nature human behaviour

Article

https://doi.org/10.1038/s41562-022-01489-2

Deep learning reveals what vocal bursts express in different cultures

Received: 28 April 2022

Accepted: 26 October 2022

Published online: 28 December 2022

Check for updates

Jeffrey A. Brooks[1,2] ✉, Panagiotis Tzirakis[1], Alice Baird[1], Lauren Kim[1], Michael Opara[1], Xia Fang[3], Dacher Keltner[1,2], Maria Monroy[2], Rebecca Corona[2], Jacob Metrick[1] & Alan S. Cowen[1,2] ✉

Human social life is rich with sighs, chuckles, shrieks and other emotional vocalizations, called 'vocal bursts'. Nevertheless, the meaning of vocal bursts across cultures is only beginning to be understood. Here, we combined large-scale experimental data collection with deep learning to reveal the shared and culture-specific meanings of vocal bursts. A total of $n = 4{,}031$ participants in China, India, South Africa, the USA and Venezuela mimicked vocal bursts drawn from 2,756 seed recordings. Participants also judged the emotional meaning of each vocal burst. A deep neural network tasked with predicting the culture-specific meanings people attributed to vocal bursts while disregarding context and speaker identity discovered 24 acoustic dimensions, or kinds, of vocal expression with distinct emotion-related meanings. The meanings attributed to these complex vocal modulations were 79% preserved across the five countries and three languages. These results reveal the underlying dimensions of human emotional vocalization in remarkable detail.

Brief emotional vocalizations such as cries, sighs, laughs, shrieks, grunts, growls, oohs and ahhs (to name but a few[1–4]) provide a ubiquitous and informationally rich scaffolding to our social lives[2,3,5–7]. Known as 'vocal bursts', they are thought to predate language, with precursors in other mammals such as stereotyped non-human primate vocalizations related to food, predators and mating[8]. Vocal bursts play a profound role in early life, with spontaneous vocalizations by parent and child forming a critical feedback loop helping infants learn to navigate challenges and opportunities in their environment[2,3]. The perceptual system is readily attuned to vocal bursts, with the understanding of subtle distinctions among laughs, coos and ahhs appearing in infants as young as 12–17 months[9,10]. And the information conveyed by vocal bursts across the lifespan can be surprisingly rich, signalling the structure of social interactions such as a speakers' rank in a social hierarchy or the quality of friendships[11,12].

Despite their centrality to human social life, our understanding of the specific emotions conveyed by vocal bursts across multiple cultures and languages is only beginning to emerge[4,13,14]. The science of emotion, long focused on facial expression, has historically underemphasized non-linguistic vocalizations (with some exceptions[15]). Vocal bursts have just recently been investigated across cultures, using methods pioneered in the study of facial expression: presenting relatively small curated samples of vocal bursts to small groups of participants[4,13,14,16]. These studies reveal vocal bursts to be at least as universal in meaning as facial expressions, demonstrating that people in many different cultures reliably attribute a range of specific meanings to different types of vocal bursts[4,17]. This early work provided evidence that 13 distinct emotion concepts could be matched with corresponding vocal bursts in more than 14 cultural groups, including two remote societies with minimal Western influence[4,12,13,16] (but see ref. [14]). However, with at most only a few hundred vocal bursts in any given study, investigations have been underpowered to precisely characterize the nuanced meanings inferred from different kinds of vocal bursts in different cultures[18,19]. The studies to date have also been underpowered to control for perceptual and linguistic confounds, such as the influence of speakers' demographics (for example, gender, age and ethnicity[20,21]) on inferred meanings. Finally, studies have been limited by their reliance on imperfect translations of emotion concepts across

[1]Research Division, Hume AI, New York, NY, USA. [2]University of California, Berkeley, Berkeley, CA, USA. [3]Zhejiang University, Hangzhou, China. ✉e-mail: jeff@hume.ai; alan@hume.ai

languages[22]. Thus, several fundamental questions about vocal bursts remain unanswered. How many distinct meanings do vocal bursts convey? How can these meanings be precisely conceptualized? How well are they preserved across cultures?

In the present investigation, we addressed these questions and the limitations of past studies using data-driven methods that produced a quantitative description of the distinct emotional meanings that vocal expressions reliably convey in diverse cultures. We did so by analysing vocal bursts inductively at a large scale (282,906 vocal bursts contributed by 4,031 participants in five countries, who self-reported the meanings of the vocalizations they were forming) and mapping dimensions of emotional meaning to their underlying structural (acoustic) dimensions with machine learning. We generated a wide range of vocalizations by starting with over 2,756 distinctive sounds that participants were instructed to imitate. We explored a wide range of meanings of the imitated vocal bursts by gathering intensity ratings corresponding to 48 emotion concepts, ranging from intense positive and negative states to more neutral states such as confusion and interest, in five countries and three languages. This allowed us to capture a wide range of meanings people attribute to emotional expressions[23–25] as well as how these meanings overlap, diverge and blend. We derived speaker-invariant measurements of vocal bursts by training a deep neural network (DNN) to predict the average meanings inferred from vocal bursts solely from their imitations by globally diverse individuals.

Importantly, we modelled cultural differences. We also avoided mistranslating emotion concepts across cultures by training the DNN to predict average meanings in each country separately. This means that the DNN had no prior knowledge of how emotion concepts translated across cultures or languages (a pattern of vocal expression found to be associated with "joy" in one country could still just as easily be found to be associated with "sadness" in another, even if they speak the same language). Finally, we assessed how many distinct dimensions of meaning were captured by the DNN using principal preserved components analysis (PPCA). Using these methods, we discovered that vocal bursts convey at least 24 distinct dimensions of meaning. These 24 dimensions of vocal expression were 79% preserved in meaning across the five countries and three languages, with 21 dimensions showing a high degree of shared meaning across all countries studied and the remaining two dimensions having similar meanings in four out of five countries. Our results capture the underlying dimensions of the meanings of emotional vocalizations within and across cultures in unprecedented detail.

Results

Vocal burst imitation of seed recordings and self-report

To derive the cross-cultural dimensional structure of the meaning of vocal bursts while addressing perceptual, linguistic and demographic sources of variation, we conducted a large-scale experiment in two phases. In the first phase of data collection (henceforth 'mimicry phase'), a total of 4,031 participants from China (n = 380; 213 female), India (n = 377, 78 female), South Africa (n = 1,155; 712 female), the USA (n = 1,492; 762 female) and Venezuela (n = 203; 65 female) completed a vocal burst mimicry task, imitating randomly sampled subsets of 2,756 vocal bursts and rating what each vocal burst meant to them before or after they imitated it (participants were able to complete up to 30 trials per survey; Methods; Supplementary Table 1 and Supplementary Fig. 1 give extended demographic information). The seed recordings included wide-ranging vocal bursts produced in laboratory settings in five countries[17], vocal bursts produced in ecological settings gathered in online video[17] and newly assembled vocal bursts extracted from Chinese and Japanese media (Methods). On the basis of past estimates of reliability of observer judgement, for each seed recording we collected mimicry responses (ratings and mimicked vocal bursts) from an average of 20.5 separate participants in each culture. This amounted to a total of 282,906 experimental trials for which we successfully obtained mimicry responses.

Before engaging in the mimicry task, participants were instructed to use their computer microphone to record themselves on each trial. On each trial, participants heard a target vocal burst and were instructed to mimic the vocal burst such that their imitation would be perceived to convey similar emotions to the original recording. On the same survey page, participants were asked to judge what they thought the vocal burst expressed by selecting from 48 emotion terms (Methods; Supplementary Table 2) and then rating each selection from 1 to 100, with values reflecting the perceived intensity of the emotion (note that these ratings were not intended to measure alternative dimensions such as valence or arousal, just the intensity of each emotion concept selected). Participants were required to select a value on a rating scale for at least one category (Supplementary Fig. 2 gives distributions of ratings of the seed vocal bursts). English terms were used in the three out of five countries where English is an official language (India, South Africa and the USA). In China, ratings were collected in Chinese and in Venezuela ratings were collected in Spanish (Supplementary Table 2 provides a complete list of terms and their translations in each language). This first phase of data collection resulted in many participant-generated 'mimic' recordings in each culture (China (n = 42,765), India (n = 44,059), South Africa (n = 66,842), the USA (n = 103,228) and Venezuela (n = 26,012)), for a total of 282,906 vocal bursts. Furthermore, the judgements provided during this phase of data collection constitute self-report ratings, given that they capture what each vocal burst meant to the person making it. Confirming our well-documented ability to mimic vocalizations, the mimicked vocal bursts are for the most part qualitatively similar to the original seed vocal bursts, except for differences in participants' vocal traits and recording quality (for examples, see Supplementary Table 3).

Independent emotion ratings of mimic recordings

We made further use of these stimuli in the second phase of data collection (henceforth 'rating-only phase'), in which we recruited an independent set of 4,998 participants from all five countries (China (n = 277; 159 female), India (n = 438; 194 female), South Africa (n = 1,401; 893 female), the USA (n = 2,475; 1,337 female) and Venezuela (n = 189; 70 female)) to complete an emotion perception task. In this rating-only task, observer participants rated audio recordings from the mimicry task from within their own country. Participants' ratings in this task thus capture culture-specific understandings of the original seed stimuli but, unlike the self-report ratings from the mimicry task, could not have been influenced by the perception of any demographic or contextual influences on the acoustic properties of the seed stimuli. Thus, their correlations with the seed stimuli represent measures of the emotional meaning of these original expressions that are invariant to certain perceptual biases and confounds (gender, age, race, acoustic context and so on) that can be problematic for studies of vocal emotional expression[20]. As in the judgement portion of the mimicry phase of data collection, participants were asked to judge each recording along 48 emotion terms (Supplementary Table 2) and select a value ranging from 1 to 100 for each term selected, with values reflecting the perceived intensity of the emotion. On average, participants in this phase of the experiment completed 77.1 trials. This amounted to a total of 577,732 judgements of all mimic recordings.

Shared dimensions of vocal expression

In our recent work computing shared dimensions of emotional experience and expression between two cultures, we used a method we are calling PPCA[26–29]. In the present study, with datasets measuring the same attributes in five different countries, we developed a generalized version of the PPCA algorithm (G-PPCA) that extracts linear combinations of attributes that maximally covary across three or more datasets (in this case, emotion judgements from five countries).

The resulting components are ordered in terms of their level of positive covariance across all five datasets (see Methods for more information on the method and significance testing).

We first investigated the cross-cultural dimensions of perceived emotion using perceptual judgements of the seed recordings collected during the mimicry phase. We applied G-PPCA to judgements of the 2,756 seed recordings across the five countries. We iteratively applied G-PPCA in a leave-one-stimulus-out manner to extract components from the judgements of all but one stimulus and then projected each country's ratings of the left-out stimulus onto the extracted components, resulting in cross-validated component scores for each country and stimulus.

To determine the statistical significance of each component, we sought to ensure that the extracted dimensions not only reflect shared structure preserved across all five countries but also reflect significantly preserved dimensions across pairs of countries in the analysis. First, we calculated the one-tailed partial Spearman correlation between corresponding component scores for each country pair, iteratively partialling out each previous component and calculating statistical significance separately for each dimension. One-tailed tests were used since we were specifically interested in dimensions that were preserved across countries. Within country pairs, P values for the 48 dimensions were false discovery rate (FDR)-corrected using the Benjamini–Hochberg procedure. We found that 28 semantic dimensions, or distinct kinds of emotion, were preserved across all five cultures in judgements of the seed recordings (Supplementary Fig. 3; see Methods section on G-PPCA significance testing for more detail).

To interpret and visualize the preserved components, we applied varimax factor rotation (Methods). Here, factor rotation extracts a simplified representation of the space that prioritizes dimensions loading onto a small set of categories. The top component loadings for each preserved component thus reflect a semantic summary of the distinct patterns of perceived emotion captured by the preserved components for visualization and communication purposes.

For the original set of vocal bursts, upon visual inspection of the rotated factor matrix (Supplementary Fig. 3) we found that each of the 28 dimensions loaded maximally on a specific category, reflecting distinct semantic dimensions of emotion perceived from vocal bursts (ordered in terms of covariance explained in ratings of vocal bursts across five countries): "satisfaction", "confusion", "pride", "triumph", "awe", "fear", "surprise (positive)", "disappointment", "disgust", "embarrassment", "interest", "ecstasy", "boredom", "determination", "romance", "guilt", "excitement", "craving", "amusement", "sexual desire", "sympathy", "surprise (negative)", "relief", "horror", "shame", "contemplation", "realization" and "tiredness".

This is consistent with prior work showing that a high-dimensional semantic space represents the shared meanings of emotion-related experience and perception (of faces, vocal bursts, prosody and music) across cultures and is better able to predict the structure of human emotion judgements above and beyond a few emotion categories or affective dimensions[27,30,31]. This work also converges with the high-dimensional structure of emotion observed in more top-down studies of emotion production and recognition[32–34]. But despite the scale of the dataset, these findings could be partly driven by the specific stimulus set and any biased sampling it reflects. In particular, the ratings that went into this analysis could in theory be influenced by confounding factors of the original audio stimuli, such as race, gender and acoustic context of each speaker, rather than solely reflecting speaker-independent vocal modulations underlying the expressions.

The potential influence of expressers' traits and context is a serious confound for all studies that rely only on perceptual judgements. For instance, distinctions in meaning inferred from different expressions could be based on the differing genders of the expressers and not the specific movements or sounds they are producing. While these biases are in part attenuated by the large samples of expressers in the present study, progress in understanding emotional expression rests on deriving more direct measures of the movements or sounds being produced. To this end, we trained a DNN to predict the meanings attributed to vocal bursts solely from imitations of the vocal bursts by randomized speakers from different countries.

Culture-specific emotion regression using a DNN

By incorporating a DNN into our analysis, we were able to link human judgements to underlying acoustics and derive a taxonomy of vocal bursts within and across cultures. DNNs are powerful machine learning algorithms that can approximate many complex natural functions and dynamics, allowing them to accomplish a wide range of tasks. Given that recent advances in computing power have allowed DNNs to approximate or exceed human performance in tasks such as classifying recordings or translating text, DNNs have also emerged as useful scientific models for understanding the psychological processes involved in these domains (for example, speech perception). Specifically, they provide a computational model of human perceptual judgements that are directly tethered to the sensory input. While a DNN is more difficult to interpret than the linear models used more traditionally within psychology (for example, multiple regression and analysis of covariance), it can much more accurately capture the relationship between human perceptual judgements and the sensory input[35–39], as this is a highly nonlinear relationship. The resulting model is directly interpretable in terms of sensory perturbations, which, in the present study, are the acoustic modulations that reliably convey specific emotional meanings. Moreover, after being fit on a suitably large training dataset, the DNN can be used for large-scale inference, thus enabling theories to be tested at a suitable scale to understand the complex, high-dimensional dynamics of a real-world information-processing task. For our purposes—trying to isolate the specific underlying acoustic dimensions of vocal expression that give rise to emotional meaning, while remaining invariant to factors that can bias human judgements—our use of a DNN was essential.

We trained a DNN to extract measures of vocal expression that were invariant to the acoustic properties specific to the person making the expression, such as the nature of their speaking voice and recording context. We did so by tasking it with predicting the average emotion judgements of each seed vocal burst in each culture solely from recordings of participants mimicking each seed vocal burst (Fig. 1). Because the seed vocal bursts were each shown to a random set of participants, this method forced the DNN to remain invariant to factors that were randomized relative to the expression being imitated (acoustic cues to demographics, context or individual variability in mimicry ability). (As a result, identity-related variations in the DNN predictions explain at most about 0.62% of the variance in human perceptual judgements; Supplementary Fig. 4.)

Furthermore, we treated the average emotion judgements within each culture (evaluated in three separate languages) as separate outputs. Thus, the DNN was not provided any prior mapping between emotion and mental state concepts and their use across countries or attempted translations across languages (English, Chinese and Spanish).

Our model architecture was comprised of three parts: (1) a frame-wise feature extractor, where audio features are extracted from each audio frame using a convolutional neural network (CNN); (2) a temporal feature extractor, where features across temporal domain are extracted using a two-layer long-short term memory (LSTM) cell; and (3) a linear layer, which outputs the final predictions (Table 1; Methods section on Deep neural network architecture).

After training, we applied the model to the seed recordings (to which it was not exposed during training). Finally, we applied a multidimensional reliability analysis method to distil the significant shared and culture-specific dimensions of vocal expression uncovered by the model[30]. Specifically, we applied PPCA between the model's

https://doi.org/10.1038/s41562-022-01489-2

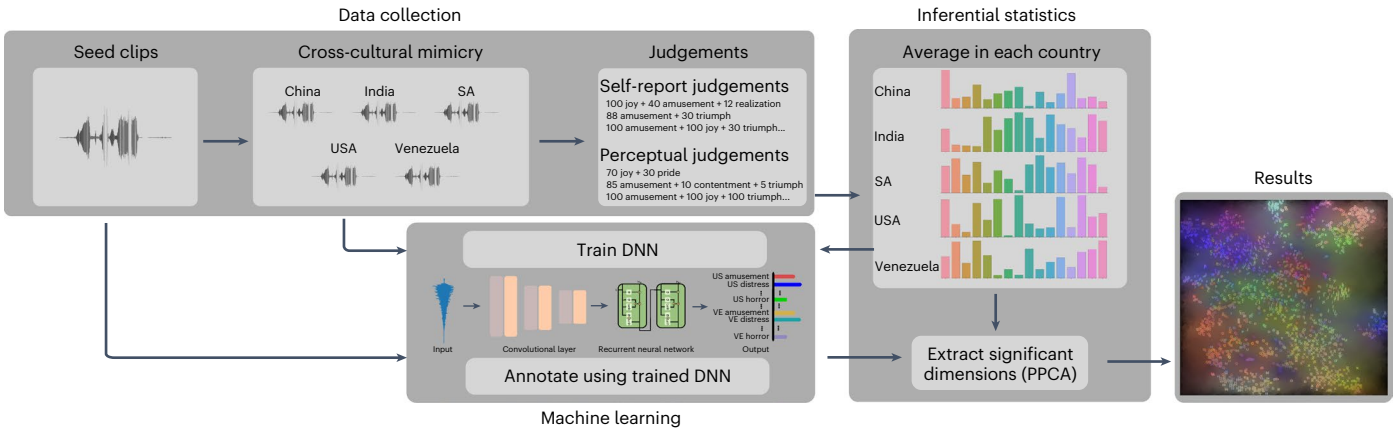

Fig. 1 | Schematic of our experimental and analytic approach. An initial set of 2,756 seed vocal bursts were rated by 8,941 participants in five countries. These participants also used a computer microphone to record themselves mimicking randomly sampled subsets of up to 30 seed vocal bursts. Thus, each of the resulting 282,906 mimic vocal bursts had a corresponding self-report judgement reflecting what the vocal burst meant to the person making it. An additional 7,879 participants provided perceptual judgements of mimic vocal bursts produced in their own country. We used a DNN to find dimensions of vocal expression that had distinct meanings within or across cultures, independent of demographic and contextual cues, by averaging all self-report and perceptual ratings corresponding to each seed vocal burst in each country and tasking the DNN with predicting these averages from the mimic vocal bursts (thereby putting a cost on the model predictions being influenced by the speaking voice of the participant forming the imitation; Methods). Finally, we evaluated the DNN on the seed vocal bursts (to which it had no exposure during training) and compared these predictions to the average human judgements of the seed vocal bursts in each country to extract the dimensions of meaning that the DNN successfully identified in distinct vocal modulations (Methods). SA, South Africa; USA, United States of America; VE, Venezuala.

Table 1 | Convolutional neural network parameters

Layer	Kernel size	Stride	Channels	Activation
Convolution	8	1	64	LeakyReLU
Max-pooling	10	10	—	—
Convolution	6	1	128	LeakyReLU
Max-pooling	8	8	—	—
Convolution	6	1	256	LeakyReLU
Max-pooling	8	8	—	—

culture-specific annotations of the seed recordings and the emotions and mental states actually inferred from the seed recordings by participants in each culture[26,27,29,40]. Given that no prior was built into the model linking the words from different languages to one another, any relationship uncovered between the emotion and mental state concepts across languages using this method implies that the concepts were used similarly to describe the same vocal expression (Fig. 1).

To assess the significance of the dimensions extracted using PPCA, we used a leave-one-out cross-validation method. Specifically, we iteratively performed PPCA between the DNN outputs and the averaged perceptual judgements of all but one of the seed vocal bursts and computed the scores of each dimension extracted by PPCA on the DNN outputs and averaged perceptual judgements of the held-out clips. We used one-tailed partial Spearman correlations to assess the relationship between the held-out DNN output dimensions and human judgements, where for each PPCA dimension we controlled for the PPCA scores on all previous dimensions. After computing *P* values, we used a conservative method of correction for FDR that combined multiple FDR-correction methods (Methods).

Using this method, we uncovered 24 significant dimensions of vocal expression that were reliably associated with distinct meanings (Figs. 2 and 3). More precisely, each of the 24 dimensions corresponds to a pattern of vocal modulation that is reliably associated with a distinct set of emotion and mental state concepts in at least one country or language (Methods section on Extracting significant DNN output dimensions). Most were discovered to have strongly shared meanings across all five countries. It is important to again emphasize that no constraints were put on the DNN to encourage this result (that is, the dimension corresponding to "happiness" in one country could just as easily have been found to correspond to "sadness" in another, even if the two countries speak the same language, if the same vocal modulations in fact had opposite meanings across cultures). In particular, distinct vocal cues were reliably associated with the same 21 emotion concepts or combinations of concepts and their most direct translations across all five countries: adoration/love/sympathy, anger/distress, boredom, concentration/contemplation/calmness, confusion/doubt, pride/triumph, disappointment, disgust, excitement/triumph, fear/horror, horror, interest, joy/amusement, pain, relief, sadness, satisfaction, sexual desire, surprise (positive or negative), admiration/aesthetic appreciation/awe and tiredness.

Of the remaining three significant dimensions of vocal expression we discovered, two were reliably associated with the same three concepts or combinations of concepts and their most direct translations across four out of five countries: awkwardness, craving/interest/satisfaction and realization. The awkwardness dimension was not significantly associated with awkwardness in India (despite being rated in English) but had a wide range of other significant meanings there, including shame (along with 11 other emotion concepts). The craving/interest/satisfaction dimension had significant loadings on at least two of the three terms in each of the five countries. Finally, the realization dimension of vocalization was not significantly associated with realization in India but instead with boredom and distress.

It is worth noting that, among the 21 dimensions whose primary meaning was strongly preserved across all five countries, there were sometimes subtle cultural differences in secondary meanings significantly associated with the same vocal modulations. For instance, the horror (but not fear) dimension was also associated with awe in India and Venezuela, with distress in China and the USA and with amusement and realization in South Africa. These findings probably reflect subtle differences in the meaning of the vocal modulations rather than in the translation of emotion concepts, given that they do not seem to be more common in the non-English-speaking countries (China and Venezuela) than in the English-speaking ones.

https://doi.org/10.1038/s41562-022-01489-2

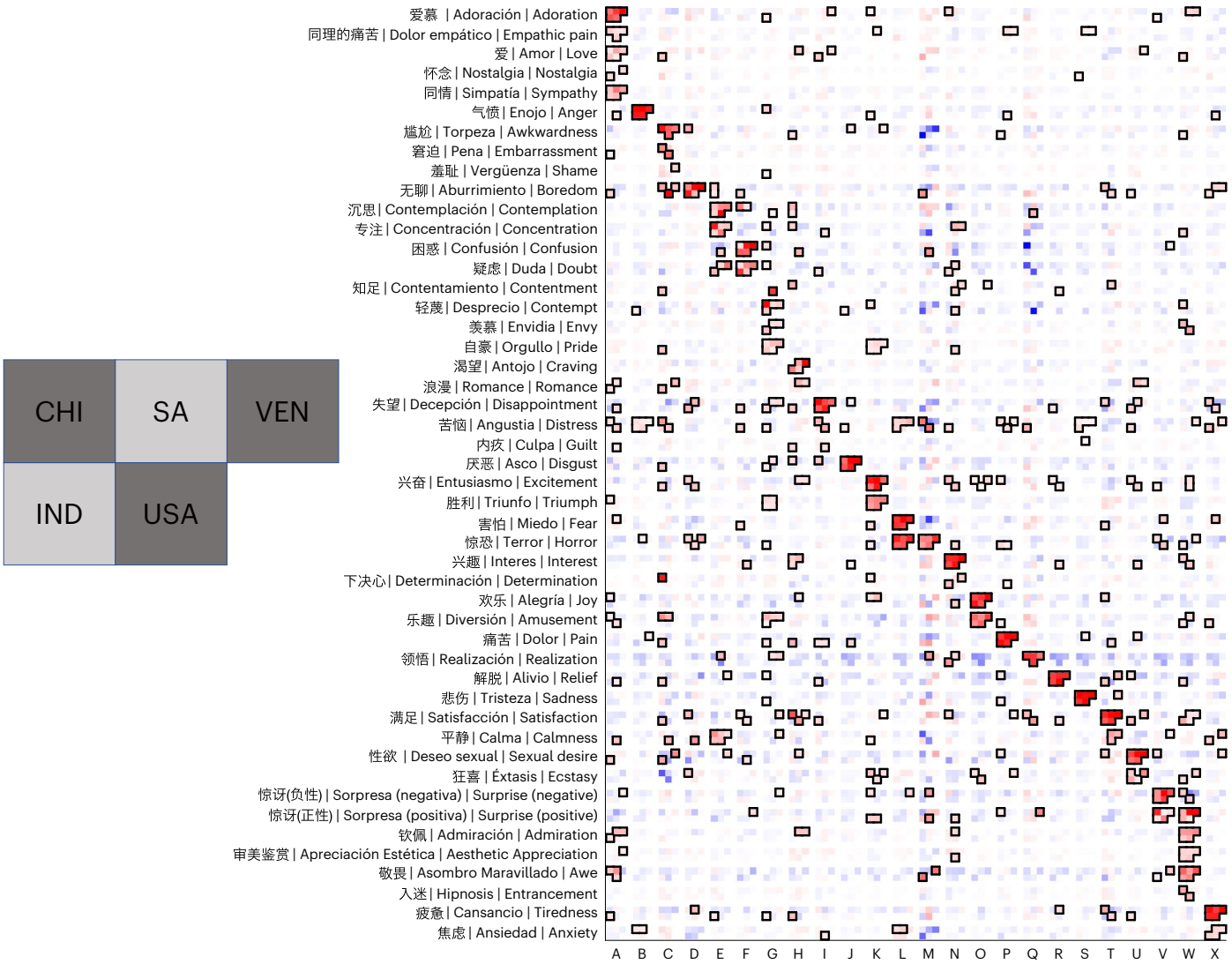

Fig. 2 | Dimensions of vocal expression that emerged as having distinct meanings within or across cultures. The meaning of the 24 dimensions of vocal expression (*x* axis, labeled A–X) that were reliably predicted by the model is captured by loadings on the 48 predicted emotion concepts that people used to judge their own expressions in their own language (*y* axis) in each of the five countries. CHI, China; SA, South Africa; VEN, Venezuala; IND, India; USA, United States of America. We evaluated the DNN on the seed vocal bursts (to which it had no exposure during training) and compared these predictions to the average human judgements of the seed vocal bursts in each country to extract the dimensions of meaning that the DNN successfully identified in distinct vocal modulations. To assess the significance of the dimensions extracted using PPCA, we used a leave-one-out cross-validation method. Specifically, we iteratively performed PPCA between the DNN outputs and the averaged perceptual judgements of all but one of the seed vocal bursts and computed the scores of each dimension extracted by PPCA on the DNN outputs and averaged perceptual judgements of the held-out clips. Finally, we concatenated and correlated the PPCA scores of the held-out DNN outputs and judgements. For each PPCA dimension, we iteratively computed two-tailed partial Spearman correlations between human judgements and DNN model annotations, controlling for the PPCA scores on all previous dimensions. To determine the significance of each dimension, we used a bootstrapping method. After computing *P* values, we used a conservative method of correction for FDR that combined multiple FDR-correction methods. We assessed the significance of the individual loadings of emotion concepts on the extracted dimensions using a bootstrapping method and, for each dimension, we applied a ForwardStop FDR-correction procedure[59] at an alpha of 0.05 to determine the number of significant loadings. Each rectangle is composed of five squares that represent the five countries (as indicated in the bottom left corner). Squares with dark outlines reflect statistically significant correlations between human judgements of the seed vocal bursts in that country and DNN model annotations, with warm colours indicating positive correlations and cool colours indicating negative correlations. Recall that the model was trained to predict judgements in each country (and language) separately. Thus, when multiple countries share statistically significant loadings on similar concepts, it indicates that the dimension of vocal expression has a similar meaning across the countries.

In total, the 24 dimensions of vocal expression—the vocal modulations we found to have reliable meanings somewhere in the world—emerged as having very similar meanings in most places in the world we studied, being 79% preserved overall in both meaning and translation across the five diverse countries where we collected data ($r = 0.89$, $r^2 = 0.79$, country-wise dimension loadings explained by the average loading), leaving the remaining 21% of the variance to be accounted for by differences in meaning across cultures, individual differences, imperfect translation across languages or sampling error (see Supplementary Fig. 5 for breakdown of preserved variance by dimension and country). Where we observed cultural differences, it is clear from an inspection of interactive Fig. 3 that the corresponding vocal expressions were still imitated very similarly across cultures, confirming that our findings reflected

https://doi.org/10.1038/s41562-022-01489-2

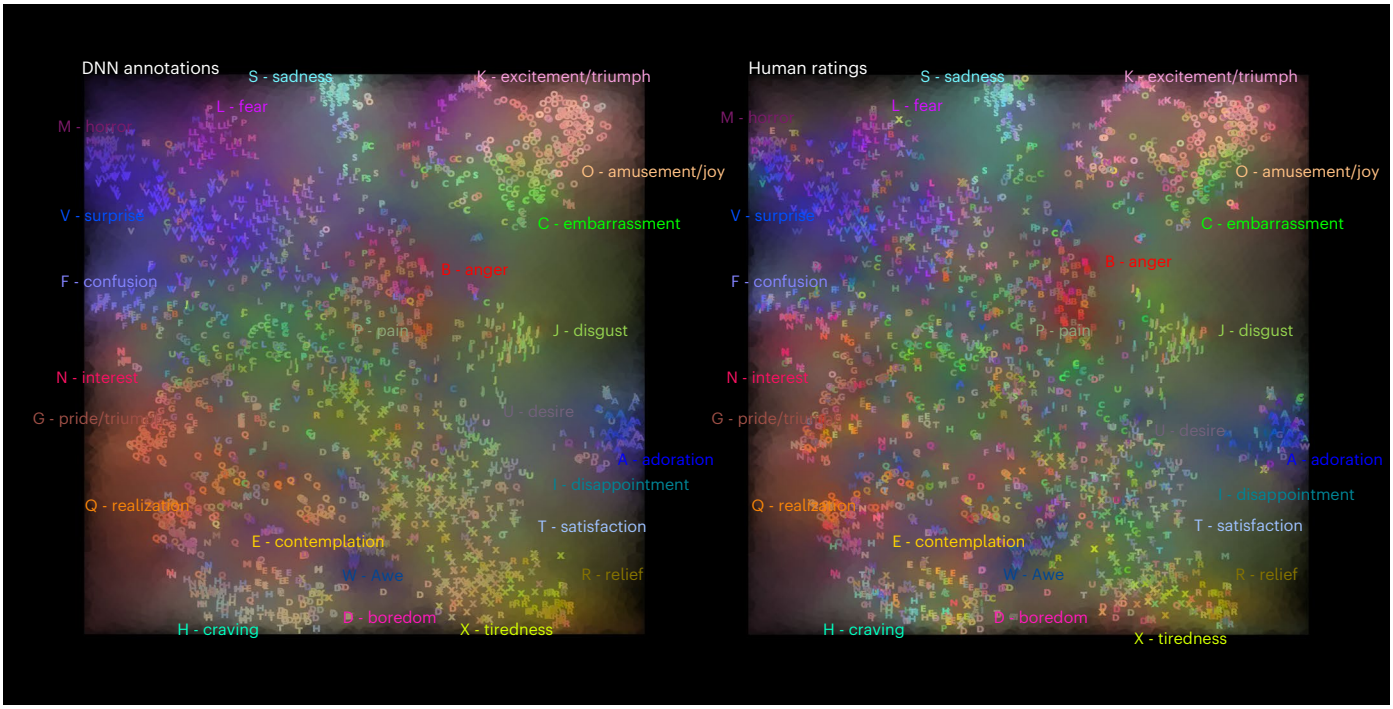

Fig. 3 | Interactive visualization of vocal bursts along the 24 acoustic dimensions of vocal modulation found to have distinct meanings within or across cultures. In an interactive visualization (https://is.gd/iHP0Qf), readers can explore all 282,906 vocal burst mimics, organized by the position of the corresponding seed vocal burst along the 24 dimensions of vocalization we uncovered. We applied *t*-SNE (Methods) to the concatenated human and DNN annotations of the seed vocal bursts to visualize their distribution along the 24 acoustic dimensions of vocal expression that we found to have distinct shared or culture-specific meanings. Here, colours in the plot on the left, as well as in the interactive visualization, represent DNN annotations projected onto the 24 dimensions. Colours in the plot on the right represent average human intensity ratings projected onto the 24 dimensions. Similarity in colour between the two plots reflects similarities between DNN predictions and human perceptual ratings.

differences in the meanings attributed to the underlying vocal expression rather than in the ability to perceive or produce a given set of vocal expression.

Finally, as observed in many previous studies[26,31,41], the emotions attributed to vocal bursts were not discrete but were heterogeneous and varied, reflecting continuous blends of meaning that were readily associated with smoothly varying vocal modulations (Fig. 3).

Robustness and potential for additional dimensions

To assess the robustness of our results, we repeated our analysis for each individual country. That is, for each country, we performed PPCA between human ratings from that country alone and the predictions of the DNN. Despite the reduced statistical power when analysing each country separately, we still extracted between 20 and 23 significant dimensions per country that were qualitatively similar to those we extracted in our main analysis (Supplementary Fig. 6).

We also examined how our results would vary had we extracted fewer or more dimensions, given that the number of significant dimensions identified may depend in part on the sample size (and resulting statistical power) of the dataset. Thus, we computed PPCA loadings for varying numbers of dimensions, from 1 to 95. We found that, as dimensionality increased, shared variance in dimension loadings across the five countries gradually reduced from 95.6% of variance shared along one dimension to 34.8% along 95 dimensions (Supplementary Fig. 7). These results suggest that more granular dimensions of vocal expression may be more culturally variable but could also reflect that the more granular dimensions are more sensitive to sampling error. In either case, the results of this analysis reinforce that the 24 significant dimensions we extracted provide a new lower bound for the complexity of vocal expression across cultures but even more granular dimensions may await further discovery.

Discussion

For decades, scientific approaches to affect and emotion have largely overlooked the critical role non-verbal vocalizations play in human social life (although see refs. [1,2]). Social interactions are structured in dynamic and nuanced ways, although, by cries, sighs, chuckles, shrieks, grunts, growls, oohs and ahhs. Here, by combining large-scale experimental data collection and machine learning, we uncover a 24-dimensional semantic space of emotions perceived from brief vocalizations across cultures. This work reveals the nuanced meanings of vocal bursts across five distinct cultures and adds to a growing literature establishing that emotional behaviours are high-dimensional, departing from traditional emotion models that posit a small number of discrete categories or affective dimensions[30,41]. The 24 dimensions of vocal expression that we uncovered by applying machine learning to over 280,000 vocal bursts were 79% preserved in meaning across the five countries and three languages we studied, with 21 dimensions showing a high degree of shared meaning and the remainder showing varying degrees of cultural specificity.

This study is not without its limitations. It is worth noting that this is not an exhaustive catalogue or taxonomy of distinct emotion concepts or acoustically distinct vocalizations around the world but the most comprehensive description to date of the distinct meanings that vocal expressions can reliably convey in a wide range of countries. The 2,756 vocal expressions that participants imitated came from multiple datasets encompassing both posed and spontaneous expressions across many contexts and cultures (Methods) but do not capture all possible non-verbal utterances. Moreover, the experimental procedure—requiring participants to intentionally mimic vocalizations—could have resulted in vocalizations with subtle differences to those made in everyday life. The concepts participants relied upon to describe the meaning of the expressions—despite

encompassing the widest range of distinct emotions for which there is evidence, to our knowledge, across any modality of emotional behaviour[30] and having previously been found to explain attributions of valence, arousal and a wide range of other proposed dimensions of appraisal and motivation[42]—may still omit other nuanced meanings about mental states and the social context that vocal expressions can potentially convey. And although the three languages included in the present study (English, Spanish and Mandarin Chinese) are spoken by ~40% of the world population (by ~25% as a first language), there are 190 other countries and ~6,500 other languages that we did not study, across which there are almost certainly other culture-specific vocal expressions. Further work is also needed to examine how vocal expression varies as a function of gender, social class and other sources of individual identity and variation.

The present study used an experimental approach to control for factors other than vocal modulation that may affect how people perceive vocal bursts. In particular, we asked 8,941 participants to each mimic a random subset of 2,756 wide-ranging vocal bursts and then trained a DNN to predict the average meanings inferred from vocal bursts solely from their imitations. This strategy required the DNN to decouple the vocal modulations underlying each vocal burst from the identity or the context of the person forming them. We also modelled cultural differences by predicting average meanings in each country separately, departing from studies that assume a one-to-one mapping of emotion concepts across cultures. Using this deeply inductive approach, we discovered how the meanings of vocal bursts converge and diverge across cultures.

Our findings relied on emotional mimicry, which a large body of work suggests is a pervasive element of human social life that structures our interactions from a very early age. Our results confirm that the human ability to mimic vocalizations extends to intentional mimicry, with most of the mimicked vocal bursts we collected being qualitatively similar to the original seed vocal bursts, except for differences in participants' vocal traits and recording quality (for examples, see Supplementary Table 3). Thus, we were able to produce an experimentally controlled set of vocal bursts suitably large for both training machine learning models and performing large-scale psychological inference. However, it is important to note that participants were instructed to mimic vocal bursts such that they would convey similar emotions to the original seed samples, meaning that the mimicry process used here differs from spontaneous mimicry observed in daily life. For instance, participants could have followed the instructions either by imitating the sound or by producing a sound that they thought would convey the same emotional meaning. Further research is needed to understand the processes underlying spontaneous and intentional emotional mimicry (but see refs. [43,44]).

Notably, our methods and findings are not necessarily specific to the emotion domain and converge with other research showing that non-linguistic vocalizations can convey a wide range of informational content (that is, information related to objects or threats in the environment, as well as speakers' intentions and cognitive states) and that, even when participants are asked to produce new vocalizations, the inferred meanings of the sounds are often shared across different cultures and languages[45]. This lends support to the idea that vocalizations can represent a broad range of meanings in the absence of language and that vocalizations such as vocal bursts may serve as an important medium for cross-cultural communication between people who lack a common language.

Our findings are consistent with semantic space theory (SST), which conceives of emotions as dimensions of a continuous, high-dimensional state space that explains the systematic variation in emotional behaviour and physiology[30]. Consistent with SST, vocal bursts were found to be (1) high-dimensional, with cross-cultural similarities that could not be reduced to a few dimensions (Fig. 2) and (2) continuous, with smooth gradients corresponding to smooth variations in meaning (Fig. 3). The 24 dimensions of vocal expression that we uncover were, across four or more cultures, associated with admiration/aesthetic appreciation/awe, adoration/love/sympathy, anger/distress, awkwardness/embarrassment, boredom, concentration/contemplation/calmness, confusion/doubt, craving/interest/satisfaction, disappointment, disgust, excitement/triumph, fear/horror, horror only, interest, joy/amusement, pain, pride/triumph, realization, relief, sadness, satisfaction, sexual desire, surprise (positive or negative) and tiredness. These dimensions of meaning largely explain the dimensions of vocal expression found to be recognized as distinct in smaller-scale studies[1,4,13,16,17,26,46]. They also largely overlap with those found to be distinguished in facial expression, encompassing the distinct facial expressions that have been found to occur in similar contexts worldwide[47] and to be depicted in consistent contexts in ancient American sculptures[28], providing further evidence that a wide range of emotions may have associated expressions with shared emotional meanings across cultures. Furthermore, they largely encompass the dimensions of evoked emotional experience found to correspond to distinctive modes of neural activity in multimodal, associative brain areas[48]. Together, these findings support what is emerging as a core tenet of SST: that one's position along continuous dimensions of emotional experience helps tune the nervous system and, via expressive behaviour, those of social observers to respond efficiently and collectively to ongoing challenges and opportunities in the environment[30].

Methods

All participants provided informed consent and all aspects of the study design and procedure were approved by Heartland IRB (HIRB project no. 031221–315).

Procedure

Participants from China (n = 380; 213 female; mean age = 23.42 yr), India (n = 377; 78 female; mean age = 26.67 yr), South Africa (n = 1,155; 712 female; mean age = 26.81 yr), the USA (n = 1,492; 762 female; mean age = 37.03 yr) and Venezuela (n = 203; 65 female; mean age = 29.97 yr) were recruited for the mimicry phase of the experiment via psychology recruitment email lists compiled by the authors and via a range of crowdsourcing platforms (Amazon Mechanical Turk, Clickworker, Prolific, Microworkers and RapidWorker). These five countries were selected because they are widely diverse in terms of culture-related values—for example individualism versus collectivism, power distance, autonomy—of interest in cross-cultural comparisons[49].

In each trial, participants heard a seed vocal burst and were instructed to use their computer microphone to record themselves mimicking the vocal burst such that their imitation would be perceived to convey similar emotions to the original recording. The rating and recording interfaces were presented on the same survey page for convenience, so participants could choose to rate each vocal burst before or after imitating it. Participants completed 30 trials per survey and could complete multiple versions of the survey, up to ten depending on the country. See Supplementary Table 1 and Supplementary Fig. 1 for further information on the breakdown of survey responses by demographics and country. Data collection was completed between April 2021 and December 2021. To collect a large-scale dataset, we released up to 300 versions of the survey for completion per country per day. The number of speakers and samples per country is a function of the number of participants who opted to participate and the number of surveys they completed over time.

The seed vocal bursts consisted of wide-ranging vocal bursts produced in laboratory settings in five countries (n = 2,032) (see ref. [17], vocal bursts produced in ecological settings gathered in online video (n = 48)[17] and newly assembled vocal bursts extracted at large from assorted Chinese (n = 86) and Japanese (n = 540) media by authors A.C. and X.F.

We excluded any recordings below 5 kilobytes (KB) (1.7% of recordings) or above 250 KB (0.0018%) in file size, which consistently indicated recording errors (empty or silent recordings, or incorrect start or end times). All other recordings gathered during the mimicry phase of the experiment were used as stimuli in the rating-only phase of the experiment. Participants in the rating-only phase of the experiment (China (n = 277; 159 female; mean age = 23.07 yr), India (n = 438; 194 female; mean age = 28.07 yr), South Africa (n = 1,401; 893 female; mean age = 26.92 yr), the USA (n = 2,475; 1,337 female; mean age = 37.14 yr) and Venezuela (n = 189; 70 female; mean age = 29.45 yr)) were given the option of responding that no vocal burst was present in the clip. Note that for the purposes of the present study, we intentionally exercised minimal quality control measures, given that (1) the data were generally found to be of high quality and (2) exclusion of data can bias results[50,51].

The 48 emotions used to rate each vocal burst were derived from a comprehensive examination of the meanings vocal bursts have been previously posited to convey[13,15–17,46,52] and the words that people regularly use to describe emotion-related experiences and expressions[23,29–31]. In both phases, participants listened to each vocal burst and were asked, "What emotions is this person feeling? Select all that apply". For each emotion selected, participants were then instructed to "rate the intensity of the emotion on a 1–100 scale".

PPCA

PPCA is a dimensionality reduction and statistical modelling technique that decomposes two high-dimensional datasets measuring the same attributes and finds linear combinations of attributes that statistically maximize the covariance explained across datasets[26,27,29,40]. In particular, like more established methods such as partial least-squares correlation analysis (PLSC) and canonical correlation analysis (CCA), PPCA examines the cross-covariance between datasets rather than the variance–covariance matrix within a single dataset. However, whereas PLSC and CCA derive two sets of latent variables, α and β, maximizing cov($X\alpha i$,$Y\beta i$) or corr($X\alpha i$,$Y\beta i$), PPCA derives only one variable: α. For an extended validation of the method including a mathematical proof, please refer to ref. [26].

Generalized PPCA

PPCA was initially conceived to maximize the covariance explained across two datasets. To apply this technique to five different datasets measuring the same 48 attributes, we developed G-PPCA which extracts linear combinations of attributes that maximally covary across three or more datasets (in this case, emotion judgements from five countries). In particular, G-PPCA maximizes the objective function sum(cov(αX,αY) for X,Y in S) where S is the set of all possible pairwise combinations of datasets.

We iteratively applied G-PPCA in a leave-one-stimulus-out manner to extract components from the judgements of all but one stimulus and then projected each country's ratings of the left-out stimulus onto the extracted components, resulting in cross-validated component scores for each country and stimulus.

G-PPCA significance testing

To determine the statistical significance of each component, we sought to ensure that the extracted dimensions not only reflect shared structure preserved across all five countries but also reflect significantly preserved dimensions across pairs of countries in the analysis. First, we calculated the one-sided partial Spearman correlation between corresponding component scores for each country pair, iteratively partialling out each previous component and calculating statistical significance separately for each dimension. One-sided tests were used since we were specifically interested in dimensions that were preserved across countries. Within country pairs, P values for the 48 dimensions were FDR-corrected using the Benjamini–Hochberg procedure. Group-level statistical significance of the generalized dimensions was determined by representing each dimension as a graph with countries as nodes and binary statistical significance (P < 0.05, FDR-corrected) as edges. We then retained dimensions whose statistical significance graph was not bipartite (that is, could not be partitioned—in this case meaning that for any given dimension the component scores for any given country were significantly positively correlated with the corresponding component scores for at least two other countries).

Factor rotation

To interpret and visualize the preserved components, we applied varimax factor rotation. Varimax is a factor rotation method that minimizes the number of factors needed to explain a variable, simplifying the structure of the factor matrix and making dimensions more interpretable. We used an implementation of varimax available in Python's statsmodels package (https://www.statsmodels.org/stable/generated/statsmodels.multivariate.factor_rotation.rotate_factors.html).

t-SNE

We also sought to establish and visualize how the dimensions of vocal emotion perception are distributed. As in previous work, we approached this by visualizing the data using t-distributed stochastic neighbour embedding (t-SNE; ref. [53]). This method projects high-dimensional data onto two nonlinear axes, such that the local distances between data points are accurately preserved while more distinct data points are separated by longer, more approximate, distances.

Deep neural network architecture

For our purposes, we use an end-to-end learning model[54], which takes the raw waveform as input and outputs predictions of the 48 × 5 country-wise average emotional expression ratings. The model architecture is comprised of three parts: (1) a CNN that extracts audio features from a single audio frame, (2) a recurrent neural network (RNN) that captures the temporal dynamics of the input signal (across the audio frames) and (3) a regressor that outputs the final predictions. The CNN consists of three layers of convolution and max-pooling operations, the RNN is a two-layer LSTM cell with 256 hidden units and the regressor is a linear layer of 240 dimensions. To prevent the model from overfitting, we use batch normalization after each convolution layer. Table 1 outlines the parameters of the audio architecture.

Model training

To train our model, we labelled each mimic vocal burst with the average of all self-report judgements of the seed vocal burst being imitated and all perceptual judgements of all the other imitations of that seed vocal burst, in each country. This required the model to learn the meaning of vocal modulations that were shared between the mimic vocal bursts and their respective seed vocal burst. Correspondingly, it put a cost on the model predictions being influenced by the speaking voice of the participant forming the imitation, since on average >99% of the judgements contributing to the average judgement were judgements by other people of other people's voices. Furthermore, averaging the self-report and perceptual ratings corresponding to each mimic vocal burst encouraged the model to learn aspects of vocal modulation that reflect both the intended meaning of a vocal burst and its perceived meaning.

We trained the model to predict judgements in each country separately from each mimic vocal burst, including judgements originating in countries different from that of the mimic vocal burst. Thus, in total, the model learned to predict 240 outputs: 48 outputs per country from the USA, Venezuela, China, India and South Africa. This training strategy took advantage of the similarity and high fidelity of imitations across all countries (interactive visualization: https://is.gd/iHP0Qf) to learn the culture-specific meanings of vocal modulations.

(Differences in how vocal bursts were imitated in different countries seemed to be rare but even where such differences occurred, it is likely that the model would learn the culture-specific meanings associated with the vocal modulations that differed across countries through a form of weak supervision[55].)

During model training, we used the Adam optimization algorithm[56] with an initial learning rate of 10^{-4} and batch size of 32 samples. The model was trained from scratch, meaning that the weights were randomly initialized using the Kaiming uniform method[57], with the biases in all layers set to zero. We tuned the hyper-parameters of the model by performing a random search, following ref. [56]. The model was trained with an early stopping strategy to avoid overfitting the training samples. Across all hyper-parameter configurations tested, the best-performing model on the validation set (with respect to our metric function, the concordance correlation coefficient (CCC), described below) was selected for further analysis.

The input audio signal was resampled to 16 kHz and segmented into frames of 0.1 s before feeding to the network. We partitioned the dataset into training (80%) and validation (20%) sets in a subject-independent manner, meaning each subject was uniquely assigned to one of the sets. To test the efficacy of our approach, we used the seed vocal burst samples as our test set.

As our objective function in the optimization process we used the mean squared error, which mathematically, given two variables x and y and N data points, is defined as follows:

$$L_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

For our purposes we trained our networks to predict 240 outputs and as such we define the overall loss function as:

$$L = \frac{1}{240}\sum_{i=1}^{240} L^i_{\mathrm{MSE}}$$

As our metric function we used the CCC which is used to select the model that has the highest performance on the validation set. Mathematically, given two variables x and y, CCC is defined as follows:

$$\rho_c = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where $\sigma_x^2 = \mathrm{var}(x), \sigma_y^2 = \mathrm{var}(y)$, $\sigma_{xy}^2 = \mathrm{cov}(x,y)$, $\mu_x = E(x)$ and $\mu_y = E(y)$. The overall score, across each of the 240 predictions, is given by:

$$S = \frac{1}{240}\sum_{i=1}^{240}\rho_c^i$$

Extracting significant DNN output dimensions

To identify dimensions of vocal expression captured by the DNN that were reliably associated with distinct meanings in one or more cultures, we applied PPCA between the 240 outputs of the DNN applied to the seed recordings and the 240 averaged perceptual judgements of the seed vocal bursts (ratings of 48 emotion concepts averaged within each of five countries). This analysis captures the acoustic dimensions of country-specific perceptual judgements of naturalistic vocal bursts. The seed vocal bursts were well-suited for this analysis because each was judged an average of 20.5 times per country (102.7 times in total), providing sufficient data for statistical testing using robust non-parametric approaches.

To assess the significance of the dimensions extracted using PPCA, we used a leave-one-out cross-validation method. Specifically, we iteratively performed PPCA between the DNN outputs and the averaged perceptual judgements of all but one of the seed vocal bursts and computed the scores of each dimension extracted by PPCA on the DNN outputs and averaged perceptual judgements of the held-out clips. Finally, we concatenated and correlated the PPCA scores of the held-out DNN outputs and judgements. To control for nonlinear monotonic dependencies between extracted dimensions, we used partial Spearman correlations, where for each PPCA dimension we controlled for the PPCA scores on all previous dimensions. To determine the significance of each dimension, we used a bootstrapping method, iteratively repeating the correlation procedure while randomly resampling the seed vocal bursts (1,000 iterations with replacement). P values were taken as one minus the proportion of times that the correlation exceeded zero across resampling iterations.

After computing P values, we used a conservative method of correction for FDR that combined multiple FDR-correction methods. Specifically, we used Benjamini–Hochberg FDR correction[58] across the first 48 PPCA dimensions (as we were interested in variations of 48 potentially distinct emotion concepts and their translations across countries) at an α of 0.05. We also separately performed a ForwardStop sequential FDR-correction procedure[59]. Finally, we determined the signal-to-noise ratio of the correlations corresponding to each PCA dimension (the correlation divided by the standard deviation computed using bootstrapping (see above)) and applied a threshold of 3 to the signal-to-noise ratio to extract more stable dimensions. We only kept dimensions that met all three of these criteria. We applied factor rotation using the varimax criterion[60] to these dimensions.

To assess the significance of the individual loadings of emotion concepts on the extracted dimensions, we used a bootstrapping method. Specifically, we performed the entire PPCA analysis repeatedly after resampling the seed vocal bursts with replacement, extracting the significant dimensions and performing factor analysis each time. For each dimension, we then tested the significance of the top n loadings, with n varying from 1 to 240, by determining how often, across resampling the iterations, there existed a dimension with all of these top n loadings pointing in the same direction. This estimates the proportion of times a dimension with these coloadings would be extracted if we repeated the entire study. We took one minus this proportion as the P value. As n varies from 1 to 240, the P value can only increase because more loadings are included in the test (and therefore the probability of all loadings pointing in the same direction decreases monotonically). For each dimension, we applied a ForwardStop FDR-correction procedure[59] at an α of 0.05 to determine the number of significant loadings.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data associated with this manuscript are available upon reasonable request to the corresponding authors.

Code availability

Code associated with this study, including the functions to perform PPCA, is available in the following Zenodo repository: https://doi.org/10.5281/zenodo.7111972.

References

1. Banse, R. & Scherer, K. R. Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* **70**, 614–636 (1996).
2. Fernald, A. in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (eds Barkow, J. et al.) 391–428 (Oxford Univ. Press, 1992).
3. Soltis, J. The signal functions of early infant crying. *Behav. Brain Sci.* **27**, 443–458 (2004).

4. Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D. & Flynn, L. M. The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion* **16**, 117–128 (2016).
5. Keltner, D. & Kring, A. M. Emotion, social function, and psychopathology. *Rev. Gen. Psychol.* **2**, 320–342 (1998).
6. Van Kleef, G. A., De Dreu, C. K. W. & Manstead, A. S. R. An interpersonal approach to emotion in social decision making: the emotions as social information model. *Adv. Exp. Social Psychol.* **42**, 45–96 (2010).
7. Bryant, G. A. in *The Handbook of Communication Science and Biology* (eds Floyd, K. & Weber, R.) 63–77 (Routledge, 2020).
8. Snowdon, C. T. in *Handbook of Affective Sciences* (eds Davidson, R. J. et al.) 457–480 (Oxford Univ. Press, 2003).
9. Wu, Y., Muentener, P. & Schulz, L. E. One- to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proc. Natl Acad. Sci. USA* **114**, 11896–11901 (2017).
10. Vouloumanos, A. & Bryant, G. A. Five-month-old infants detect affiliation in colaughter. *Sci. Rep.* **9**, 4158 (2019).
11. Smoski, M. & Bachorowski, J.-A. Antiphonal laughter between friends and strangers. *Cogn. Emot.* **17**, 327–340 (2003).
12. Bryant, G. A. et al. Detecting affiliation in colaughter across 24 societies. *Proc. Natl Acad. Sci. USA* **113**, 4682–4687 (2016).
13. Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl Acad. Sci. USA* **107**, 2408–2412 (2010).
14. Gendron, M., Roberson, D., van der Vyver, J. M. & Barrett, L. F. Cultural relativity in perceiving emotion from vocalizations. *Psychol. Sci.* **25**, 911–920 (2014).
15. Scherer, K. R. in *Emotions in Personality and Psychopathology* (ed. Izard, C. E.) 493–529 (Springer, 1979).
16. Laukka, P. et al. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Front. Psychol.* **4**, 353 (2013).
17. Cowen, A. S., Elfenbein, H. A., Laukka, P. & Keltner, D. Mapping 24 emotions conveyed by brief human vocalization. *Am. Psychol.* **74**, 698–712 (2019).
18. Jolly, E. & Chang, L. J. The flatland fallacy: moving beyond low-dimensional thinking. *Top. Cogn. Sci.* **11**, 433–454 (2019).
19. Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychol. Sci.* **26**, 354–356 (2015).
20. Whiting, C. M., Kotz, S. A., Gross, J., Giordano, B. L. & Belin, P. The perception of caricatured emotion in voice. *Cognition* **200**, 104249 (2020).
21. Monroy, M., Cowen, A. S. & Keltner, D. Intersectionality in emotion signaling and recognition: the influence of gender, ethnicity, and social class. *Emotion* https://doi.org/10.1037/emo0001082 (2022).
22. Jackson, J. C. et al. Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
23. Rozin, P. & Cohen, A. B. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion* **3**, 68–75 (2003).
24. Hejmadi, A., Davidson, R. J. & Rozin, P. Exploring Hindu Indian emotion expressions: evidence for accurate recognition by Americans and Indians. *Psychol. Sci.* **11**, 183–186 (2000).
25. Russell, J. A., Suzuki, N. & Ishida, N. Canadian, Greek, and Japanese freely produced emotion labels for facial expressions. *Motiv. Emot.* **17**, 337–351 (1993).
26. Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R. & Keltner, D. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nat. Hum. Behav.* **3**, 369–382 (2019).
27. Cowen, A. S., Fang, X., Sauter, D. & Keltner, D. What music makes us feel: at least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl Acad. Sci. USA* **117**, 1924–1934 (2020).
28. Cowen, A. S. & Keltner, D. Universal facial expressions uncovered in art of the ancient Americas: a computational approach. *Sci. Adv.* **6**, eabb1005 (2020).
29. Demszky, D. et al. GoEmotions: a dataset of fine-grained emotions. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 4040–4054 (ACL, 2020).
30. Cowen, A. S. & Keltner, D. Semantic space theory: a computational approach to emotion. *Trends Cogn. Sci.* **25**, 124–136 (2021).
31. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl Acad. Sci. USA* **114**, E7900–E7909 (2017).
32. Cordaro, D. T. et al. The recognition of 18 facial-bodily expressions across nine cultures. *Emotion* **20**, 1292–1300 (2020).
33. Cordaro, D. T. et al. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* **18**, 75–93 (2018).
34. Keltner, D., Sauter, D., Tracy, J. & Cowen, A. Emotional expression: advances in basic emotion theory. *J. Nonverbal Behav.* **43**, 133–160 (2019).
35. Peterson, J. C., Abbott, J. T. & Griffiths, T. L. Adapting deep network features to capture psychological representations. In *Proc. of the 48th Annual Conference of the Cognitive Science Society* 2363–2368 (2016).
36. Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A. & Suchow, J. W. Deep models of superficial face judgments. *Proc. Natl Acad. Sci. USA* **119**, e2115228119 (2022).
37. Peters, B. & Kriegeskorte, N. Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* **5**, 1127–1144 (2021).
38. Storrs, K. R., Anderson, L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021). https://doi.org/10.1101/2020.04.07.026120
39. Lake, B. M., Zaremba, W., Fergus, R. & Gureckis, T. M. Deep neural networks predict category typicality ratings for images. In *Proc. 37th Annual Meeting of the Cognitive Science Society* (eds Noelle, D. C. et al.) 1243–1248 (The Cognitive Science Society, 2015); https://cogsci.mindmodeling.org/2015/papers/0219/paper0219.pdf
40. Cowen, A. S. & Keltner, D. Universal emotional expressions uncovered in art of the ancient Americas: a computational approach. *Sci. Adv.* **6**, eabb1005 (2020).
41. Cowen, A., Sauter, D., Tracy, J. L. & Keltner, D. Mapping the passions: toward a high-dimensional taxonomy of emotional experience and expression. *Psychol. Sci. Public Interest* **20**, 69–90 (2019).
42. Cowen, A. S. & Keltner, D. What the face displays: mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol.* **75**, 349–364 (2020).
43. Hess, U. & Fischer, A. Emotional mimicry: why and when we mimic emotions. *Soc. Pers. Psychol. Compass* **8**, 45–57 (2014).
44. Fischer, A. & Hess, U. Mimicking emotions. *Curr. Opin. Psychol.* **17**, 151–155 (2017).
45. Ćwiek, A. et al. Novel vocalizations are understood across cultures. *Sci. Rep.* **11**, 10108 (2021).
46. Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L. & Abramson, A. The voice conveys specific emotions: evidence from vocal burst displays. *Emotion* **9**, 838–846 (2009).
47. Cowen, A. S. et al. Sixteen facial expressions occur in similar contexts worldwide. *Nature* **589**, 251–257 (2021).
48. Horikawa, T., Cowen, A. S., Keltner, D. & Kamitani, Y. The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *iScience* **23**, 101060 (2020).

49. Hofstede, G. Dimensionalizing cultures: the Hofstede model in context. *Online Read. Psychol. Cult.* **2**, 8 (2011).
50. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
51. Roth, P. L. Missing data: a conceptual review for applied psychologists. *Pers. Psychol.* **47**, 537–560 (1994).
52. Juslin, P. N. & Laukka, P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* **129**, 770–814 (2003).
53. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
54. Tzirakis, P., Zhang, J. & Schuller, B. W. End-to-end speech emotion recognition using deep neural networks. In *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5089–5093 (IEEE, 2018). https://doi.org/10.1109/icassp.2018.8462677
55. Zamani, H. & Croft, W. B. On the theory of weak supervision for information retrieval. In *Proc. 2018 ACM SIGIR International Conference on Theory of Information Retrieval* 147–154 (Association for Computing Machinery, 2018).
56. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations (ICLR)* (ICLR, 2015).
57. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE, 2015).
58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
59. G'Sell, M. G., Wager, S., Chouldechova, A. & Tibshirani, R. Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. B* **78**, 423–444 (2016).
60. Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958).

Acknowledgements

This work was supported by Hume AI as part of its effort to advance emotion research using computational methods. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.S.C. and D.K. designed the experiment. L.K., M.O., X.F., M.M., R.C., J.M. and A.S.C. implemented the study design and collected data. J.A.B., P.T., A.B. and A.S.C. analysed data. J.A.B. and A.S.C. interpreted results and created figures. J.A.B. and A.S.C. drafted the manuscript. All authors provided critical revisions and approved the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01489-2.

Correspondence and requests for materials should be addressed to Jeffrey A. Brooks or Alan S. Cowen.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Jeffrey A. Brooks
Alan S. Cowen

Last updated by author(s): 9/27/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collected was conducted using web-based tasks implemented using custom and proprietary code.
Data analysis	Data analysis was conducted using custom code written in Python 3.9 and MATLAB 2021a. Key analysis functions and scripts are available in the following Zenodo repository: 10.5281/zenodo.7111972

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data associated with this manuscript are available upon reasonable request to the corresponding authors.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A large-scale quantitative mimicry-based experimental investigation into emotional vocalizations
Research sample	Participants in the mimicry phase of the experiment were 4,031 individuals from China (n = 380; 213 female; mean age = 23.42), India (n = 377; 78 female; mean age = 26.67), South Africa (n = 1,155; 712 female; mean age = 26.81), the United States (n = 1,492; 762 female; mean age = 37.03), and Venezuela (n = 203; 65 female; mean age = 29.97). Participants in the rating-only phase of the experiment were 4,780 individuals from China (n = 277; 159 female; mean age = 23.07), India (n = 438; 194 female; mean age = 28.07), South Africa (n = 1,401; 893 female; mean age = 26.92), the United States (n = 2,475; 1,337 female; mean age = 37.14), and Venezuela (n = 189; 70 female; mean age = 29.45). These five countries were selected to be highly representative and widely diverse in terms of culture-related values – e.g. individualism vs. collectivism, power distance, autonomy – of interest in cross-cultural comparisons.
Sampling strategy	Participants were randomly sampled from crowdsourcing platforms (see Recruitment, below). To collect a large-scale dataset, we released up to 300 versions of the survey for completion per country per day. The number of speakers and samples per country is a function of the number of participants who opted to participate and the number of surveys they completed over time.
Data collection	All data was collected using remote web-based computer tasks which participants were instructed to complete without anyone else in the room. In the mimicry phase of the experiment, stimulus ratings as well as participant-recorded vocalizations (which participants were instructed to record using their computer's microphone) were collected. In the rating-only phase of the experiment, stimulus ratings were collected.
Timing	Data was collected between April, 2021 and December, 2021.
Data exclusions	We excluded any recordings below 5 KB (1.7% of recordings) or above 250 KB (0.0018%) in file size, which consistently indicated recording errors (empty or silent recordings, or incorrect start or end times).
Non-participation	No participants who provided informed consent subsequently dropped out or declined participation.
Randomization	No experimental groups were created. Participants in the mimicry phase of the experiment completed 30 trials per survey (each reflecting a random sample from over 2,500 vocal bursts) and could complete multiple versions of the survey. Participants in the ratings-only phase of the experiment could complete as many ratings (randomized stimuli from the mimicry phase of the experiment) and completed an average of 77.1 trials.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants in both phases of the experiment were recruited via psychology recruitment email lists compiled by the authors and via a range of crowdsourcing platforms (Amazon Mechanical Turk, Clickworker, Prolific, Microworkers, and RapidWorker). The number of speakers and samples per country is a function of the number of participants who opted to participate and the number of surveys they completed over time. While all participants are self-selected due to interest and motivation to participate in research studies, this is unlikely to introduce bias into the results due to the scale and diversity of the sample. Further, studies have shown that online workers provide high-quality data that yield replicable findings (Piolacci & Chandler, 2014; Current Directions in Psychological Science).
Ethics oversight	The study protocol was approved by Heartland commercial IRB (HIRB Project No. 031221-315).

Note that full information on the approval of the study protocol must also be provided in the manuscript.