, and visualizations; other materials; and code for the naïve listener experiment are available at https://github.com/themusiclab/infant-speech-song. The audio corpus is available at https://doi.org/10.5281/zenodo.5525161. The preregistration for the auditory analyses is at https://osf.io/5r72u. Readers may participate in the naïve listener experiment by visiting https://themusiclab.org/quizzes/ids.

Acknowledgments

This research was supported by the Harvard University Department of Psychology (M.M.K. and S.A.M.); the Harvard College Research Program (H.L-R.); the Harvard Data Science Initiative (S.A.M.); the National Institutes of Health Director's Early Independence Award DP5OD024566 (S.A.M. and C.B.H.); the Academy of Finland Grant 298513 (J.A.); the Royal Society of New Zealand Te Apārangi Rutherford Discovery Fellowship RDF-UOA1101 (Q.D.A., T.A.V.); the Social Sciences and Humanities Research Council of Canada (L.K.C.); the Polish Ministry of Science and Higher Education grant N43/DBS/000068 (G.J.); the Fogarty International Center (P.M., A.S., C.D.P.); the National Heart, Lung, and Blood Institute, and the National Institute of Neurological Disorders and Stroke Award D43 TW010540 (P.M., A.S.); the National Institute of Allergy and Infectious Diseases Award R15-AI128714-01 (P.M.); the Max Planck Institute for Evolutionary Anthropology (C.T.R., C.M.); a British Academy Research Fellowship and Grant SRG-171409 (G.D.S.); the Institute for Advanced Study in Toulouse, under an Agence nationale de la recherche grant, Investissements d'Avenir ANR-17-EURE-0010 (L.G., J. Stieglitz); the Fondation Pierre Mercier pour la Science (C.S.); and the Natural Sciences and Engineering Research Council of Canada (S.E.T.). We thank the participants and their families for providing recordings; L. Sugiyama, for supporting pilot data collection; J. Du, E. Pillsworth, P. Wiessner, and J. Ziker, who collected or attempted to collect additional recordings; S. Atwood, A. Bergson, Z. Jurewicz, D. Li, L. Lopez, E. Radytė, and S. Ccari Cutipa for research assistance; and J. Kominsky, L. Powell, and L. Yurdum for feedback on the manuscript.

Author contributions

- S.A.M. and M.M.K. conceived of the research, provided funding, and coordinated the recruitment of collaborators and creation of the corpus.
- S.A.M. and M.M.K. designed the protocol for collecting vocalization recordings with input from D.A., who piloted it in the field.
- L.G., A.G., G.J., C.T.R., M.B.N., A.M., L.K.C., S.E.T., J. Song, M.K., A.S., T.A.V., Q.D.A., J.A., P.M., A.S., C.D.P., G.D.S., S.K., M.S., S.A.C., J.Q.P., C.S., J. Stieglitz, C.M., R.R.S., and B.M.W collected the field recordings.
- S.A.M., C.M.B., and J. Simson designed and implemented the online experiment.
- C.J.M. and H.L-R. processed all recordings and designed the acoustic feature extraction with S.A.M. and M.M.K.; C.M.B. provided associated research assistance.
- C.M. designed the fieldsite questionnaire with assistance from M.B. and C.J.M., who collected the data from the principal investigators.
- C.B.H. and S.A.M. led analyses, with additional contributions from C.J.M., M.B., and D.K., and M.M.K.
- C.B.H. and S.A.M. designed the figures.
- C.B.H. wrote computer code, with contributions from S.A.M., C.J.M., and M.B.
- C.J.M., H.L-R., M.M.K., and S.A.M. wrote the initial manuscript.
- C.B.H. and S.A.M. wrote the revision, with contributions from C.J.M. and M.B., and all authors approved it.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ethics

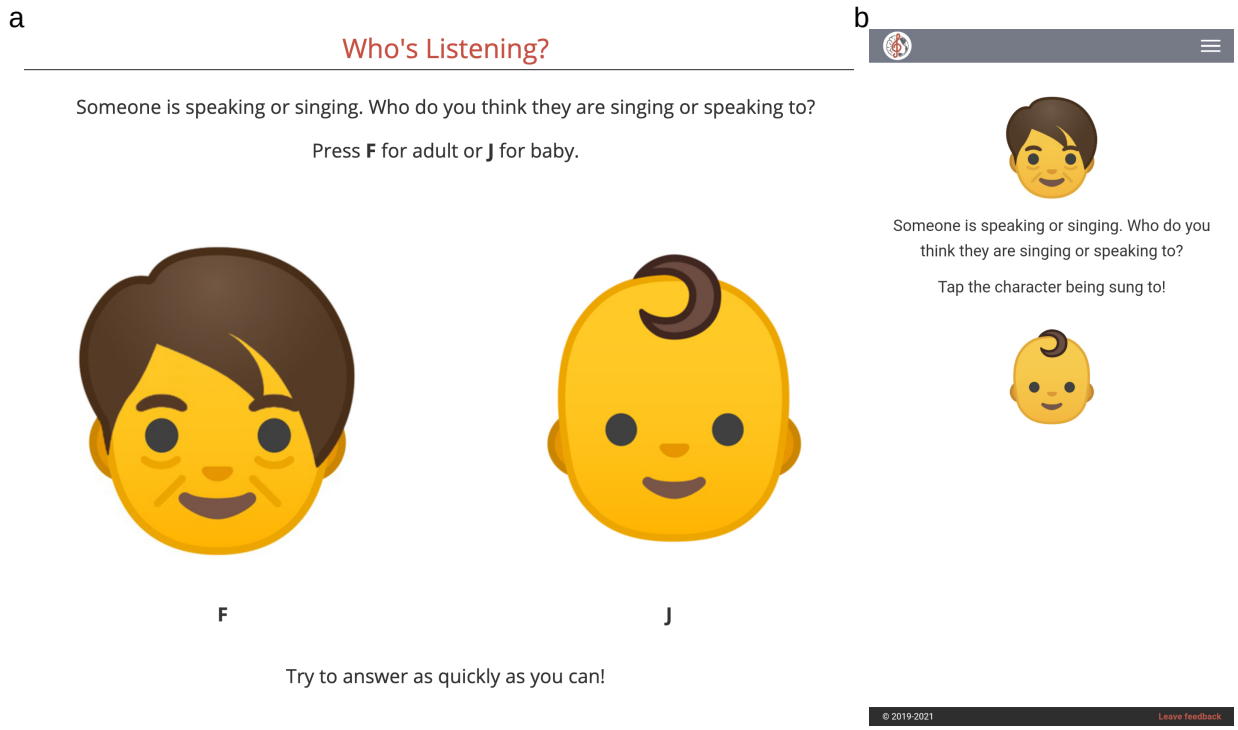
Informed consent was obtained from all participants. Ethics approval for the naïve listener experiment was provided by the Committee on the Use of Human Subjects, Harvard University's Insitutional Review Board (protocol #IRB17-1206). Ethics approval for the collection of recordings and their use in research was decentralized; each collaborating research arranged ethics approval with their local institution.

Additional information

The authors declare no competing interests.
Supplementary information is available for this paper.
Correspondence and requests for materials should be addressed to S.A.M.

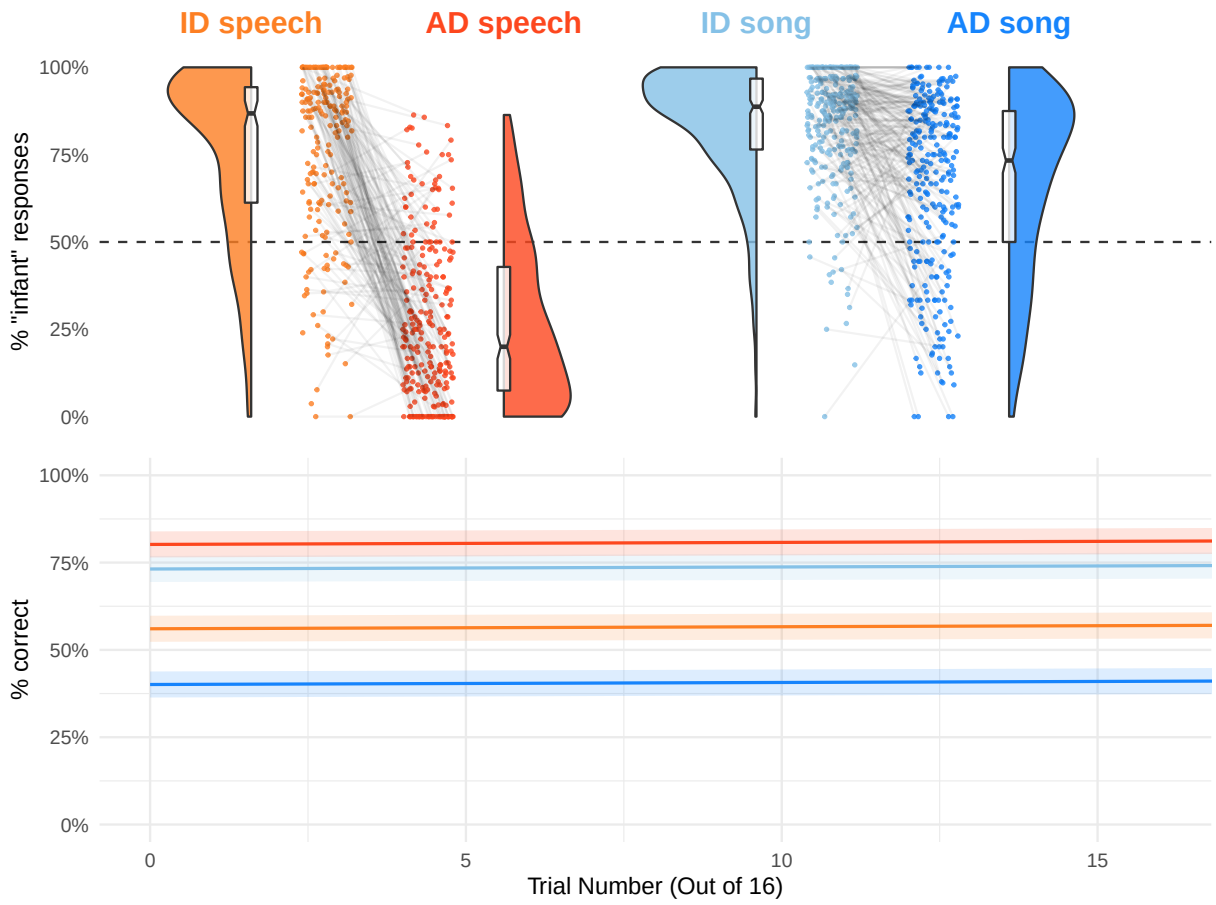
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Information



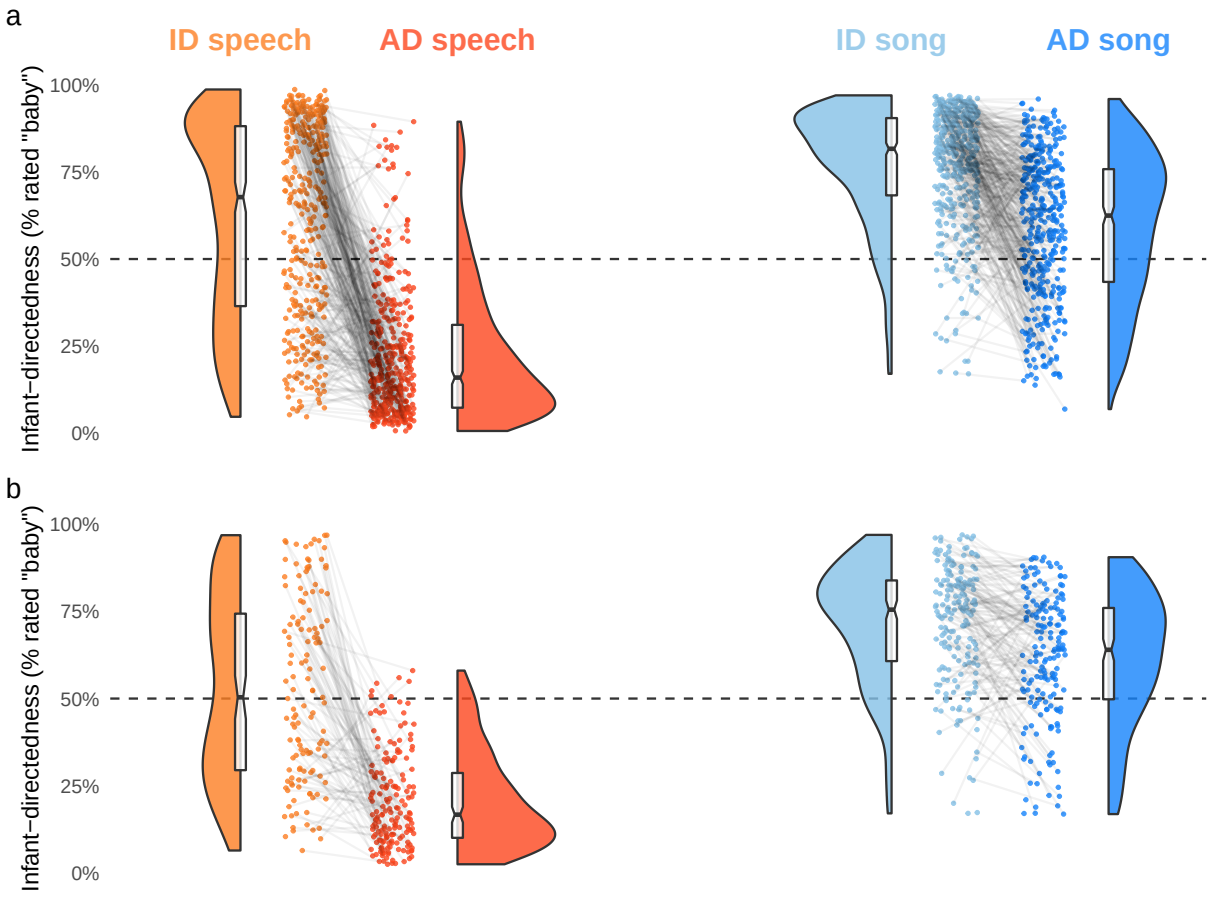
Extended Data Fig. 1 | Screenshots from the naïve listener experiment. On each trial, participants heard a randomly selected vocalization from the corpus and were asked to quickly guess to whom the vocalization was directed: an adult or a baby. The experiment used large emoji and was designed to display comparably on desktop computers (**a**) or tablets/smartphones (**b**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



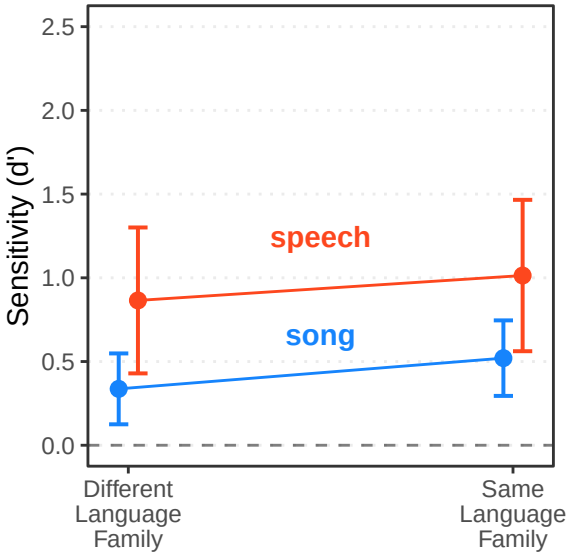
Extended Data Fig. 2 | The main effects in the naïve listener experiment are not attributable to learning. a, This panel repeats Fig. 2a, but only uses data from each participant's first trial, to avoid the possibility of any learning effects over the course of their participation. See further details in the caption to Fig. 2a. **b**, Over the course of multiple trials in the experiment, which contained corrective feedback, participants' accuracy was estimated to increase by only 0.06% per trial ($p < .001$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



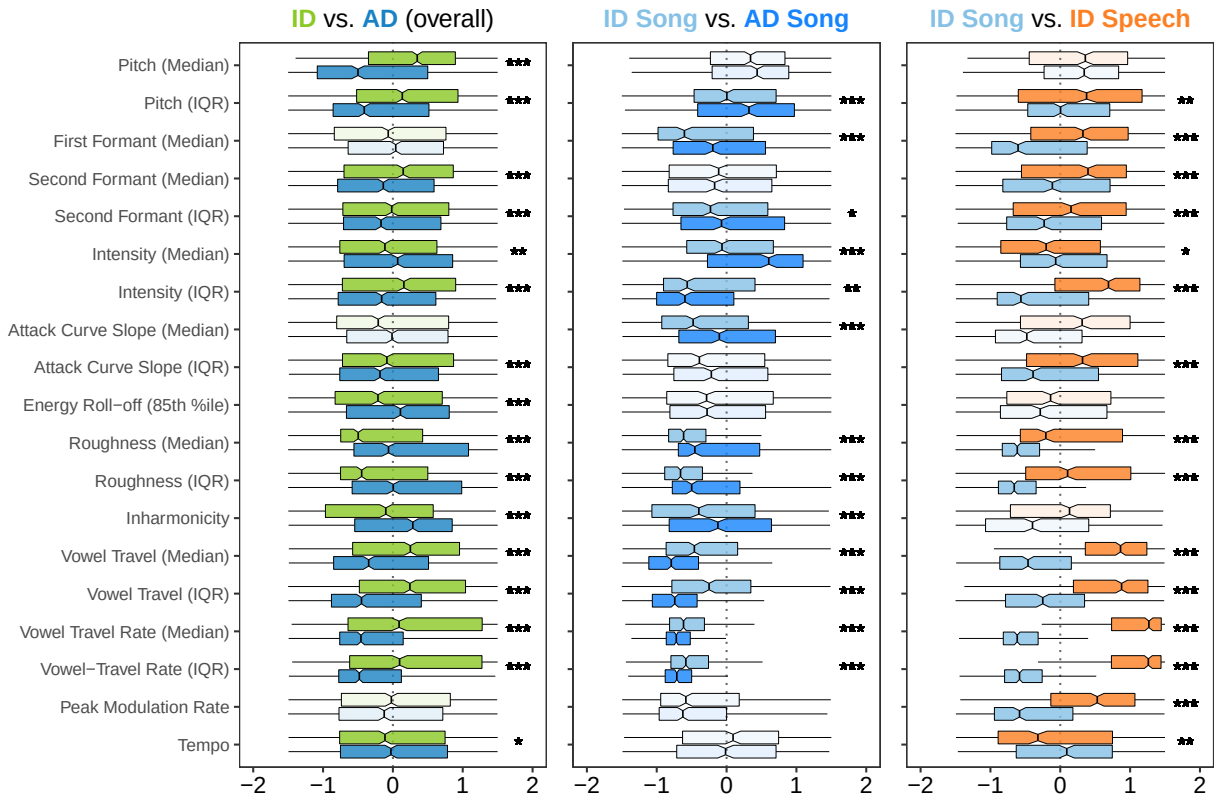
Extended Data Fig. 3 | The main effects in the naïve listener experiment are robust to alternative exclusion criteria. a, This panel repeats Fig. 2a, but including analysis of *all* recordings, even those that have audible confounds (e.g., a crying infant). **b**, This panel again repeats Fig. 2a, but excluding all English-language recordings (i.e., mostly from the Wellington, Toronto and San Diego sites). In both cases, the main effects repeat. See further details in the caption to Fig. 2a.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



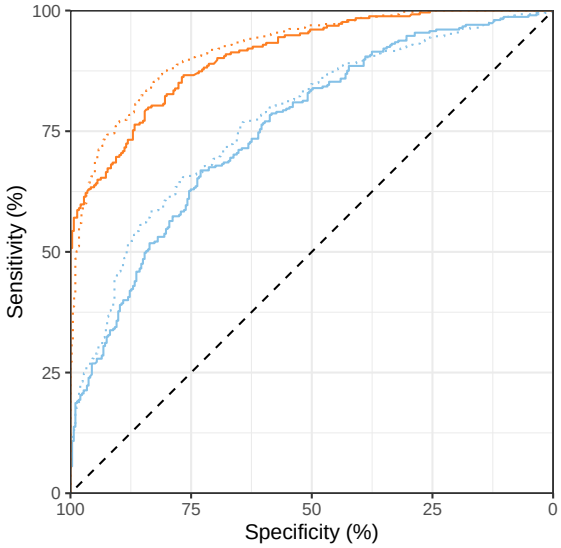
Extended Data Fig. 4 | Relation between listener language and vocalist language is minimally predictive of sensitivity to infant-directedness. We split all trials from the main experiment into two groups: those where the native language of the listener was in the same Glottolog language family as the language of the vocalization excerpt (but was not the same language; n = 270,221), and those that were not (n = 110,565). The plot shows the estimated marginal effects of a mixed-effects model predicting *d*-prime values across language and music examples, after adjusting for fieldsite. Relatedness had only a modest effect on identification accuracy.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



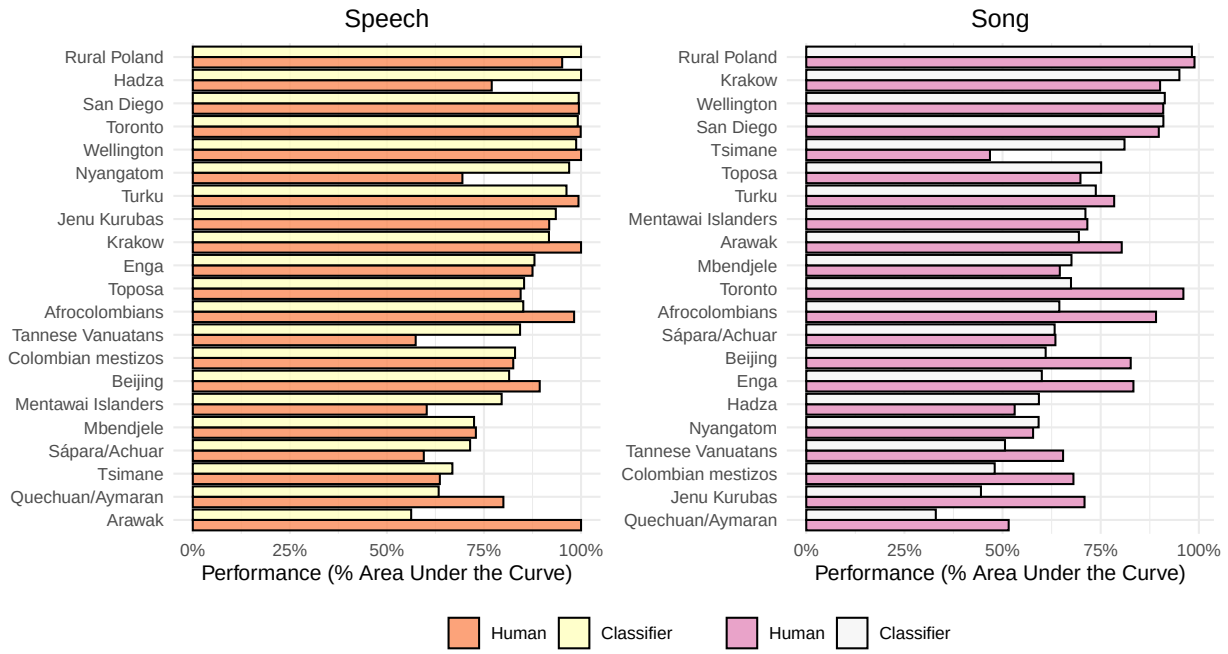
Extended Data Fig. 5 | Exploratory-confirmatory selected acoustic features for pre-registered analyses. The preregistered analyses included comparisons of the acoustic features of infant-directed vocalizations, regardless of whether they included speech or song. For the reasons discussed in the Main Text and Footnote 1, and per the results reported in Fig. 3, 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. 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, following multi-level mixed-effects models; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Prespecified hypotheses about each comparison are posted in the project GitHub repository.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 6 | Machine classifier and naïve listeners perform similarly. Receiver operator characteristic curves show a close match between the model and naïve listeners, after accounting for response bias. The dotted lines represent the model and the solid lines represent human performance, for both speech (orange) and song (blue). The area under the curve values (AUC) visualized in Fig. 4 are derived from this ROC curve; they summarize performance as a single value as the area under the curve.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 7 | Comparison of machine classifier and naïve listener performance across fieldsites. The bar graphs summarize the models' accuracy for classifying infant-directedness in speech and song, using the same receiver operator characteristic curve approach reported in Fig. 4 and Extended Data Fig. 6, but disambiguated across the 21 fieldsites.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Speech			Song		
Fieldsite	d'	95% CI	n vocalists	d'	95% CI	n
Mentawai Islanders	0.048	[-0.389 0.486]	6	-0.085	[-0.304 0.133]	13
Tannese Vanuatans	-0.008	[-0.676 0.66]	2	0	[-0.297 0.296]	10
Tsimane	0.234	[0.038 0.43]	11	0.037	[-0.103 0.178]	12
Sápara/Achuar	0.051	[-0.257 0.36]	10	0.052	[-0.137 0.242]	11
Mbendjele	0.252	[-0.053 0.558]	3	0.111	[-0.073 0.295]	10
Quechuan/Aymaran	0.530	[0.168 0.891]	3	0.143	[-0.064 0.35]	6
Hadza	0.551	[0.239 0.862]	10	0.152	[-0.044 0.348]	9
Nyangatom	0.519	[0.045 0.993]	5	0.175	[-0.049 0.398]	7
Toposa	0.580	[0.212 0.948]	8	0.24	[0.049 0.43]	6
Jenu Kurubas	0.877	[0.465 1.289]	10	0.306	[0.087 0.525]	11
Enga	0.849	[0.437 1.261]	2	NA	NA	0
Colombian mestizos	0.920	[0.257 1.583]	5	0.396	[0.061 0.731]	7
Beijing	1.184	[0.684 1.684]	26	0.48	[0.234 0.727]	28
Afrocolombians	1.188	[0.883 1.494]	4	0.539	[0.355 0.723]	9
Arawak	1.526	[1.181 1.871]	1	0.605	[0.398 0.811]	6
Krakow	1.781	[1.588 1.975]	7	0.707	[0.561 0.854]	7
Turku	1.981	[1.688 2.273]	16	0.729	[0.55 0.909]	14
Rural Poland	1.722	[1.474 1.97]	10	0.747	[0.58 0.915]	7
San Diego	2.329	[2.061 2.597]	13	0.99	[0.825 1.155]	17
Wellington	2.363	[1.834 2.893]	20	1.006	[0.769 1.244]	26
Toronto	2.563	[2.343 2.784]	27	1.135	[0.99 1.28]	23

Extended Data Table 1. d-prime values quantifying sensitivity to infant-directedness in speech and song, independent of response bias, for each fieldsite. Values are estimated as random 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Feature	Variables	Label	Description	Significance
id		filename		
mir_attack	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR, Distance	Attack Curve Slope	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.
mir_roughness	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR, Distance	Roughness	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.
mir_rolloff85	Whole	85th Energy Percentile	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.
mir_inharmonicity	Whole	Inharmonicity	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.
mir_tempo	Whole	Tempo	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.
mir_pulseclarity	Whole	Pule Clarity	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.
npvi_total	Whole	nPVI 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.
npvi_phrase	Whole	nPVI 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(continued)

Feature	Variables	Label	Description	Significance
tm_std_hz	StD	Temporal Modulation	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 rhythmic variability in a recording, with a lower standard deviation leaning towards more monorhythmic signals.
praat_f0	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	Pitch	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.
praat_f0travel	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	Pitch Space	The distance between f0 at each .03125/sec interval to the next.	Pitch space provides a dynamic measure of pitch's range over time.
praat_pitch_rate	Whole, Med, IQR	Pitch Rate	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.
praat_vowtrav	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	Vowel Space	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.
praat_vowtrav_rate	Whole, Med, IQR	Vowel Space Travel Rate	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.
praat_intensity	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR, Distance	Amplitude	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
praat_intensitytravel	Mean, Med, StD, Range, Max, 1st Quart, 3rd Quart, IQR	Amplitude Space	The distance between amplitude at each .03125/sec interval to the next.	Intensity space provides a dynamic measure of intensity's range over time.
praat_intensity_rate	Whole, Med, IQR	Amplitude Rate	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.
praat_f1	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	1st Formant	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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(continued)

Feature	Variables	Label	Description	Significance
praat_f2	Mean, Med, StD, Range, Min, Max, 1st Quart, 3rd Quart, IQR	Second Formant	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.
meta_length		File duration	The length of the unedited sound files	
meta_edit_length		Concatenated file duration	The length of the concatenated versions of the sound files	

Extended Data Table 2. Codebook for acoustic features. Variable names are stubs, i.e., in the datasets, suffixes are added to denote summary statistics. Abbreviations: infant-directed (ID); adult-directed (AD).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparison	Feature	Statistic	β	SE	z	p
ID Speech vs. AD Speech	Pitch (F0)	Inter-Quartile Range	0.996	0.063	15.937	<0.001
	First Formant	Median	0.169	0.067	2.518	0.012
	Intensity	Median	0.393	0.067	5.877	<0.001
	Attack Curve Slope	Median	0.214	0.067	3.175	0.001
	Acoustic Roughness	Median	-0.262	0.065	-4.040	<0.001
	Inharmonicity	Whole	-0.288	0.066	-4.345	<0.001
	Acoustic Roughness	Inter-Quartile Range	-0.153	0.062	-2.469	0.014
	Vowel Travel	Median	0.200	0.054	3.675	<0.001
		Inter-Quartile Range	0.287	0.058	4.954	<0.001
	Vowel-Travel Rate	Median	0.764	0.048	15.979	<0.001
		Inter-Quartile Range	0.775	0.049	15.891	<0.001
	Pitch (F0)	Median	1.250	0.055	22.829	<0.001
	Second Formant	Median	0.305	0.068	4.461	<0.001
		Inter-Quartile Range	0.297	0.069	4.321	<0.001
	Attack Curve Slope	Inter-Quartile Range	0.372	0.068	5.485	<0.001
	Energy Roll-off (85 %ile)	Whole	-0.340	0.068	-5.016	<0.001
ID Song vs. AD Song	Pitch (F0)	Inter-Quartile Range	-0.099	0.028	-3.520	<0.001
	First Formant	Median	-0.111	0.030	-3.690	<0.001
	Intensity	Median	-0.206	0.030	-6.878	<0.001
	Attack Curve Slope	Median	-0.145	0.030	-4.829	<0.001
	Acoustic Roughness	Median	-0.140	0.029	-4.831	<0.001
	Inharmonicity	Whole	-0.091	0.030	-3.068	0.002
	Acoustic Roughness	Inter-Quartile Range	-0.128	0.028	-4.617	<0.001
	Vowel Travel	Median	0.185	0.024	7.589	<0.001
		Inter-Quartile Range	0.224	0.026	8.617	<0.001
	Vowel-Travel Rate	Median	0.081	0.021	3.800	<0.001
		Inter-Quartile Range	0.094	0.022	4.290	<0.001
	Pitch (F0)	Median	-0.034	0.025	-1.388	0.165
	Second Formant	Median	0.012	0.031	0.400	0.689
		Inter-Quartile Range	-0.059	0.031	-1.923	0.054
	Attack Curve Slope	Inter-Quartile Range	-0.046	0.030	-1.515	0.130
	Energy Roll-off (85 %ile)	Whole	0.005	0.030	0.174	0.862

Extended Data Table 3. Regression results from confirmatory analyses (corresponding with Fig. 3). The features tested here were limited to those with significant differences in the exploratory analyses. Statistics are from post-hoc linear combinations following multi-level mixed-effects models. Abbreviations: infant-directed (ID); adult-directed (AD).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Extended Data 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 no statistically significant difference in responses ($b = -0.011, p = .13$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Principal Component 1		Principal Component 2		Principal Component 3	
Feature	Weighting	Feature	Weighting	Feature	Weighting
praat_intensitytravel_mean	-0.198	praat_intensity_mean	0.262	praat_f0_mean	-0.310
praat_intensitytravel_rate_median	-0.197	praat_intensity_median	0.257	praat_f0_third_quart	-0.304
praat_f0travel_rate_IQR	-0.197	praat_intensity_third_quart	0.254	praat_f0_median	-0.296
praat_f0_travel_rate_default	-0.195	praat_intensity_first_quart	0.235	praat_f0_first_quart	-0.256
praat_intensitytravel_rate_IQR	-0.192	mir_roughness_3q	0.220	praat_f0_IQR	-0.217
praat_intensitytravel_median	-0.187	mir_roughness_iqr	0.219	mir_roughness_1q	0.197
praat_f0travel_third_quart	-0.185	mir_roughness_std	0.202	mir_roughness_med	0.174
praat_f0travel_IQR	-0.185	mir_roughness_med	0.198	praat_f0_std	-0.164
praat_intensitytravel_first_quart	-0.185	praat_intensity_max	0.181	praat_intensitytravel_range	-0.147
praat_intensitytravel_rate_default	-0.185	mir_roughness_range	0.171	praat_intensitytravel_max	-0.147
praat_intensitytravel_third_quart	-0.184	mir_roughness_max	0.171	praat_f0travel_first_quart	-0.142
praat_voweltravel_rate_median	-0.181	praat_intensity_min	0.167	praat_intensity_third_quart	-0.142
praat_f0travel_mean	-0.180	mir_roughness_mean	0.163	mir_roughness_mean	0.139
praat_intensitytravel_IQR	-0.178	praat_f1_first_quart	0.154	praat_intensity_mean	-0.137
praat_voweltravel_rate_IQR	-0.177	mir_roughness_1q	0.144	praat_intensity_median	-0.136
praat_voweltravel_rate_default	-0.175	praat_voweltravel_IQR	-0.135	mir_roughness_3q	0.134
praat_f0travel_rate_median	-0.174	praat_voweltravel_third_quart	-0.131	mir_roughness_iqr	0.129
praat_voweltravel_mean	-0.167	praat_f1_std	-0.130	mir_rolloff85	0.127
praat_intensitytravel_std	-0.158	praat_f2_third_quart	-0.128	praat_intensity_first_quart	-0.123
praat_voweltravel_median	-0.158	praat_f2_mean	-0.126	tm_std_hz	-0.115
praat_voweltravel_first_quart	-0.157	praat_intensitytravel_max	0.125	tm_std_idx	-0.115
praat_voweltravel_std	-0.155	praat_intensitytravel_range	0.124	mir_inharmonicity	0.114
praat_f0travel_std	-0.149	praat_f1_min	0.123	praat_f0_max	-0.108
praat_f0travel_median	-0.149	praat_f2_median	-0.121	praat_f2_IQR	-0.108
praat_voweltravel_third_quart	-0.149	praat_f1_range	-0.117	praat_intensity_max	-0.106
praat_voweltravel_IQR	-0.140	praat_f1_median	0.116	praat_f2_min	0.103
praat_intensity_IQR	-0.125	praat_voweltravel_mean	-0.116	praat_intensitytravel_std	-0.102
tm_peak_hz	-0.111	praat_f2_range	-0.116	praat_f1_mean	0.100
tm_peak_idx	-0.111	praat_f2_max	-0.114	praat_f2_std	-0.098
npvi_total	0.100	praat_f1_max	-0.102	praat_voweltravel_rate_default	-0.096

Extended Data Table 5. Factor loadings of the top three principal components reported in Main Text Fig. 3.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Morton, E. S. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist* **111**, 855–869 (1977).

2. Endler, J. A. Some general comments on the evolution and design of animal communication systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* **340**, 215–225 (1993).

3. Owren, M. J. & Rendall, D. Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology* **10**, 58–71 (2001).

4. Fitch, W. T., Neubauer, J. & Herzel, H. Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour* **63**, 407–418 (2002).

5. Wiley, R. H. The evolution of communication: Information and manipulation. *Animal Behaviour* **2**, 156–189 (1983).

6. Krebs, J. & Dawkins, R. Animal signals: Mind-reading and manipulation. in *Behavioural Ecology: An Evolutionary Approach* (eds. Krebs, J. & Davies, N.) 380–402 (Blackwell, 1984).

7. Karp, D., Manser, M. B., Wiley, E. M. & Townsend, S. W. Nonlinearities in meerkat alarm calls prevent receivers from habituating. *Ethology* **120**, 189–196 (2014).

8. Slaughter, E. I., Berlin, E. R., Bower, J. T. & Blumstein, D. T. A test of the nonlinearity hypothesis in great-tailed grackles (Quiscalus mexicanus). *Ethology* **119**, 309–315 (2013).

9. Wagner, W. E. Fighting, assessment, and frequency alteration in Blanchard's cricket frog. *Behavioral Ecology and Sociobiology* **25**, 429–436 (1989).

10. Ladich, F. Sound production by the river bullhead, Cottus gobio L. (Cottidae, Teleostei). *Journal of Fish Biology* **35**, 531–538 (1989).

11. Filippi, P. *et al.* Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences* **284**, (2017).

12. Lingle, S. & Riede, T. Deer mothers are sensitive to infant distress vocalizations of diverse mammalian species. *The American Naturalist* **184**, 510–522 (2014).

13. Custance, D. & Mayer, J. Empathic-like responding by domestic dogs (Canis familiaris) to distress in humans: An exploratory study. *Animal Cognition* **15**, 851–859 (2012).

14. Magrath, R. D., Haff, T. M., McLachlan, J. R. & Igic, B. Wild birds learn to eavesdrop on heterospecific alarm calls. *Current Biology* **25**, 2047–2050 (2015).

15. Lea, A. J., Barrera, J. P., Tom, L. M. & Blumstein, D. T. Heterospecific eavesdropping in a nonsocial species. *Behavioral Ecology* **19**, 1041–1046 (2008).

16. Piantadosi, S. T. & Kidd, C. Extraordinary intelligence and the care of infants. *Proceedings of the National Academy of Sciences* **113**, 6874–6879 (2016).

17. Soltis, J. The signal functions of early infant crying. *Behavioral and Brain Sciences* **27**, 443–458 (2004).

18. Fernald, A. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. in *The adapted mind: Evolutionary psychology and the generation of culture* (eds. Barkow, J. H., Cosmides, L. & Tooby, J.) 391–428 (Oxford University Press, 1992).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Burnham, E., Gamache, J. L., Bergeson, T. & Dilley, L. Voice-onset time in infant-directed speech over the first year and a half. in **19**, 060094 (ASA, 2013).

20. Fernald, A. & Mazzie, C. Prosody and focus in speech to infants and adults. *Developmental Psychology* **27**, 209–221 (1991).

21. Ferguson, C. A. Baby talk in six languages. *American Anthropologist* **66**, 103–114 (1964).

22. Audibert, N. & Falk, S. Vowel space and f0 characteristics of infant-directed singing and speech. in *Proceedings of the 19th international conference on speech prosody* 153–157 (2018).

23. Kuhl, P. K. *et al.* Cross-language analysis of phonetic units in language addressed to infants. *Science* **277**, 684–686 (1997).

24. Englund, K. T. & Behne, D. M. Infant directed speech in natural interaction: Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research* **34**, 259–280 (2005).

25. Fernald, A. The perceptual and affective salience of mothers' speech to infants. in *The origins and growth of communication* (1984).

26. Falk, S. & Kello, C. T. Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* **163**, 80–86 (2017).

27. Bryant, G. A. & Barrett, H. C. Recognizing intentions in infant-directed speech: Evidence for universals. *Psychological Science* **18**, 746–751 (2007).

28. Piazza, E. A., Iordan, M. C. & Lew-Williams, C. Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology* **27**, 3162–3167 (2017).

29. Trehub, S. E., Unyk, A. M. & Trainor, L. J. Adults identify infant-directed music across cultures. *Infant Behavior and Development* **16**, 193–211 (1993).

30. Trehub, S. E., Unyk, A. M. & Trainor, L. J. Maternal singing in cross-cultural perspective. *Infant Behavior and Development* **16**, 285–295 (1993).

31. Mehr, S. A., Singh, M., York, H., Glowacki, L. & Krasnow, M. M. Form and function in human song. *Current Biology* **28**, 356–368 (2018).

32. Mehr, S. A. *et al.* Universality and diversity in human song. *Science* **366**, 957–970 (2019).

33. Trehub, S. E. Musical predispositions in infancy. *Annals of the New York Academy of Sciences* **930**, 1–16 (2001).

34. Trehub, S. E. & Trainor, L. Singing to infants: Lullabies and play songs. *Advances in Infancy Research* **12**, 43–78 (1998).

35. Trehub, S. E. *et al.* Mothers' and fathers' singing to infants. *Developmental Psychology* **33**, 500–507 (1997).

36. Thiessen, E. D., Hill, E. A. & Saffran, J. R. Infant-directed speech facilitates word segmentation. *Infancy* **7**, 53–71 (2005).

37. Trainor, L. J. & Desjardins, R. N. Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review* **9**, 335–340 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Werker, J. F. & McLeod, P. J. Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue Canadienne de Psychologie* **43**, 230–246 (1989).

39. Ma, W., Fiveash, A., Margulis, E. H., Behrend, D. & Thompson, W. F. Song and infant-directed speech facilitate word learning. *Quarterly Journal of Experimental Psychology* **73**, 1036–1054 (2020).

40. Falk, D. Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and Brain Sciences* **27**, 491–502 (2004).

41. Mehr, S. A. & Krasnow, M. M. Parent-offspring conflict and the evolution of infant-directed song. *Evolution and Human Behavior* **38**, 674–684 (2017).

42. Mehr, S. A., Kotler, J., Howard, R. M., Haig, D. & Krasnow, M. M. Genomic imprinting is implicated in the psychology of music. *Psychological Science* **28**, 1455–1467 (2017).

43. Kotler, J., Mehr, S. A., Egner, A., Haig, D. & Krasnow, M. M. Response to vocal music in Angelman syndrome contrasts with Prader-Willi syndrome. *Evolution and Human Behavior* **40**, 420–426 (2019).

44. Cirelli, L. K., Jurewicz, Z. B. & Trehub, S. E. Effects of maternal singing style on mother–infant arousal and behavior. *Journal of Cognitive Neuroscience* (2019). doi:10.1162/jocn_a_01402

45. Cirelli, L. K. & Trehub, S. E. Familiar songs reduce infant distress. *Developmental Psychology* (2020). doi:10.1037/dev0000917

46. Mehr, S. A., Krasnow, M. M., Bryant, G. A. & Hagen, E. H. Origins of music in credible signaling. *Behavioral and Brain Sciences* 1–41 (2020). doi:10.1017/S0140525X20000345

47. Grieser, D. L. & Kuhl, P. K. Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology* **24**, 14 (1988).

48. Fisher, C. & Tokura, H. Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development* **67**, 3192–3218 (1996).

49. Broesch, T. L. & Bryant, G. A. Prosody in Infant-Directed Speech Is Similar Across Western and Traditional Cultures. *Journal of Cognition and Development* **16**, 31–43 (2015).

50. Farran, L. K., Lee, C.-C., Yoo, H. & Oller, D. K. Cross-Cultural Register Differences in Infant-Directed Speech: An Initial Study. *PLOS ONE* **11**, e0151518 (2016).

51. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and Brain Sciences* **33**, 61–83 (2010).

52. Yarkoni, T. The generalizability crisis. *Behavioral and Brain Sciences* (2019). doi:10.1017/S0140525X20001685

53. Broesch, T. & Bryant, G. A. Fathers' Infant-Directed Speech in a Small-Scale Society. *Child Development* **89**, e29–e41 (2018).

54. Ochs, E. & Schieffelin, B. Language acquisition and socialization. *Culture theory: Essays on mind, self, and emotion* 276–320 (1984).

55. Ratner, N. B. Phonological rule usage in mother-child speech. *Journal of Phonetics* **12**, 245–254 (1984).

56. Schieffelin, B. B. *The give and take of everyday life: Language, socialization of Kaluli children.* (CUP Archive, 1990).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57. Ratner, N. B. & Pye, C. Higher pitch in BT is not universal: Acoustic evidence from Quiche Mayan. *Journal of child language* **11**, 515–522 (1984).

58. Pye, C. Quiché mayan speech to children. *Journal of child language* **13**, 85–100 (1986).

59. Heath, S. B. *Ways with words: Language, life and work in communities and classrooms.* (cambridge university Press, 1983).

60. Trehub, S. E. Challenging infant-directed singing as a credible signal of maternal attention. *Behavioral and Brain Sciences* (2021).

61. Räsänen, O., Kakouros, S. & Soderstrom, M. Is infant-directed speech interesting because it is surprising? – Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition* **178**, 193–206 (2018).

62. Cristia, A. & Seidl, A. The hyperarticulation hypothesis of infant-directed speech. *Journal of child language* **41**, 913–934 (2014).

63. Kalashnikova, M., Carignan, C. & Burnham, D. The origins of babytalk: Smiling, teaching or social convergence? *Royal Society Open Science* **4**, 170306 (2017).

64. Kitamura, C., Thanavishuth, C., Burnham, D. & Luksaneeyanawin, S. Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development* **24**, 372–392 (2001).

65. Fernald, A. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development* **60**, 1497–1510 (1989).

66. Broesch, T., Rochat, P., Olah, K., Broesch, J. & Henrich, J. Similarities and Differences in Maternal Responsiveness in Three Societies: Evidence From Fiji, Kenya, and the United States. *Child Development* **87**, 700–711 (2016).

67. ManyBabies Consortium. Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science* **3**, 24–52 (2020).

68. Soley, G. & Sebastian-Galles, N. Infants' expectations about the recipients of infant-directed and adult-directed speech. *Cognition* **198**, 104214 (2020).

69. Byers-Heinlein, K. *et al.* A Multilab Study of Bilingual Infants: Exploring the Preference for Infant-Directed Speech. *Advances in Methods and Practices in Psychological Science* 30 (2021).

70. Fernald, A. *et al.* A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* **16**, 477–501 (1989).

71. Kitamura, C. & Burnham, D. Pitch and Communicative Intent in Mother's Speech: Adjustments for Age and Sex in the First Year. *Infancy* **4**, 85–110 (2003).

72. Kitamura, C. & Lam, C. Age-Specific Preferences for Infant-Directed Affective Intent. *Infancy* **14**, 77–100 (2009).

73. Hilton, C., Crowley, L., Yan, R., Martin, A. & Mehr, S. *Children infer the behavioral contexts of unfamiliar foreign songs.* (PsyArXiv, 2021). doi:10.31234/osf.io/rz6qn

74. Yan, R. *et al. Across demographics and recent history, most parents sing to their infants and toddlers daily.* (PsyArXiv, 2021). doi:10.31234/osf.io/fy5bh

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

75. Custodero, L. A., Rebello Britto, P. & Brooks-Gunn, J. Musical lives: A collective portrait of American parents and their young children. *Journal of Applied Developmental Psychology* **24**, 553–572 (2003).

76. Mendoza, J. K. & Fausey, C. M. Everyday music in infancy. *Developmental Science* (2021). doi:10.31234/osf.io/sqatb

77. Konner, M. Aspects of the developmental ethology of a foraging people. in *Ethological Studies of Child Behaviour* (ed. Blurton Jones, N. G.) 285–304 (Cambridge University Press, 1972).

78. Marlowe, F. *The Hadza hunter-gatherers of Tanzania*. (University of California Press, 2010).

79. Bainbridge, C. M. *et al.* Infants relax in response to unfamiliar foreign lullabies. *Nature Human Behaviour* (2021). doi:10.1038/s41562-020-00963-z

80. Hagen, E. H. & Bryant, G. A. Music and dance as a coalition signaling system. *Human Nature* **14**, 21–51 (2003).

81. Corbeil, M., Trehub, S. E. & Peretz, I. Singing delays the onset of infant distress. *Infancy* **21**, 373–391 (2016).

82. Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L. & Poeppel, D. Human screams occupy a privileged niche in the communication soundscape. *Current Biology* **25**, 2051–2056 (2015).

83. Friedman, J., Hastie, T. & Tibshirani, R. Lasso and elastic-net regularized generalized linear models. Rpackage version 2.0-5. (2016).

84. Fitch, W. T. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America* 11 (1997).

85. Blumstein, D. T., Bryant, G. A. & Kaye, P. The sound of arousal in music is context-dependent. *Biology Letters* **8**, 744–747 (2012).

86. Reber, S. A. *et al.* Formants provide honest acoustic cues to body size in American alligators. *Scientific Reports* **7**, 1816 (2017).

87. Reby, D. *et al.* Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B: Biological Sciences* **272**, 941–947 (2005).

88. Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J. & Mehler, J. An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General* **117**, 21–33 (1988).

89. Werker, J. F. & Lalonde, C. E. Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology* **24**, 672 (1988).

90. Polka, L. & Werker, J. F. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance* **20**, 421–435 (1994).

91. Trainor, L. J., Clark, E. D., Huntley, A. & Adams, B. A. The acoustic basis of preferences for infant-directed singing. *Infant Behavior and Development* **20**, 383–396 (1997).

92. Tsang, C. D., Falk, S. & Hessel, A. Infants prefer infant-directed song over speech. *Child Development* **88**, 1207–1215 (2017).

93. McDermott, J. H., Schultz, A. F., Undurraga, E. A. & Godoy, R. A. Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature* **535**, 547–550 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.09.032995; this version posted October 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

94. Fernald, A. & Simon, T. Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* **20**, 104–113 (1984).

95. Trehub, S. E., Hill, D. S. & Kamenetsky, S. B. Parents' sung performances for infants. *Canadian Journal of Experimental Psychology* **51**, 385–396 (1997).

96. Kirby, K. R. *et al.* D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. *PLOS ONE* **11**, e0158391 (2016).

97. Leeuw, J. R. de. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* **47**, 1–12 (2015).

98. Hartshorne, J. K., Leeuw, J. de, 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).

99. Boersma, P. W. Praat: Doing phonetics by computer. (2019).

100. 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).

101. Patel, A. D. Musical rhythm, linguistic rhythm, and human evolution. *Music Perception* **24**, 99–104 (2006).

102. Mertens, P. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. in (2004).

103. Buyens, W., Moonen, M., Wouters, J. & Dijk, B. van. A model for music complexity applied to music preprocessing for cochlear implants. in 971–975 (IEEE, 2017).

104. Ding, N. *et al.* Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews* **81**, (2017).

105. Yale, C. & Forsythe, A. B. Winsorized regression. *Technometrics* **18**, 291–300 (1976).