Acoustic regularities in infant-directed vocalizations across cultures

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

[1]Department of Psychology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Psychology, University of Washington, Seattle, WA 98105, USA
[3]Department of Politics, Princeton University, Princeton, NJ 08544, USA
[4]Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA
[5]Department of Environmental Health, Faculty of Health Sciences, Jagiellonian University Medical College, 31-531 Krakow, Poland
[6]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
[7]School of Psychology, Victoria University of Wellington, Wellington 6012, New Zealand
[8]Department of Psychology, University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada
[9]Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada
[10]Department of Mathematics, Univesity of California Los Angeles, Los Angeles, CA 90095, USA
[11]Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA
[12]School of Psychology, University of Auckland, Auckland 1010, New Zealand
[13]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, D-07745 Jena, Germany
[14]Department of Psychology, Åbo Akademi, 20500 Turku, Finland
[15]Department of Health Promotion Sciences, Mel & Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ 85724, USA
[16]Division of Infectious Diseases, College of Medicine, University of Arizona, Tucson, AZ 85724, USA
[17]Department of Family & Community Medicine, College of Medicine, University of Arizona, Tucson, AZ 85724, USA
[18]Public Health Research Institute of India, Yadavgiri, Mysore 560020, India
[19]Department Of Epidemiology, Stempel School Of Public Health, Florida International University, Miami, FL 33157, USA
[20]Public Health Research Institute of India, Mysuru 570020, India
[21]Department of Anthropology, Ball State University, Muncie, IN 47306, USA
[22]Department of Anthropology, University College London, WC1H 0BW London, UK
[23]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[24]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85281, USA
[25]Division of Anthropology, California State University, Fullerton, CA 92831, USA
[26]Institute of Evolutionary Medicine, University of Zurich, 8006 Zürich, Switzerland
[27]Université Toulouse 1 Capitole, 31080 Toulouse Cedex 6, France
[28]Institute for Advanced Study in Toulouse, 31080 Toulouse Cedex 6, France
[29]Department of Anthropology, University of California, Davis, Davis, CA 95616, USA
[30]Future Generations University, Circle Ville, WV 26807, USA
[31]Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[32]Data Science Initiative, Harvard University, Cambridge, MA 02138, USA

*Corresponding author. Emails: cmoser@g.harvard.edu (C.J.M.); sam@wjh.harvard.edu (S.A.M.)

Abstract

Humans often produce vocalizations for infants that differ from vocalizations for adults. Is this property common across societies? The forms of infant-directed vocalizations may be shaped by their function in parent-infant communication. If so, infant-directed song and speech should be differentiable from adult-directed song and speech on the basis of their acoustic features, and this property should be relatively invariant across cultures. To test this hypothesis, we built a corpus of 1,614 recordings of infant- and adult-directed singing and speech produced by 411 people living in 21 urban, rural, and small-scale societies. We studied the corpus in a massive online experiment and in a series of acoustic analyses. Naïve listeners ($N = 13{,}218$) reliably identified infant-directed vocalizations as infant-directed, and adult-directed speech (but not songs) as adult-directed, at rates far higher than chance. Ratings of infant-directed song were the most accurate and the most consistent across all societies; infant-directed speech was accurately identified on average, but inconsistently across societies. To determine the mechanisms underlying these results, we extracted many acoustic features from each recording and identified those that most reliably characterize infant-directed song and speech across cultures, via preregistered exploratory-confirmatory analyses and machine classification. The features distinguishing infant- and adult-directed song and speech concerned pitch, rhythmic, phonetic, and timbral attributes; a hypothesis-free classifier with cross-validation across societies reliably identified all vocalization types, with highest accuracy for infant-directed song. Last, we isolated 12 acoustic features that were predictive of perceived infant-directedness; of these, two pitch attributes (median F_0 and its variability) were by far the most explanatory. These findings demonstrate cross-cultural regularities in infant-directed vocalizations that are suggestive of universality; moreover, infant-directed song appears to be more cross-culturally stereotyped than infant-directed speech, informing hypotheses of the functions and evolution of both.

Keywords: *vocalization, human infants, human parents, music, speech, form and function, cross-cultural*

1 Background

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]. This is especially true of vocal signals, where form-function links have been demonstrated across many species, including nonhuman primates[3], meerkats[7], grackles[8], frogs[9], and fish[10].

The link between form and function in vocalizations is also evident from listeners' behavior. For example, humans[11], red deer[12], and canines[13] reliably detect the intentions of heterospecific signalers on the basis of the sounds of their signals. A classic demonstration of this fact is the ability of some species to eavesdrop on the alarm signals of other species, whether or not their own species has an extended vocal repertoire[14,15].

In humans, an area of particular importance for effectively transmitting vocal signals is between parents and infants. This is because human infants are especially helpless to manage their own nutrition, safety, and comfort. Infants use a distinctive alarm signal to elicit care from those around them — they cry[16]. In response, adults and children produce infant-directed vocalizations, which are known to differ reliably from adult-directed vocalizations in at least some societies, in the form of speech[17,18] or song[19–21].

Are the forms and functions of infant-directed vocalizations linked, like the vocal signals of many other species? Fernald[22] noted that a number of features of infant-directed vocalizations observed in Western societies follow Wiley's criteria for signal detection in biological systems[5]. Many others have proposed ways in which infant-directed and adult-directed speech might differ; for example, when compared to adult-directed speech, infant-directed speech may have longer voice-onset times[23]; higher pitch[24,25]; more formant variability[26]; longer and more carefully articulated vowels[27,28], with an upwards-shifted vowel space[29]; more repetition, with longer pitch curves[30]; and more temporal amplitude variability[31]. Many of these features are predicted by functional accounts of stereotyped infant-directed speech, which propose that it facilitates word segmentation[32], distinction of sound categories[33], the elicitation of infant attention[34], or parent-infant communication at a distance[35].

Infants appear to be receptive to at least some of these features, across at least some languages. For example, the ManyBabies Consortium study of 2,329 monolingual infants found reliable preferences for North American English infant-directed speech (relative to North American English adult-directed speech), even when, for more than half of the infants, North American English was not their native language[36]. Infants also have expectations about the infant-directed speech they hear: they look longer at videos of infant-directed speech being directed to an adult-like character, relative to videos of infant-directed speech being directed to an infant-like character, across several languages[37].

Whether or not infant-directedness is characterized by universal acoustic features is unknown, however. Infant-directed speech has rarely been studied outside of Western, Educated, Industrialized, Rich, or Democratic (WEIRD) societies[38], despite a longstanding interest in cross-cultural regularities in infant development[39,40]. No corpora have systematically measured the acoustics of infant-directed speech across a variety of societies, and the pattern of results in smaller studies is unclear.

The prosody of infant-directed speech is similar across tonal and non-tonal languages[41,42]; across French, Italian, German, Japanese, and British and American English[43]; and across Fijian, Kenyan, and North American adults[44]. Across North American English, Swedish, and Russian, infant-directed speech includes vowel accentuation to a more extreme extent than does adult-directed speech[28]. Adults from the Shuar, a South American hunter-horticulturalist group, accurately distinguish infant- from adult-directed speech in recordings of North American English mothers[17]; they do so, in part, on the basis of pitch. This finding echoes reports of raised pitch in Lebanese infant-directed speech[45]. In contrast, the infant-directed speech of fathers in a small-scale Vanuatuan society is rather different in pitch and speech rate than that of North American fathers[46]. And the timbre of infant-directed speech differs from adult-directed speech in ten languages, though with very small samples of speakers[18]. (Note that several studies of the frequency of occurrence of infant-directed speech have been conducted in non-WEIRD and small-scale societies[47,48], but these address a separate question from what acoustic features characterize infant-directed vocalizations when they do occur).

In the domain of music, Mehr and Krasnow proposed that infant-directed song emerged through arms-race co-evolution as an honest signal of parental attention, with acoustic forms elaborated from other vocalizations, such as non-human primate contact calls, so as to provide infants with reliable information that they were being kept safe[49]. This idea is supported by at least three forms of evidence. First, infant-directed song modulates infant arousal, whether the songs are familiar[50] or not[51], and delays the onset of infant distress longer than does infant-directed speech[52]. Second, people with genomic imprinting disorders, which are characterized by altered parental investment behaviors, such as those related to food consumption[53,54], also have altered music perception ability and responses to music[55,56]. Last, consistent with classic ideas in the psychology of music[57–59] substantial evidence demonstrates that lullabies, one typical form of infant-directed song, are a human universal: singing is associated with infant care across the ethnographies of a representative sample of human small-scale societies, even after correcting for reporting biases[21], and parents use singing to calm infants in several of the most genetically distant human societies, the Hadza, Mbuti, and !Kung San hunter-gatherers of East, Central, and South Africa[60–62]. Other forms of infant-directed song, like excitatory play songs and singing games for children, also appear to be widespread[21,63], and parents produce them often[64].

The universality of infant-directed song is also supported by evidence showing that its acoustics differ from those of other forms of music. For example, naïve listeners reliably identify lullabies as infant-directed in a cross-culturally representative sample of vocal music, both when rating multiple functions (e.g., rating the songs more highly as "used to soothe a baby" than "used for dancing"[20]) and in a forced-choice classification task[21]. This finding echoes earlier work, wherein adult listeners were able to distinguish lullabies from love songs recorded in some foreign societies[19]. And machine classifiers reliably distinguish lullabies from healing, dance, and love songs based only on pitches and rhythms of the vocalizations, as opposed to acoustic features merely associated with the vocalization, such as the sound of an infant crying[21].

In sum, while infant-directed song and speech seem to *appear* universally, the ways in which they are acoustically distinct from other vocalizations are not fully understood, nor is it clear whether those acoustic distinctions are themselves universal. This makes it difficult to evaluate the theories of the functions of infant-directed vocalizations mentioned above[32–35,49,57–59], all of which imply the presence of universal acoustic structure in infant-directed speech or song.

To explore these questions, we built a corpus of infant-directed song, infant-directed speech, adult-directed song, and adult-directed speech from a diverse set of 21 human societies. Each participant provided all four recordings, enabling within-person analyses of the differences between the vocalization types. The corpus is open-access at https://osf.io/m5yn2. Here, we report tests of the cross-cultural regularity of the acoustics of infant-directed song and speech, studied via (1) a large-scale listener experiment, where naïve adults recruited online from many countries were asked to discriminate between infant-directed and adult-directed vocalizations in the corpus; and (2) a series of acoustic analyses, to determine reliably-occurring differences in the production and perception of infant-directed vocalizations worldwide.

2 Vocalization corpus

We built a corpus of recordings of infant-directed song, infant-directed speech, adult-directed song, and adult-directed speech. Participants ($N = 411$) living in 21 societies (Figure 1 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 and/or grandparents (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.

For infant-directed song and infant-directed speech, participants sang or spoke to their infant as if they were fussy, where "fussy" could refer to anything from frowning or mild whimpering to a full tantrum (note that each language had its own word for "fussy", suggesting that participants had an intuitive understanding of it). For most participants (90%) an infant was physically present during the recording (the infants were 48% female; mean age 11.4 mo; SD = 0.6 mo; range: 0.5-48). When an infant was not present, participants were

Figure 1. Societies from which vocalizations were recorded. Diamonds denote urban societies; circles denote rural or small-scale societies.

asked to imagine that they were vocalizing to their own infant or grandchild, and simulated their infant-directed vocalizations. For adult-directed song, participants sang a song that was not intended for infants; they also stated what that song was for (e.g., "a celebration song"). For adult-directed speech, participants spoke to the researcher about a topic of their choice (e.g., they described their daily routine).

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[65].

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-degree 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.

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[66], 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. Further details of the acoustic analyses are in the Supplementary Information.

3 Naïve listener experiment

We used the citizen science platform https://themusiclab.org to play excerpts of each item in the corpus to listeners who were unaware of the type of vocalization they heard and who were presumably unfamiliar with many of the societies in which the vocalizations were recorded. This experiment is similar in style to other studies of form and function in vocalization[11,19–21].

Region	Sub-Region	Society	Language	Language Family	Subsistence Type	N
Africa	Central Africa	Mbendjele BaYaka	Mbendjele	Niger-Congo	Hunter-Gatherer	60
	Eastern Africa	Hadza	Hadza	Hadza	Hunter-Gatherer	38
		Nyangatom	Nyangatom	Nilotic	Pastoralist	56
		Toposa	Toposa	Nilotic	Pastoralist	60
Asia	East Asia	Beijing	Mandarin	Sino-Tibetan	Urban	124
	South Asia	Jenu Kurubas	Kannada	Dravidian	Other	80
	Southeast Asia	Mentawai Islanders	Mentawai	Austronesian	Horticulturalist	60
Europe	Eastern Europe	Krakow	Polish	Indo-European	Urban	44
		Rural Poland	Polish	Indo-European	Intensive Agriculturalists	55
	Scandinavia	Turku	Finnish & Swedish	Uralic and Indo-European	Urban	80
North America	North America	San Diego	English (USA)	Indo-European	Urban	116
		Toronto	English (Canadian)	Indo-European	Urban	198
Oceania	Melanesia	Tannese Vanuatuans	Bislama	Indo-European Creole	Horticulturalist	90
		Enga	Enga	Trans-New Guinea	Horticulturalist	22
	Polynesia	Wellington	English (New Zealand)	Indo-European	Urban	228
South America	Amazonia	Arawak	English Creole	Indo-European	Other	48
		Tsimane	Tsimane	Moseten-Tsimane	Horticulturalist	51
		Sápara & Achuar	Quechua & Achuar	Quechuan & Jivaroan	Horticulturalist	59
	Central Andes	Quechua	Quechua	Quechuan	Agro-Pastoralist	49
	Northwestern South America	Afrocolombians	Spanish	Indo-European	Horticulturalist	53
		Colombian Mestizos	Spanish	Indo-European	Commercial Economy	43

Table 1. Societies from which recordings were gathered. N refers to the total number of recordings from each site, not the number of participants.

3.1 Methods

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 jsPsych[67] experiment distributed via Pushkin[68] to both desktop and mobile web browsers. Participants (N = 13,218; gender: 4,405 female, 7,043 male, 176 other, 1,594 did not disclose; age: median 31 years, interquartile range 23-43) listened to at least 1 and at most 16 vocalizations drawn at random from the corpus, for a total of 164,759 ratings (infant-directed song: n = 47,798; infant-directed speech: n = 38,913; adult-directed song: n = 41,277; adult-directed speech: n = 37,071). This yielded over 100 ratings per vocalization (median = 117; interquartile range 107-154) and thousands of ratings for each society (median = 6,394; interquartile range: 4,664–9,569). Most participants (n = 7,241) played the full game, listening to all 16 songs. Participants self-reported living in 109 different countries and speaking 96 different native languages; roughly half the participants were native English speakers from the United States. We excluded excerpts less than 10 seconds in duration from the online experiment, studying 1405 excerpts in total (with representation from all societies).

Participants were asked to classify each vocalization as either infant- or adult-directed (Figure S1), 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). As soon as they made a choice, playback stopped. They were given corrective feedback along with a score at the end of the experiment. Because each instance of the experiment included a new random draw of recordings, we did not exclude participants who disclosed that they had played it more than once ($n = 279$); note, however, that given a random draw of 16 vocalizations from the truncated corpus of 1405 in each instance of the experiment, repeat plays for the 279 participants who played more than once are expected to be rare.

We analyzed the patterns of successful identification of vocalization target across the full corpus and within each society, using both the raw identification accuracy and *d*-prime scores. We also analyzed response time from the onset of each recording, for the subset of responses that were accurate, to explore the speed with which participants made accurate inferences about vocalization types.

3.2 Results

We computed an average score for each vocalization, by averaging across all listeners, and used them as the raw data for the following analyses. Corpus-wide, scores were above chance level, at 65.3% correct (SD = 14.8%, 95% CI: [63.9%, 66.8%]; $t = 20.9$, $p < .0001$, one-sample *t*-test relative to 50.0%). Accuracy varied substantially, however, as a function of the vocalization type (Figure 2A): infant-directed song was identified most accurately (79.7% correct), followed by adult-directed speech (75.4%), and infant-directed speech (68.0%); all these were well above chance (ps $< .0001$). In contrast, adult-directed song was reliably classified *incorrectly*, with only 38.4% accuracy (below chance at $p < .0001$). Here there was also substantial consistency across societies, with all but 2 showing an identical ordering of identification accuracy (in these remaining 2 societies, Wellington and San Diego, infant-directed speech was the highest-accuracy vocalization type).

To examine the degree to which these results held worldwide, we collapsed scores for the vocalizations from each society, in isolation, and analyzed each vocalization type independently (Figure S2; n.b., this analysis substantially reduces the sample size, as some societies had very few recordings available in the naïve listener experiment).

For infant-directed song, the result replicated robustly across societies: infant-directed songs were identified as infant-directed at a significantly higher rate than chance in 19 of 21 sites. In the remaining two societies, perceived infant-directedness trended above chance (Papua New Guinea: $M = .603$; Quechua: $M = .689$) but these sites had only 6 and 5 infant-directed songs, respectively, making it difficult to interpret their non-significant test statistics. Similarly, adult-directed speech was reliably identified as adult-directed in 19 of 21 sites, with trending results in the remaining two sites (Arawak: $M = .552$, $N = 2$; Sápara/Achuar: $M = .605$, $N = 13$).

These results contrast, however, with the identification of infant-directed speech: here, accuracy replicated in only 9 societies, fewer than half of those represented in the corpus. The societies where the naïve listeners failed to identify infant-directed speech accurately tended to be small-scale, including the Hadza, Tsimane, Mbendjele, Toposa, Nyangatom, and Mentawai Islanders (see Figure S2).

To ensure that the above findings were not attributable to response biases, we repeated the overall result using a *d*-prime analysis, which measures accuracy after adjusting for the base rates of response, which were skewed somewhat toward infant-directedness (approximately 60% of items were classified as infant-directed, despite only half actually being infant-directed). This analysis confirmed the main finding reported above (infant-directed song: $d' = 1.11$; adult-directed speech: $d' = 1.30$; infant-directed speech: $d' = 0.93$;

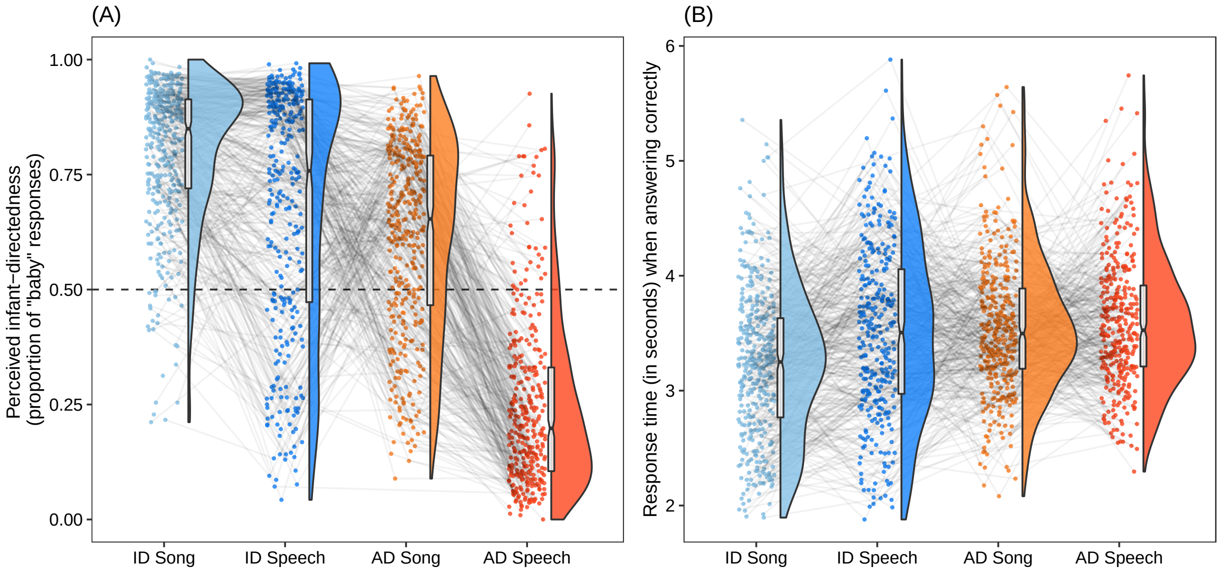


Figure 2. Results of the naïve listener experiment. (A) Listeners accurately identify infant-directed song and infant-directed speech as directed towards infants, and adult-directed speech as directed towards adults; however, they do not identify adult-directed song as directed toward adults. The horizontal dotted line represents chance level of 0.50. (B) When responding correctly, listeners are fastest to identify infant-directed song, followed by infant-directed speech, adult-directed song, and adult-directed speech. In both panels, the points indicate averages for each recording; the gray lines connecting the points indicate the groups of vocalizations produced by the same participant; the half-violins are kernel density estimations; and the boxplots represent the medians, interquartile ranges, and 95% confidence intervals (indicated by the notches). Abbreviations: infant-directed (ID); adult-directed (AD).

adult-directed song: *d*' = -0.07; *d*' scores greater than 0 represent significant results after adjusting for false positives).

Given theoretically-derived predictions that specifically concern the function of infant-directed singing[49], and following our preregistered analysis plan (at https://osf.io/5r72u) for acoustic feature comparisons across vocalization types, we tested for differences in perceived infant-directedness across three comparisons of the vocalizations: (1) infant-directed vs. adult-directed vocalizations, overall; (2) infant-directed song vs. adult-directed song; and (3) infant-directed song vs. infant-directed speech.

In all cases, we analyze *within-voice* differences in perceived infant-directedness (e.g., for all voices, comparing the proportion of "baby" responses for infant-directed songs to infant-directed speech produced by the same voice). This procedure ensures that participant-wise differences in voice characteristics cannot account for differences in the perceived infant-directedness of each vocalization.

We found substantial support for all three predictions (Figure 2A). Perceived infant-directedness was higher in infant-directed vocalizations (proportion of "baby" responses; $M = .743$, $SD = .187$, 95% CI [.724, .762]) than adult-directed vocalizations, overall ($M = .448$, $SD = .182$, 95% CI [.430, .467]; $t(372) = 20.8$, $p < .0001$, $d = 2.07$, paired *t*-test); higher in infant-directed song ($M = .799$, $SD = .152$, 95% CI [.783, .815]) than adult-directed song ($M = .615$, $SD = .208$, 95% CI [.593, .637]; $t(348) = 13.4$, $p < .0001$, $d = 1.29$); and higher in infant-directed song ($M = .806$, $SD = .152$, 95% CI [.789, .824]) than infant-directed speech ($M = .688$, $SD = .263$, 95% CI [.659, .718]; $t(301) = 8.92$, $p < .0001$, $d = 0.83$).

Response time analyses paralleled these findings (Figure 2B). When restricting the sample to correct responses, participants answered more quickly for infant-directed vocalizations (in seconds, $M = 3.34$, $SD = 0.61$, 95% CI [3.28, 3.40]) than adult-directed vocalizations ($M = 3.58$, $SD = 0.46$, 95% CI [3.53, 3.62];

$t(372) = 6.27$, $p < .0001$, $d = 0.54$, paired t-test); more quickly for infant-directed song ($M = 3.24$, $SD = 0.65$, 95% CI [3.17, 3.31]) than adult-directed song ($M = 3.54$, $SD = 0.59$, 95% CI [3.47, 3.60]; $t(348) = 6.99$, $p < .0001$, $d = 0.70$); and more quickly for infant-directed song ($M = 3.19$, $SD = 0.64$, 95% CI [3.12, 3.27]) than infant-directed speech ($M = 3.50$, $SD = 0.76$, 95% CI [3.41, 3.58]; $t(301) = 6.89$, $p < .0001$, $d = 0.70$). Because web-based participants may halt their participation during a trial (producing extremely long response times) or answer quickly at random (producing extremely short response times), in these analyses we removed observations below the 1st and above the 99th percentiles. Also note that in these and the previous paragraph's analyses, summary statistics vary across the comparisons, because a small number of participants did not provide all four of the vocalization types, and because recordings with a duration of less than 10 seconds were excluded from the online experiment. Effect sizes (ds) were computed using the overall standard deviation of accuracy, for consistency across tests.

3.3 Interim discussion

The naïve listener experiment provides evidence that infant-directed vocalizations from around the world are discriminable from adult-directed vocalizations. This effect was most consistent for infant-directed song, which was reliably identified within each society represented in the corpus; while infant-directed speech was reliably identified on average, its society-wise results were less consistent.

Why are listeners so good at identifying infant-directed song? Cross-cultural identification of infant-directedness in music might be due to universal acoustic cues, as predicted from functional accounts of infant-directed vocalizations. In the rest of this paper, we analyze the acoustic features that most reliably characterize infant-directed song, using both confirmatory and hypothesis-free methods, and test the degree to which these features explain overall ratings in the naïve listener experiment.

4 Analysis of acoustic features

We studied a broad range of acoustic features in each vocalization, using Praat[66], MIRtoolbox[69], discrete Fourier transforms for rhythmic variability[70], and normalized pairwise variability indices[71]. The acoustic 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. We extracted a variety of summary variables for each feature, producing 94 variables in total. For example, in the domain of pitch, we included 9 summaries of the feature F_0 (mean, median, minimum, maximum, range, standard deviation, first quartile, third quartile, and interquartile range), and similar summaries for F_1 and F_2, change in F_0, and so on. A codebook for all features is in Table S1.

We ran three sets of analyses. First, we randomly selected half the recordings in the corpus for exploratory analyses, confirming the results on the other half of the corpus, so as to reduce the risk of Type I error. Of particular interest in these analyses were the set of confirmatory hypotheses that we preregistered, following the exploratory analysis, based on functional theories of infant-directed vocalization[32–35,49,57–59] and general principles of signal detection[5].

Second, we used an hypothesis-free machine learning tool, least absolute shrinkage and selection operator (LASSO) classification[72]. To assess how distinct each vocalization type was, in terms of its acoustic features, we evaluated classification accuracy with a cross-validation procedure in which each society's recordings were classified using statistical models trained on the 20 *other* societies. This design allows us to gauge whether acoustic patterns are consistent cross-culturally (following prior research using a similar classification task[21]). The algorithm also includes a variable selection step to identify the specific acoustic features that most reliably characterize each vocalization type across the 21 societies.

Third, we explored the degree to which the convergent results of the first two analyses — namely, the acoustic features that most reliably characterized infant-directed song and infant-directed speech — can explain the results of the naïve listener experiment. We regressed an infant-directedness score for each recording on the acoustic features that predicted infant-directedness in *both* analyses, using a strict inclusion criterion and a conservative correction for multiple tests, to determine the core set of acoustic features characterizing infant-directedness worldwide.

4.1 Exploratory-confirmatory analyses

In exploratory analyses, we fitted a multi-level mixed-effects model for each acoustic feature, adjusting for subject and society and using three predictors: (1) target (infant-directed or adult-directed); (2) utterance type (song or speech); and (3) their interaction. For each model, we tested three linear combinations, to examine differences between (1) infant- and adult-directed vocalizations, overall; (2) infant-directed song and adult-directed song; and (3) infant-directed song and infant-directed speech. This procedure, which was preregistered, mirrors the pairwise comparison analyses in the naïve listener experiment. The linear combinations were evaluated with one-tailed z-tests, using an alpha level of .05. We did not correct for multiple tests in these analyses because the exploratory-confirmatory design restricts the number of tests to those with a strong directional prediction. We did all this with half the corpus, weighted by participant.

In the course of the exploratory analyses, we noted a small number of extreme outliers, typically attributable to anomalies in the recording environment (e.g., loud wind). As such, before running confirmatory analyses, we Winsorized all features at the lowest and highest 5 percentile ranks, and also restricted the set of features analyzed to those less sensitive to extreme observations (e.g., using the median as a measure of central tendency rather than the mean). These data were used for all subsequent analyses. This decision had no impact on the interpretation of results, but is preferable to trimming extreme values[73]; an alternate method, imputing extreme values with the mean observation for each feature, yielded comparable results.

We ran confirmatory models on the subset of acoustic features that were found to distinguish vocalization types in exploratory findings (Table S2), using the other half of the corpus. We were particularly interested in those features for which we had a preregistered directional prediction. These included predictions derived from Mehr and Krasnow[49], suggesting that infant-directed song may universally have longer attack envelopes and pitch contours than infant-directed speech, as well as slower amplitude decay, lower F_0, clearer signal-to-noise parameters, and greater vowel prolongation and stability; slower tempo[22], differential rhythmic variance[70,74], less roughness[75], and shifted vowel spaces[29,76]. The full list of theoretically-motivated hypotheses is at the preregistration (https://osf.io/5r72u) and is summarized in Table S3.

The exploratory-confirmatory procedure yielded 46 significant differences across the three comparison types, confirming some of the preregistered predictions, in terms of pitch, formant, timbre, and temporal features (Figure 3 and Table S4). For example, relative to adult-directed vocalizations, infant-directed vocalizations had a higher pitch and wider pitch variability, faster rates of pitch change and more variability in those rates, and a wider pitch space; a faster rate of vowel space change and more variability in that space; more intensity changes and more variability in intensity; a lower energy profile; and lower inharmonicity. We found similar differences in the other two comparison types, including a few additional acoustic features, such as the normalized pairwise variability index (nPVI, a measure of durational contrast) and attack slopes (a measure of the amplitude change in the onset of acoustic events). The full results are in Table S4.

4.2 Hypothesis-free classification

To validate the results of the exploratory-confirmatory models, we used a hypothesis-free LASSO-regularized categorical classifier[72] to identify the four different vocalization types on the basis of their acoustic features alone. Cross-cultural accuracy was assessed using society-wise leave-one-out cross-validation, as in previous research[21]. We then rotated the held-out society 20 more times, to analyze accuracy across all 21 societies. The classifier used acoustic features standardized within-voices, eliminating between-voice variability in the acoustic features.

The classifier accurately identified 70.5% of held-out recordings from unseen societies ([62.9%, 78.0%]; 95% CIs from corrected and resampled t-tests[77]), far above chance level of 25%. This finding justifies a strong claim of corpus-wide consistency: to predict vocalization types in a given society, the classifier only used information available from *other* societies, and did so with a high degree of accuracy (Figure 4A).

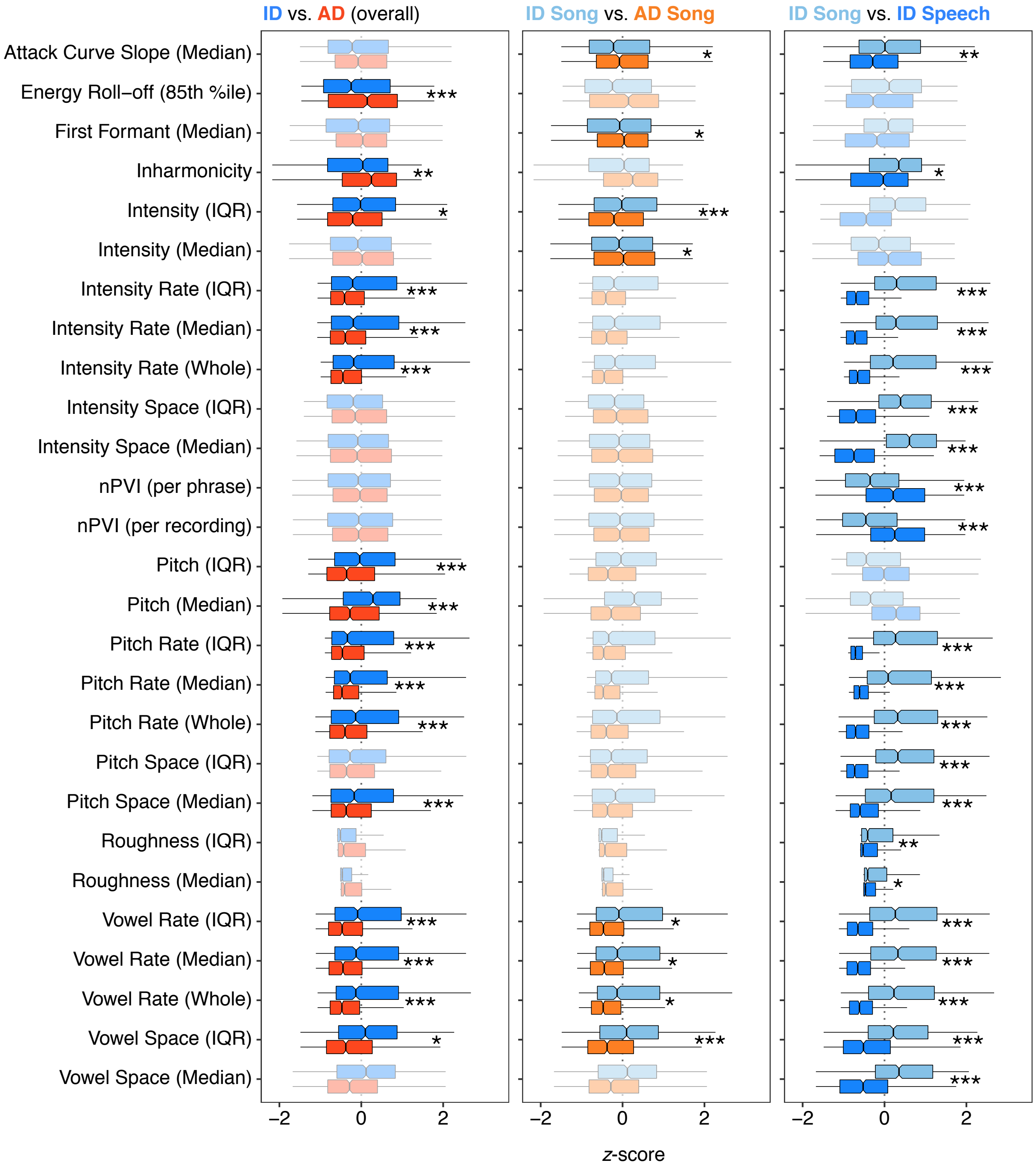


Figure 3. Confirmatory results. The boxplots represent the 25 acoustic features with a significant difference in at least one main comparison (e.g., infant-directed song vs. infant-directed speech, in the right panel), in both the exploratory and confirmatory analyses. All variables are normalized across 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. Abbreviations: infant-directed (ID); adult-directed (AD). Significance values are computed via linear combinations, following multi-level mixed-effects models. ***$p < .001$; **$p < .01$; *$p < .05$

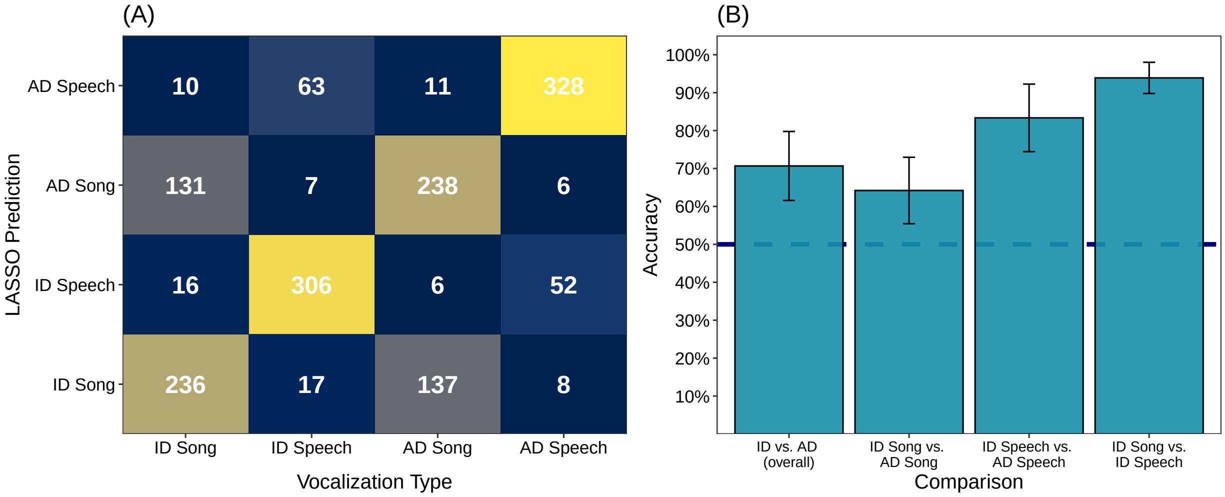


Figure 4. Accuracy of hypothesis-free classifiers. (A) The confusion matrix for the four-way categorical LASSO classifier shows successful classification in all four vocalization types. When misclassifying, the model is more likely to confuse the target (infant or adult) than the vocalization type (song or speech). (B) The bar graph displays the accuracy of each of the pairwise classifiers; all pairwise classifications were above chance level of 50% (denoted by the horizontal dotted line). Error bars denote 95% confidence intervals from corrected and re-sampled *t*-tests. Abbreviations: infant-directed (ID); adult-directed (AD).

The confusion matrix also reveals patterns of *misidentification*: in the 29.5% of recordings that are misidentified, the model rarely classifies songs as speech (or vice versa), but sometimes confuses the utterance target within the correct vocalization type. For example, infant-directed songs are more than 10 times more likely to be classified inaccurately as adult-directed songs than to be classified inaccurately as adult-directed speech — but nevertheless, the model accurately identifies them as infant-directed songs most of the time (60.0% relative to chance level of 25%).

To identify the acoustic features that most reliably differentiate pairs of vocalization types, we continued with a logistic LASSO classifier to test the same three pairwise comparisons as in the exploratory-confirmatory analyses and the analysis of the naïve listener experiment: (1) infant-directed vs. adult-directed vocalizations, overall; (2) infant-directed song vs. adult-directed song; and (3) infant-directed song vs. infant-directed speech. We also ran a fourth pairwise comparison, between infant-directed speech and adult-directed speech, as an exploratory analysis.

The classifiers performed strikingly well (Figure 4B; infant-directed vs. adult-directed vocalizations, overall: 70.7% [61.6%, 79.8%]; infant-directed song vs. adult-directed song: 64.2% [55.4%, 73.0%]; infant-directed song vs. infant-directed speech: 93.9% [89.8%, 98.0%]). Infant-directed speech was also reliably distinguished from adult-directed speech (83.4% [74.4%, 92.3%]).

Last, we examined the acoustic features identified by the variable selection step of the LASSO procedure, which most reliably predict vocalization type across all 21 societies. These are reported in Table 2.

There was substantial overlap between the results of the two approaches (Table 2): out of 31 features selected by the LASSO classifier, 22 were supported by at least one exploratory-confirmatory result, and of those, 6 were preregistered. Consistent with the exploratory-confirmatory analyses, the acoustic features that reliably distinguished between each vocalization form concerned pitch, formant, timbre, and temporal features; in some cases, these included additional variables, such as pulse clarity (the strength of the beats, detected via music information retrieval) and temporal modulation (the frequency decomposition of the amplitude envelope, or how quickly loudness changes).

Feature	Statistic	ID [+] vs. AD [−] (overall)	ID Song [+] vs. AD Song [−]	ID Song [+] vs. ID Speech [−]	ID Speech [+] vs. AD Speech [−]
Attack Curve Slope	IQR	0.155	0.182	.	0.108
	Median^{pre}	−0.139	**−0.373**	**−0.352**	0.176
Inharmonicity	Whole^{pre}	**−0.125**	−0.204	**−0.029**	−0.04
Pulse Clarity	Whole^{pre}	**0.161**	0.069	0.336	0.19
85th Energy Percentile	Whole^{pre}	**−0.243**	−0.216	.	−0.152
Roughness	IQR	−0.162	−0.159	−0.151	.
	Median	0.178	.	**−0.520**	0.002
Tempo	Whole^{pre}	.	0.047	0.12	−0.007
nPVI per Phrase	Whole^{pre}	−0.053	−0.061	**−0.021**	.
Pitch	IQR	**0.093**	−0.16	.	0.386
	Median	**0.738**	0.097	0.259	1.276
Pitch Space	IQR	**−0.112**	**−0.105**	**−0.782**	.
	Median	**0.108**	−0.216	**−0.909**	0.128
Pitch Rate	IQR	**0.146**	−0.052	**−0.735**	0.123
	Median	**0.178**	0.306	.	.
First Formant	IQR	0.032	0.024	.	.
	Median	−0.115	**−0.114**	**−0.369**	.
	Range	−0.23	−0.328	−0.121	−0.009
Second Formant	Median	0.042	−0.149	0.082	0.176
Intensity	IQR	**0.471**	0.295	**−0.225**	0.456
	Median	−0.406	**−0.511**	0.595	.
Intensity Space	IQR	−0.72	−0.543	.	−0.523
	Median	−0.436	−0.154	**−0.368**	−0.295
Intensity Rate	IQR	**0.466**	.	.	0.08
	Median	.	0.6	.	.
Vowel Space	IQR	0.51	**0.911**	.	.
	Median	**0.032**	**0.062**	.	.
Vowel Travel Rate	IQR	**0.234**	**0.567**	.	.
	Median	.	**−1.033**	**−1.256**	0.984
Temporal Modulation	Peak^{pre}	0.166	0.138	.	.
	SD^{pre}	**0.069**	0.005	**0.045**	0.03

Table 2. Acoustic features that reliably differentiate the four vocalization types, selected via LASSO classification with cross-validation across societies. The table reports coefficients from penalized logistic regressions using acoustic features (standardized within-voices). Changes in the values of the coefficients produce changes in the predicted log-odds ratio, so the values in the table can be interpreted as in a logistic regression. The features supported by convergent evidence from the exploratory-confirmatory analyses are in bold; those that were preregistered are marked with a superscript *pre*. Abbreviations: infant-directed (ID); adult-directed (AD).

4.3 Convergent analysis: Predicting listener intuitions from acoustic features

Last, we examined the degree to which the naïve listener's perceptions of infant-directedness were explicable from the primary acoustic features identified by the exploratory-confirmatory and hypothesis-free analyses of the corpus. To reduce the risk of introducing false-positives in a large dataset, we only analyzed acoustic features that had convergent evidence from at least one summary statistic in both the exploratory-confirmatory and LASSO analyses, in at least one comparison type. In these analyses, we collapsed across all vocalization types and attempted to predict only whether naïve listeners rated a given vocalization as infant- or adult-directed (regardless of society or vocalization type). This yielded 21 features. To justify a strong interpretation of potential relations between these 21 features and infant-directedness in the corpus, we regressed each vocalization's average infant-directedness score on each of the 21 features individually, using a strict Bonferroni-adjusted alpha level of .0024.

This procedure yielded 12 features that were significantly predictive of listeners' perceptions of infant-directedness after this selection procedure (Figure 5 and Table S5). The most reliably associated feature, by far, was pitch: median F_0 (Figure 5A) and its variability (Figure 5B) each accounted for about 30% of the variability in perceived infant-directedness; other features related to infant-directedness included intensity space (Figure 5C), temporal modulation (Figure 5D), roughness (Figure 5E), and inharmonicity (Figure 5F).

Last, we entered all 12 features into a multiple linear regression. These features explained 45.0% of the variability in perceived infant-directedness ($F(12, 1081) = 73.7, p < .0001$). When entered into the regression together, 5 of the 12 features had significant partial effects (median F_0: $\beta = 0.30$; F_0 IQR: $\beta = 0.33$; median intensity travel rate: $\beta = -0.17$; roughness IQR: $\beta = -0.12$; median F_1: $\beta = -0.09$). Thus, while 12 core acoustic features are reliably associated with infant-directedness across the corpus, there is nonetheless substantial additional variability in the infant-directedness of vocalizations that is left unexplained.

5 Discussion

We provide convergent evidence for widespread regularities in the acoustic design of infant-directed vocalizations, in both the domains of language and music. Naïve listeners reliably identified infant-directed vocalizations as infant-directed, despite the fact that the vocalizations were largely of unfamiliar geographic and linguistic origin, and more consistently in song than in speech. A series of hypothesis- and data-driven analyses showed consistent acoustic distinctions between infant-directed and adult-directed vocalizations overall, between infant-directed and adult-directed song, and between infant-directed song and infant-directed speech. These acoustic distinctions together explained nearly half the variability in listeners' perceptions of infant-directedness.

The most consistent ways in which infant-directed vocalizations differ from adult-directed vocalizations, worldwide, concern pitch: nearly every comparison revealed differences in pitch, pitch space, and pitch rate (Figure 3), and, moreover, F_0 median and interquartile range explained by far the largest proportion of variability in listeners' perceived infant-directedness (Figure 5). But other acoustic features also reliably distinguished infant-directed vocalizations from adult-directed vocalizations, infant-directed song from adult-directed song, and infant-directed song from infant-directed speech — albeit in subtler ways that the LASSO classifier detected more reliably than did naïve listeners. These features included rhythmic, phonetic, and timbral characteristics of the vocalizations, such as temporal modulation, durational contrast, roughness, inharmonicity, and intensity space (Figure 4, Table 2, and Table S4).

Simply put: across many voices from many cultures producing many speech and song utterances, infant-directed vocalizations tend to sound different than adult-directed vocalizations. The differences are salient enough for naïve listeners to detect, because they are characterized by a core set of acoustic dimensions — more consistently in infant-directed song than in infant-directed speech. Taken together, these findings suggest a link between form and function in the design of infant-directed vocalizations.

Surprisingly, however, naïve listeners' intuitions about infant-directed speech were far less consistent across societies than their intuitions about infant-directed song. Corpus-wide, both vocalization types were iden-

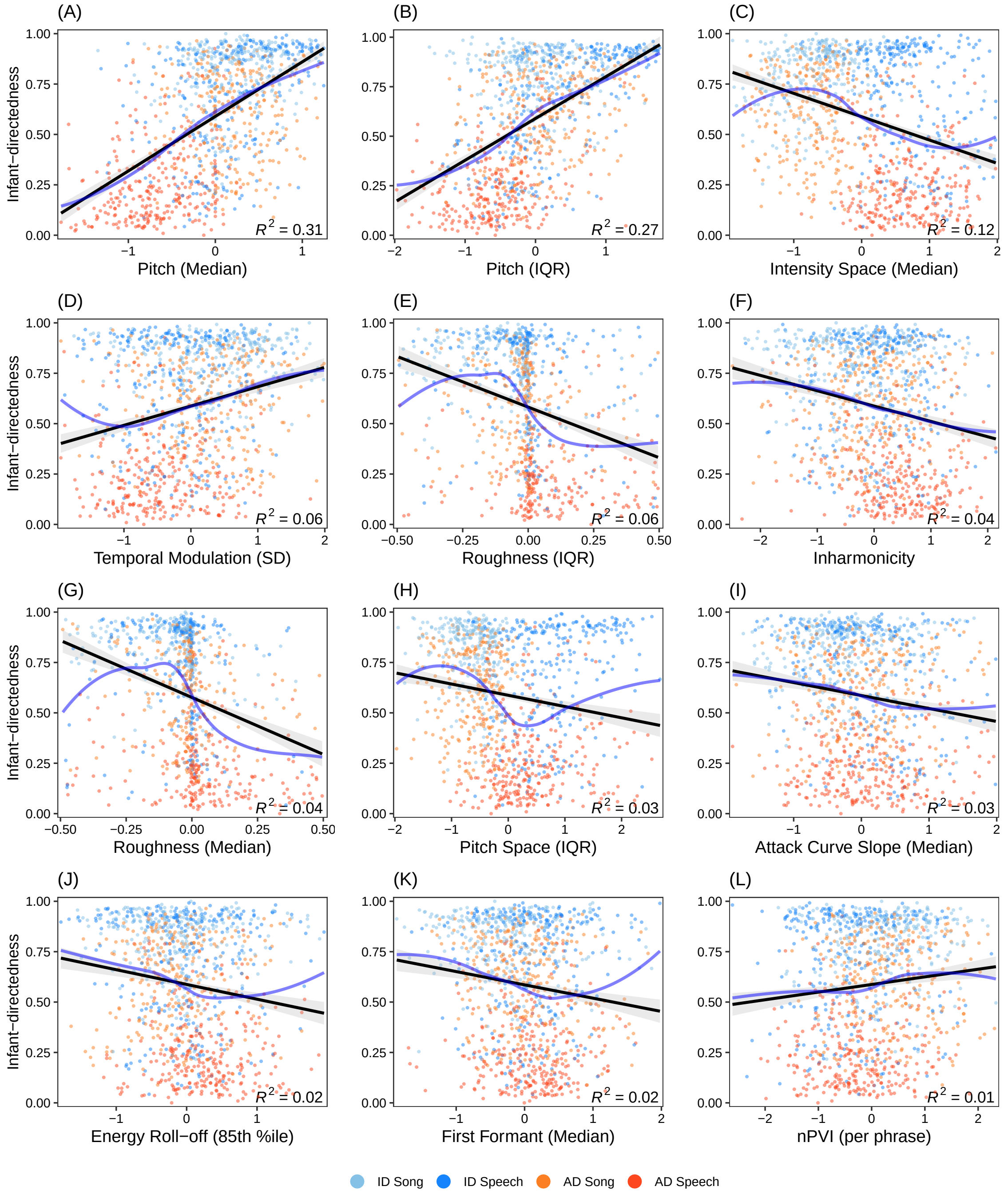


Figure 5. Twelve acoustic features reliably predict infant-directedness across societies. (A-L) The scatterplots each correspond to a single acoustic feature (indicated on the x-axis). They represent the average naïve listener ratings of infant-directedness for each recording in the corpus (measured by the average proportion of "baby" responses in the online experiment), as a function of each acoustic feature (normalized across participants and centered within participants). The features plotted here survived a Bonferroni correction for 21 tests and, further, were included only if they were supported by convergent evidence from both LASSO and exploratory-confirmatory analyses. The black line represents the linear model corresponding to the reported R^2, which is significant at $p < .0024$; the gray shaded area a 95% confidence interval; and the blue line a LOESS regression. The x-axes of some panels are truncated to facilitate visualization.

tified well above chance level, but this analysis masked some cross-cultural variability: when analyzing performance within each society independently, infant-directed song was always identified reliably, but infant-directed speech was identified reliably less than half the time. Moreover, those societies where the naïve listeners failed to identify infant-directed speech tended to be small-scale, contrasting with typical “citizen scientist” participants, who are recruited mostly from industrialized populations. This suggests that the corpus-wide identification rate for infant-directed speech is inflated by the listeners’ familiarity with the style of infant-directed speech found in societies similar to their own — and raises the intruiging possibility that infant-directed speech is more variable, worldwide, than is infant-directed song.

This research leaves open at least four questions. First, while the results point in the direction 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), the participants whose vocalizations we studied do not constitute a representative sample of humans. As such, no strong claims of universality are justified concerning the acoustic structure of infant-directedness. This issue could be addressed by (a) studying larger, representative samples of infant-directed vocalizations; (b) using phylogenetic methods to examine whether people in societies that are very distantly related nonetheless produce similar infant-directed vocalizations; (c) testing perceived infant-directedness in a more diverse sample of listeners, to more accurately characterize cross-cultural variability in the *perception* of infant-directedness; and (d) testing listener intuitions among groups with reduced exposure to a given set of infant-directed vocalizations, such as very young infants or people from distantly related small-scale societies.

Second, despite a large body of work in bioacoustics examining the structure of vocal signals[1–3,3–15], it is not yet clear the extent to which the variability in acoustic features identified here is unique to humans, or whether it reflects more general principles underlying cross-species regularities in vocal signals. It is notable, for example, that many of the acoustic features that are reduced in infant-directed vocalization (Table 2) are associated with harsh, nonlinear sounds commonly accentuated in alarm calls across species[4,78]. Comparative studies may help to disentangle the ways in which human vocal signals are shaped in ways that are different from other animals, or not.

Third, our findings say little about the *content* of infant-directed vocalizations, which are known to vary widely: song and speech are used in a wide variety of contexts with infants, of which soothing (the type of vocalization we elicited from participants) is just one. One curious finding reported here, where naïve listeners reliably characterize adult-directed song *inaccurately* as infant-directed, may bear on this question — perhaps this simply reflects a predisposition in our listeners to finding solo, mostly female voices, as soothing — given a wider variety of contexts for the solo singing, perhaps the naïve listeners would have responded differently. Similarly, the sounds of arousing or alerting infant-directed speech and soothing infant-directed speech are likely to differ consistently from one another across cultures[22], just as different forms of infant-directed song differ from one another (e.g., lullabies vs. play songs[63]). Future studies should determine the degree of generality of the present findings across a wider variety of contexts.

Last, the corpus-building approach used here may help to empirically test theories on the origins and functions of music and speech in infancy. For example, if infant-directed song communicates the costly investment of parental attention[55], then infant-directed song should feature increased flashiness and variability in salient acoustic characteristics for infants — consistent with the present findings of higher energy in second formants (important for vowel recognition[79]) and faster travel over a vowel space. Moreover, the relation between infant-directedness and the sounds of vowels is consistent with classic experimental evidence demonstrating infants’ robust perceptual sensitivity to vowels[79–81]. In contrast, cross-cultural variability in infant-directed speech found in the naïve listener experiment weighs against any universality prediction from functional accounts of infant-directed speech[32–35]; however, given the relatively high accuracy of the LASSO classifiers in distinguishing infant- from adult-directed speech across the societies studied, more research is needed to clarify those aspects of infant-directed speech that are culturally invariant.

Whatever the answers to these questions, the results presented here demonstrate that infant-directed vocalizations — and especially infant-directed song — are a fundamental aspect of human communication, characterized by acoustic regularities across many cultures.

Data, code, and materials availability

Data and code are available at https://github.com/themusiclab/infant-vocal; the corpus is available at https://osf.io/m5yn2; the preregistration is at https://osf.io/5r72u; and readers may participate in the naïve listener experiment at https://themusiclab.org/quizzes/ids.

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. 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. C.M.B. and S.A. provided essential research assistance. 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 in collaboration with S.A.M. and M.M.K. S.A.M. led analyses, with contributions from C.J.M., D.K., and M.M.K. S.A.M. made the figures. C.J.M., H.L-R., M.M.K., and S.A.M. wrote the manuscript and all authors approved it.

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.

Competing interests

We declare we have no competing interests.

Funding

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.); 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 and National Heart, Lung, and Blood Institute, and the National Institute of Neurological Disorders and Stroke Award D43 TW010540 (P.M., C.D.P.); the National Institute of Allergy and Infectious Diseases Award R15-AI128714-01 (P.M.); the Max Planck Institute for Evolutionary Anthropology (C.T.R.); 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 (J. Stieglitz); and the Natural Sciences and Engineering Research Council of Canada (S.E.T.).

Acknowledgments

We thank the participants and their families for providing recordings; D. Amir, who sparked the idea for this research in conversation with S.A.M. at the 2016 Annual Conference of the Human Behavior & Evolution Society; J. Du, E. Pillsworth, L. Sugiyama, P. Wiessner, and J. Ziker, who collected or attempted to collect additional recordings; A. Bergson, Z. Jurewicz, D. Li, L. Lopez, and E. Radytė for research assistance; and M. Bertolo, J. Kominsky, and L. Yurdum for feedback on the manuscript.

Supplementary Information

Details of acoustic feature extraction

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 females 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

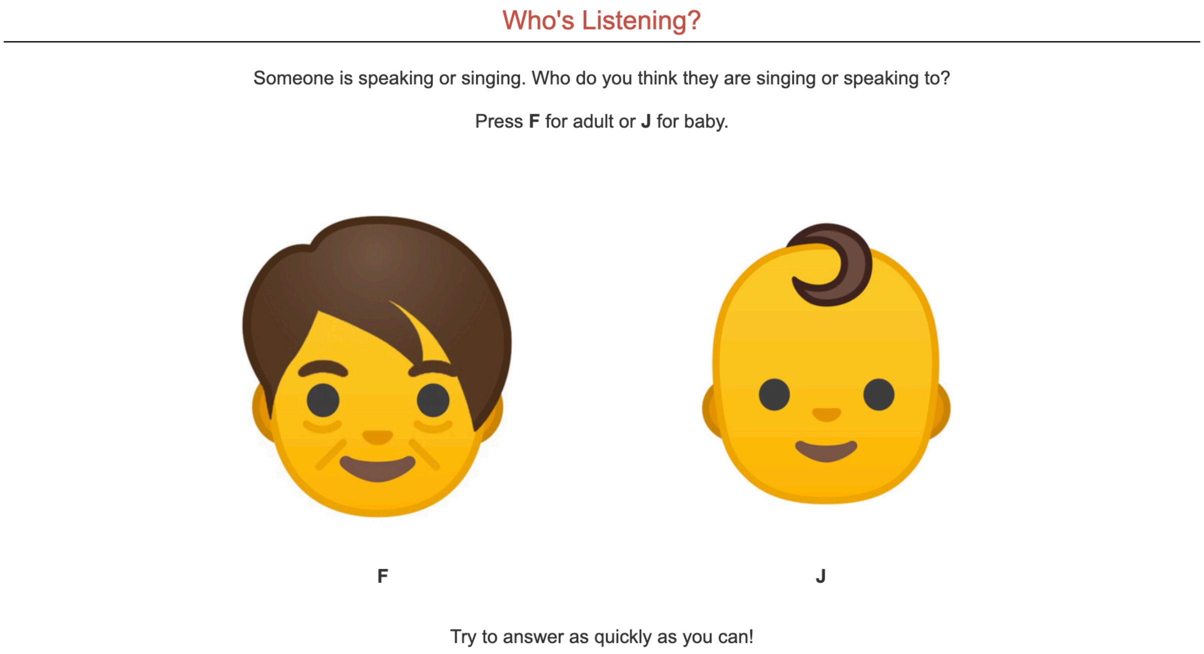
All MIRtoolbox (v. 1.7.2) features were extracted with default parameters[69]. *mirattackslope* returns a list of all attack slopes detected, so final analyses were done on summary features (e.g., mean, median, etc.). Final analyses were also done on summary features for *mirroughness*, which returns time series data of roughness measures in 50ms windows. We RMS-normalized the mean of *mirroughness* following[82]. 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[83] 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[70]. We used an automated syllable- and phrase-detection algorithm to extract events[71]. 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.


Who's Listening?

Someone is speaking or singing. Who do you think they are singing or speaking to?

Press **F** for adult or **J** for baby.

Try to answer as quickly as you can!

Figure S1. Screenshot from the naïve listener experiment (desktop computer version). 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 an infant.

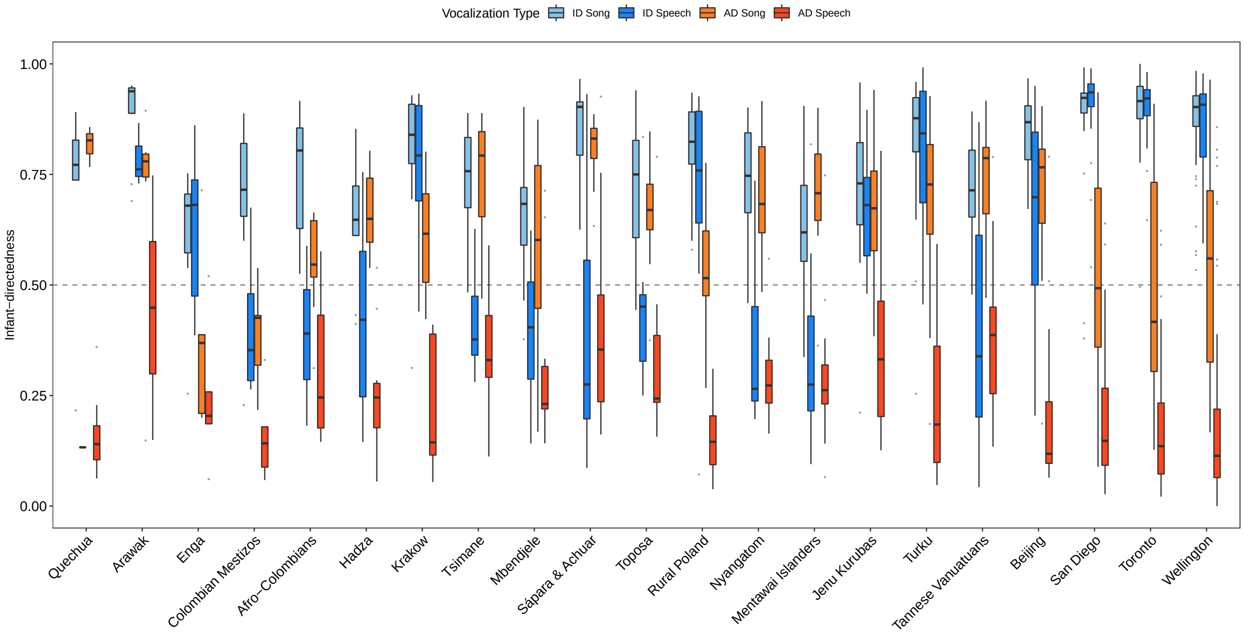


Figure S2. Perceived infant-directedness, analyzed separately for each society. For each vocalization type, the boxplots indicate the within-society median (horizontal black line), interquartile range (box), 1.5 × IQR (whiskers), and outliers (gray points). The societies are ordered from the smallest to largest number of recordings (from left to right). Abbreviations: infant-directed (ID); adult-directed (AD).

Variable	Label	Description
id	filename	
mir_attack	Attack Curve Slope	MIRtoolbox detects events in the audio; for a subset of those it can compute an attack slope, which is the slope of the line from the beginning of the event to its peak.
mir_roughness	Roughness	A roughness measure based on the dissonant beating patterns produced by interference frequencies in the spectrum of the sound. MIRtoolbox produces a roughness curve; following Buyens et al. (2017), we reduce this to a single measure by taking the RMS-normalized mean.
mir_rolloff85	85th Energy Percentile	One way to estimate the amount of high frequency in the signal consists in finding the frequency such that a certain fraction of the total energy is contained below that frequency. This ratio is fixed by default to .85 (following Tzanetakis and Cook, 2002), other have proposed .95 (Pohle, Pampalk and Widmer, 2005).
mir_inharmonicity	Inharmonicity	mirinharmonicity "estimates the inharmonicity, i.e., the amount of partials that are not multiples of the [automatically detected] fundamental frequency, as a value between 0 and 1. More precisely, the inharmonicity considered here takes into account the amount of energy outside the ideal harmonic series." (MIRtoolbox manual)
mir_tempo	Tempo	MIRtoolbox tempo detection with default parameters. Based on MIRtoolbox's event detection. Outputs a single number.
mir_pulseclarity	Pule Clarity	Estimates the rhythmic clarity, indicating the strength of the beats estimated by the mirtempo function.
npvi_total	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.
npvi_phrase	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.
tm_std_hz	Temporal Modulation	The temporal modulations spectrum is the frequency decomposition of the amplitude envelope of a signal. This measures how loud something is at any given moment, and then we measure how fast the loudness changes. Trivial example: if the song is someone singing a note every second, the spectrum will have a peak at 1Hz. If the song is someone singing 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. We're interested in the standard deviation of the spectrum, which we're construing as how exaggerated the peak is.
praat_f0	Pitch	The pitch (f0) in Hertz for each song
praat_pitch_rate	Pitch Rate	The pitch rate is a measure of pitch change over unit time. In essence, the pitch rate gives us a measure of pitch curve smoothness (a lower value corresponds to a smoother curve).
praat_vowtrav	Vowel Space	The euclidian distance travelled in vowel space. This is equivalent to distance between two formants.
praat_vowtrav_rate	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 travel.
praat_intensity	Amplitude	A measure of amplitude (loudness) in decibels
praat_intensity_rate	Amplitude Rate	A measure of decay in intensity curves in each song measured as change in intensity over rate in time.
praat_f1	First Formant	The frequency in Herz of the first formant at each (.03125/sec) point
praat_f2	Second Formant	The frequency in Herz of the second formant at each (.03125/sec) point
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

Table S1. 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).

Comparison	Feature	Statistic	*Est.*	*SE*	*z*	*p*
ID vs. AD (overall)	85th Energy Percentile	Whole	-744.65	155.62	-4.79	< .001
	Attack Curve Slope	Median	0.41	0.20	2.03	0.043
	First Formant	Maximum	-172.06	35.97	-4.78	< .001
		Range	-186.41	38.91	-4.79	< .001
	Inharmonicity	Whole	-0.01	0.00	-4.28	< .001
	Intensity	IQR	0.68	0.30	2.22	0.026
		Minimum	0.86	0.38	2.27	0.023
	Intensity Rate	Whole	-4.42	0.48	-9.25	< .001
		Whole	2.99	0.43	6.92	< .001
	Intensity Space	Mean	0.62	0.11	5.79	< .001
		St. Dev.	1.76	0.26	6.65	< .001
	Pitch	First Quartile	27.88	4.04	6.91	< .001
		Third Quartile	59.44	11.28	5.27	< .001
		IQR	31.52	8.55	3.69	< .001
		Mean	42.19	6.91	6.11	< .001
		Median	45.47	7.34	6.19	< .001
		Minimum	8.13	2.72	2.99	0.003
		St. Dev.	13.00	3.64	3.57	< .001
	Pitch Rate	Whole	-37.30	4.34	-8.59	< .001
		Whole	23.36	4.62	5.05	< .001
	Pitch Space	First Quartile	0.51	0.10	5.18	< .001
		Mean	3.24	1.34	2.42	0.015
		Median	1.61	0.33	4.87	< .001
		St. Dev.	6.99	2.35	2.98	0.003
	Second Formant	Maximum	-114.81	25.77	-4.46	< .001
		Median	35.63	12.51	2.85	0.004
		Range	-115.51	33.30	-3.47	0.001
	Vowel Space	Third Quartile	46.81	15.12	3.10	0.002
		IQR	45.23	12.97	3.49	< .001
		Mean	38.13	10.68	3.57	< .001
		St. Dev.	51.71	10.80	4.79	< .001
	Vowel Space Travel Rate	Whole	212.31	37.85	5.61	< .001
ID Song vs. AD Song	Attack Curve Slope	First Quartile	-0.45	0.21	-2.12	0.034
		Median	-0.80	0.41	-1.97	0.049
	First Formant	Median	-19.66	9.70	-2.03	0.043
	Inharmonicity	Whole	-0.01	0.00	-2.15	0.032
	Intensity	First Quartile	-1.95	0.55	-3.57	< .001
		Third Quartile	-1.45	0.50	-2.88	0.004
		Maximum	-1.13	0.51	-2.22	0.027
		Mean	-1.60	0.48	-3.35	0.001
		Median	-1.63	0.51	-3.18	0.001
		Minimum	-0.80	0.31	-2.59	0.01
	nPVI Recording	Whole	-2.14	0.86	-2.50	0.012
	Pitch	Minimum	-9.00	3.00	-3.00	0.003
	Tempo	Whole	5.80	2.75	2.11	0.035
	Temporal Modulation	Peak	0.65	0.32	2.03	0.042
	Vowel Space	Third Quartile	27.90	11.00	2.54	0.011
		IQR	24.94	9.58	2.60	0.009
		Mean	20.29	6.74	3.01	0.003
		St. Dev.	18.94	6.31	3.00	0.003
	Vowel Space Travel Rate	Whole	23.44	11.22	2.09	0.037
ID Song vs. ID Speech	Attack Curve Slope	First Quartile	-0.67	0.26	-2.59	0.01

(continued)

Comparison	Feature	Statistic	*Est.*	*SE*	*z*	*p*
		Third Quartile	-1.85	0.37	-5.05	< .001
		IQR	-1.19	0.27	-4.39	< .001
		Mean	-1.11	0.27	-4.13	< .001
		Median	-1.17	0.32	-3.65	< .001
	First Formant	First Quartile	-24.19	7.30	-3.31	0.001
		Third Quartile	-57.27	19.77	-2.90	0.004
		Maximum	112.08	30.98	3.62	< .001
		Mean	-39.98	10.94	-3.66	< .001
		Median	-41.69	11.88	-3.51	< .001
		Minimum	-26.11	5.25	-4.97	< .001
		Range	138.18	34.18	4.04	< .001
	Inharmonicity	Whole	-0.01	0.00	-3.10	0.002
	Intensity	First Quartile	1.29	0.48	2.68	0.007
		IQR	-1.44	0.35	-4.15	< .001
		Minimum	-0.94	0.35	-2.64	0.008
		St. Dev.	-0.60	0.20	-2.92	0.003
	Intensity Space	First Quartile	-0.29	0.03	-9.56	< .001
		Third Quartile	-1.68	0.23	-7.30	< .001
		IQR	-1.39	0.20	-6.80	< .001
		Mean	-1.73	0.15	-11.74	< .001
		Median	-0.76	0.08	-9.48	< .001
		St. Dev.	-2.66	0.29	-9.08	< .001
	nPVI Phrase	Whole	7.21	1.27	5.67	< .001
	nPVI Recording	Whole	5.68	1.34	4.25	< .001
	Pitch	Maximum	-23.98	11.46	-2.09	0.036
		St. Dev.	-11.25	5.10	-2.21	0.027
	Pitch Space	First Quartile	-0.54	0.16	-3.38	0.001
		Third Quartile	-14.25	1.78	-8.02	< .001
		IQR	-13.71	1.81	-7.57	< .001
		Maximum	-23.15	11.48	-2.02	0.044
		Mean	-16.16	1.50	-10.76	< .001
		Median	-2.97	0.31	-9.70	< .001
		Range	-23.15	11.48	-2.02	0.044
		St. Dev.	-18.79	2.56	-7.35	< .001
	Pulse Clarity	Whole	0.02	0.01	3.44	0.001
	Roughness	Third Quartile	-13.00	3.99	-3.26	0.001
		Distance	-746.17	224.00	-3.33	0.001
		IQR	-12.96	3.91	-3.32	0.001
		Mean	-177.13	41.50	-4.27	< .001
		Median	-2.55	0.96	-2.66	0.008
		St. Dev.	-54.89	18.84	-2.91	0.004
	Second Formant	Maximum	83.42	27.09	3.08	0.002
		Median	-49.14	21.99	-2.23	0.025
		Minimum	-69.20	23.20	-2.98	0.003
		Range	152.58	41.31	3.69	< .001
	Temporal Modulation	St. Dev.	0.53	0.06	8.23	< .001
	Vowel Space	First Quartile	-24.33	3.59	-6.77	< .001
		Third Quartile	-97.33	14.50	-6.71	< .001
		IQR	-73.02	11.76	-6.21	< .001
		Mean	-82.31	9.28	-8.87	< .001
		Median	-47.59	6.97	-6.83	< .001
		St. Dev.	-83.54	10.56	-7.91	< .001
	Vowel Space Travel Rate	Whole	-298.47	32.34	-9.23	< .001

Table S2. Significant results from exploratory analyses, using post-hoc linear combinations following multi-level mixed-effects models. Abbreviations: infant-directed (ID); adult-directed (AD).

Feature	Variable	ID vs. AD	ID Song vs. AD Song	ID Song vs. ID Speech
85th Energy Percentile	Whole	$-^{1}$	$-$	$-$
Attack Curve Slopes	Median	$-$	$-^{1}$	$-^{1}$
Attack Curve Slopes	Mean	$-$	$-$	$-^{1}$
First Formant	Mean	$-$	$-$	$-^{1}$
First Formant	Max	$-^{1}$	$-$	$-^{0}$
Inharmonicity	Whole	$-^{1}$	$-$	$-^{1}$
Intensity	Minimum	$-^{1}$	$-$	$-^{1}$
Intensity Rate	Whole	$-^{0}$	$-$	$-^{1}$
nPVI per Phrase	Whole	$+$	$+$	$+^{1}$
nPVI per Recording	Whole	$+$	$+$	$+^{1}$
Pitch	Mean	$+^{1}$	$-$	$-$
Pitch Space	Mean	$-$	$-$	$-^{1}$
Pitch Rate	Whole	$-^{0}$	$-$	$-^{1}$
Pulse Clarity	Whole	$+^{1}$	$+$	$+$
Roughness	Mean	$-$	$-$	$-^{1}$
Second Formant	Mean	$-$	$-$	$-$
Second Formant	Max	$-^{1}$	$-^{0}$	$-^{0}$
Tempo	Whole	$-$	$-$	$-$
Temporal Modulation	St. Dev.	$-$	$-$	$-^{0}$
Temporal Modulation	Peak	$-^{0}$	$-$	$-^{1}$
Vowel Space	Mean	$+^{1}$	$+^{1}$	$+$
Vowel Space Travel Rate	Whole	$+^{1}$	$+^{1}$	$+$

Table S3. Preregistered predictions. Predictions that were supported by the exploratory-confirmatory analyses are marked ^{1} while predictions which were significantly falsified in the opposite direction are marked ^{0}. Abbreviations: infant-directed (ID); adult-directed (AD).

Comparison	Feature	Statistic	*Est.*	*SE*	*z*	*p*
ID vs. AD (overall)	85th Energy Percentile	Whole	-665.11	182.20	-3.65	< .001
	Inharmonicity	Whole	-0.01	0.00	-3.03	0.002
	Intensity	IQR	0.46	0.18	2.51	0.012
	Intensity Rate	Whole	2.07	0.43	4.81	< .001
		IQR	2.08	0.52	4.04	< .001
		Median	0.85	0.21	4.05	< .001
	Pitch	IQR	26.27	5.89	4.46	< .001
		Median	41.55	7.64	5.44	< .001
	Pitch Rate	Whole	13.20	3.30	4.00	< .001
		IQR	12.61	3.29	3.84	< .001
		Median	3.12	0.66	4.70	< .001
	Pitch Space	Median	1.19	0.25	4.73	< .001
	Vowel Space	IQR	30.83	12.52	2.46	0.014
	Vowel Space Travel Rate	Whole	144.97	29.12	4.98	< .001
		IQR	179.47	41.49	4.33	< .001
		Median	71.18	15.08	4.72	< .001
ID Song vs. AD Song	Attack Curve Slope	Median	-0.44	0.19	-2.31	0.021
	First Formant	Median	-12.58	6.02	-2.09	0.037
	Intensity	IQR	-1.73	0.24	-7.17	< .001
		Median	-1.20	0.54	-2.22	0.026
	Vowel Space	IQR	26.84	7.02	3.82	< .001
	Vowel Space Travel Rate	Whole	24.82	11.50	2.16	0.031
		IQR	39.07	16.13	2.42	0.015
		Median	11.72	5.80	2.02	0.043
ID Song vs. ID Speech	Attack Curve Slope	Median	-0.81	0.25	-3.29	0.001
	First Formant	Median	-33.81	6.47	-5.23	< .001
	Inharmonicity	Whole	-0.01	0.00	-2.02	0.043
	Intensity Rate	Whole	-3.92	0.36	-10.89	< .001
		IQR	-5.03	0.43	-11.61	< .001
		Median	-2.11	0.17	-12.25	< .001
	Intensity Space	IQR	-1.33	0.16	-8.46	< .001
		Median	-0.83	0.07	-11.32	< .001
	nPVI per Phrase	Whole	4.39	1.14	3.87	< .001
	nPVI per Recording	Whole	4.81	0.88	5.45	< .001
	Pitch Rate	Whole	-28.11	2.54	-11.05	< .001
		IQR	-31.97	2.56	-12.51	< .001
		Median	-5.78	0.56	-10.25	< .001
	Pitch Space	IQR	-9.99	0.77	-12.94	< .001
		Median	-2.70	0.23	-11.63	< .001
	Roughness	IQR	-6.63	2.04	-3.25	0.001
		Median	-1.52	0.73	-2.07	0.038
	Vowel Space	IQR	-55.11	7.82	-7.05	< .001
		Median	-31.27	3.98	-7.87	< .001
	Vowel Space Travel Rate	Whole	-227.44	21.26	-10.70	< .001
		IQR	-310.78	27.95	-11.12	< .001
		Median	-124.53	11.05	-11.27	< .001

Table S4. Significant results from confirmatory analyses, after Winsorization and excluding variables with extreme observations (e.g., using median and IQR instead of mean and standard deviation), using post-hoc linear combinations following multi-level mixed-effects models.

Feature	$F(1, 1094)$	p	R^2
Pitch (Median)	489.9	5.27×10^{-90}	0.309
Pitch (IQR)	411.6	6.30×10^{-78}	0.273
Intensity Space (Median)	149.5	2.61×10^{-32}	0.120
Temporal Modulation	74.5	2.16×10^{-17}	0.064
Roughness (IQR)	69.9	1.86×10^{-16}	0.060
Inharmonicity	51.1	1.59×10^{-12}	0.045
Roughness (Median)	49.2	4.05×10^{-12}	0.043
Pitch Space (IQR)	29.2	7.89×10^{-8}	0.026
Attack Curve Slope (Median)	29.0	8.74×10^{-8}	0.026
Energy Roll-off (85th %ile)	26.6	2.92×10^{-7}	0.024
First Formant (Median)	19.4	1.14×10^{-5}	0.017
nPVI (per phrase)	13.1	3.01×10^{-4}	0.012

Table S5.
Omnibus tests from simple linear regressions of perceived infant-directedness (from the naive listener experiment) on each of 12 acoustic features validated by exploratory-confirmatory and LASSO analyses. All tests are significant at the Bonferroni-corrected alpha level of .0024.

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

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

18. 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).

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

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

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

22. 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).

23. Burnham, E., Gamache, J. L., Bergeson, T. & Dilley, L. Voice-onset time in infant-directed speech over the first year and a half. in *Proceedings of Meetings on Acoustics ICA2013* **19**, 060094 (ASA, 2013).

24. Fernald, A. Prosody in speech to children: Prelinguistic and linguistic functions. *Annals of Child Development* **8**, 43–80 (1991).

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

26. 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).

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

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

29. 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).

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

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

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

33. 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).

34. 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).

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

36. 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).

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

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

39. Bowlby, J. *Attachment and Loss (Vol. I: Attachment).* (Basic Books, 1969).

40. Konner, M. *The evolution of childhood: Relationships, emotion, mind.* (Belknap Press of Harvard University Press, 2010).

41. 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).

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

43. 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).

44. 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).

45. 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**, (2016).

46. Broesch, T. & Bryant, G. A. Fathers' infant-directed speech in a small-scale society. *Child Development* **89**, e29–e41 (2018).

47. Cristia, A., Dupoux, E., Gurven, M. & Stieglitz, J. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development* **90**, 759–773 (2019).

48. Konner, M. Infancy among the Kalahari desert San. in *Culture and Infancy: Variations in the Human Experience* (eds. Leiderman, H., Tulkin, S. R. & Rosenfeld, A. H.) 287–328 (Academic Press, 1977).

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

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

51. Bainbridge, C. *et al.* Infants relax in response to unfamiliar foreign lullabies. *PsyArXiv* (2020). doi:10.31234/osf.io/xcj52

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

53. Cassidy, S. B. & Driscoll, D. J. Prader-Willi syndrome. *European Journal of Human Genetics* **17**, 3–13 (2008).

54. Williams, C. A. *et al.* Angelman syndrome 2005: Updated consensus for diagnostic criteria. *American Journal of Medical Genetics* **140**, 413–418 (2006).

55. 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).

56. 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).

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

58. Peretz, I. The nature of music from a biological perspective. *Cognition* **100**, 1–32 (2006).

59. McDermott, J. & Hauser, M. The origins of music: Innateness, uniqueness, and evolution. *Music Perception* **23**, 29–59 (2005).

60. Fan, S. *et al.* African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology* **20**, 82 (2019).

61. 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).

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

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

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

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

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

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

68. Hartshorne, J. K., de Leeuw, J., Goodman, N., Jennings, M. & O'Donnell, T. J. A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods* **51**, 1782–1803 (2019).

69. 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).

70. Patel, A. D., Iversen, J. R. & Rosenberg, J. C. Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America* **119**, 3034 (2006).

71. Mertens, P. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. in *Speech Prosody 2004, International Conference* (2004).

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

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

74. Salselas, I. & Herrera, P. Development of perception and representation of rhythmic information: Towards a computational model. in *2010 9th International Conference on Development and Learning* (IEEE, 2010).

75. 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).

76. Diehl, R. L. Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 965–978 (2007).

77. Nadeau, C. & Bengio, Y. Inference for the generalization error. *Machine Learning* **52**, 239–281 (2003).

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

79. 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).

80. 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).

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

82. Buyens, W., Moonen, M., Wouters, J. & van Dijk, B. A model for music complexity applied to music preprocessing for cochlear implants. in *2017 25th European Signal Processing Conference (EUSIPCO)* 971–975 (IEEE, 2017).

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