Human Listeners' Perception of Behavioural Context and Core Affect Dimensions in

Chimpanzee Vocalizations

Roza G. Kamiloğlu[1], Katie E. Slocombe[2], Daniel B. M. Haun[3], Disa A. Sauter[1]

[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands;

[2]Department of Psychology, University of York, York, UK; [3]Max Planck Institute for

Evolutionary Anthropology, Leipzig, Germany

Author Note: Correspondence concerning this article should be addressed to Roza G.

Kamiloğlu, Department of Psychology, University of Amsterdam, REC G, Nieuwe

Achtergracht 129B, 1001 NK, Amsterdam, The Netherlands.

Contact: r.g.kamiloglu@uva.nl

Abstract

Vocalizations linked to emotional states are partly conserved among phylogenetically related species. This continuity may allow humans to accurately infer affective information from vocalizations produced by chimpanzees. In two pre-registered experiments, we examine human listeners' ability to infer behavioural contexts (e.g., discovering food) and core affect dimensions (arousal and valence) from 155 vocalizations produced by 66 chimpanzees in 10 different positive and negative contexts at high, medium, or low arousal levels. In Experiment 1, listeners ($n = 310$), categorised the vocalizations in a forced-choice task with 10 response options, and rated arousal and valence. In Experiment 2, participants ($n = 3120$) matched vocalizations to production contexts using Yes/No response options. The results show that listeners were accurate at matching vocalizations of most contexts in addition to inferring arousal and valence. Judgments were more accurate for negative as compared to positive vocalizations. An acoustic analysis demonstrated that, listeners made use of *brightness* and *duration* cues, and relied on *noisiness* in making context judgements, and *pitch* to infer core affect dimensions. Overall, the results suggest that human listeners can infer affective information from chimpanzee vocalizations beyond core affect, indicating phylogenetic continuity in the mapping of vocalizations to behavioural contexts.

Keywords: arousal, behavioural context, chimpanzee, human, valence, vocalization

When we hear a hissing cat or a person laughing, we may be able to infer information from these vocalizations, including both the individual's affective state and the kind of situation they are in. In 1872, Darwin [1] hypothesised that emotional vocal expressions have ancient evolutionary roots and that they are based on shared mechanisms across mammalian species. In the research on phylogenetic continuity of emotional vocalizations that has followed since then, researchers have primarily focused on vocal production; this work has established considerable similarities in the acoustic features linked to affective information in different animal groups. In an extensive review, Briefer [2] notes consistent acoustic correlates of core affect dimensions such as arousal (physiological alertness or attentiveness [3]) and valence (degree of positivity or negativity [3]) in vocalizations across mammalian species. Across species, there is thus consistency in the acoustic features that characterise arousal and valence.

In inferring affective information from vocalizations, perceivers might be able to make use of consistencies in affective vocalizations. When listening to conspecific or heterospecific vocalizations, accurate perception of the producer's affective state is beneficial for the perceiver in many contexts [4]. Indeed, inferring affective states from conspecific vocalizations can be essential for the perceiver in contexts including parental behaviour and sexual partner selection. Going beyond conspecific vocalizations, listening to heterospecific vocalizations can be used to gather information about the producer's inner state which can facilitate adaptive behaviour in various contexts, like being attacked by a predator. This ability might be based on inherent capacities to perceive phylogenetically conserved acoustic regularities [4].

Human perception of affective information from heterospecific vocalizations

Most of what we know about human listeners' perception of affective information in heterospecific vocalizations comes from studies on core affect. This work has showed that

humans can accurately infer arousal and valence from vocalizations of many different species [5-13]. However, acoustic features of mammalian vocalizations vary systematically across different types of behavioural contexts such as threats, food, and play, that do not only vary in terms of arousal and valence [14,15]. Perceptual mechanisms may exist that allow human listeners to infer richer affective information from particular types of behaviours than inferences of core valence [16]. For instance, humans associate cats' purring with contentment and dogs' yelping with distress. However, it is not straightforward to map the affective information in heterospecific expressions onto human emotion categories, and there is a clear risk of anthropomorphising those species. An alternative approach is to examine mappings between vocalizations and behavioural contexts as an indirect route to inferring affective states.

Only a few studies to date have tested human listeners' perception of behavioural contexts from heterospecific vocalizations. The results have shown that listeners can correctly classify the production context of dogs' barks [17], cats' meows [18], and the vocalizations of pigs [19]. However, previous studies are limited to domesticated animals that are distantly related to humans. Here, we seek to examine human listeners' ability to infer behavioural context and core affect dimensions from vocalizations of chimpanzees (*Pan troglodytes*), one of the genetically closest living relatives to humans.

In studies comparing animal vocalizations produced in positive and negative contexts, humans have consistently been found to be better at identifying affective information produced in negative contexts [8,17-19]. For instance, listeners correctly identified arousal levels in silver fox vocalizations only when they were produced in negative contexts [8]. In the current study, we therefore examine whether human listeners' perception of behavioural context and core affect dimensions is more accurate for negative as compared to positive chimpanzee vocalizations.

The Present Study

Drawing on two complementary approaches to phylogenetic continuity in emotional expressions, we sought to test the hypotheses that 1) human listeners can accurately infer the type of behavioural context in which chimpanzee vocalizations were produced [16]; and that 2) human listeners can correctly judge arousal and valence from chimpanzee vocalizations [12]. We included chimpanzee vocalizations produced in a wide range of different positive and negative behavioural contexts at high, medium, or low arousal levels.

In Experiment 1, participants were asked to complete a forced-choice context categorisation task for each vocalization, and to rate arousal and valence. We predicted that listeners would be able to categorise the behavioural contexts and to judge the arousal level and valence from the vocalizations at better-than-chance levels. However, the 10-way forced-choice context categorisation task was challenging for participants, and so Experiment 2 employed a simpler paradigm. It tested whether participants could match the vocalizations to a corresponding behavioural context when selecting from two options (match vs. no match). We predicted that listeners would be able to match vocalizations to their respective production contexts at better-than-chance levels. Finally, in both experiments we expected that accuracy would be better for vocalizations produced in negative, as compared to positive, contexts.

In order to investigate the features shaping human listeners' perception of affective information from chimpanzee vocalizations, we conducted an acoustic analysis. First, we examined whether behavioural context, arousal level, and valence would be reflected in the acoustic structure of the vocalizations. Second, we tested which acoustic features would predict humans' perceptual judgements. The hypotheses, methods (including exclusion criteria), and data analysis plan for both experiments were preregistered on the Open Science Framework (osf.io/mkde8) before data collection was commenced.

Experiment 1: Categorisation of Behavioural Contexts and Judgements of Arousal Level and Valence

In Experiment 1, we tested whether human listeners would be able to accurately 1) categorise the behavioural context in which the chimpanzee vocalizations were produced by selecting from 10 context categories; and 2) judge the arousal level (high, medium, low) and valence (positive, negative) of these vocalizations.

Participants

The sample size was predetermined by a power analysis using G*Power 3.1 [20] for a t-test given $d = 0.2$, power = 0.80, $\alpha = 0.005$. The power analysis was conducted based on the context categorisation task, since we expected it to be the most difficult for participants. This categorisation task included separate tests for 10 behavioural context categories; thus Bonferroni-corrected alpha level was used ($\alpha = 0.005$ [0.05/10]), and so 296 participants were required to detect a small effect size. To ensure that the study was not underpowered, data was collected from 14 additional participants to allow for potential exclusions (see Statistical Analyses for exclusion criteria). Consequently, 310 participants (195 female, $M_{age} = 22.08$, $Sd_{age} = 3.39$, range = 18-38 years old) took part in the experiment. All reported having no hearing impairments and no experience working with or studying chimpanzees. Participants were recruited via the University of Amsterdam, Department of Psychology's research pool, and flyers distributed across the university campus. The average duration of the main experiment was 27.43 minutes ($Sd = 9.75$), and participation was compensated with monetary reward or course credit.

Materials and Procedure

Stimuli. In the practice trials, two chimpanzee vocalizations taken from findsounds.com were used as stimuli. In the main task, the stimuli were 155 vocalizations produced by 66 individual chimpanzees in 10 types of behavioural contexts including

positive and negative contexts at high, medium, and low arousal levels (see Table 1). The

behavioural contexts were recorded by author K.E.S. in real time, alongside the sound

recordings of vocalizations, and K.E.S., an expert in chimpanzee vocal communication,

provided classifications of the arousal levels and valence of each call type (see Table 1).

Descriptions of behavioural contexts and classification of each context based on arousal level

and valence, together with number of stimuli, are listed in Table 1; details of the recording

setups are provided in Supplementary Materials Text 1S. A representative vocalization for

each context can be found in Supplementary Materials Audio 1S. All recordings were

normalized for peak amplitude using AudaCity software (http://audacity.sourceforge.net)

before the experiment.

Table 1. *Behavioural contexts and core affect dimensions of chimpanzee vocalizations*

	Positive (n = 80)	Negative (n = 75)	No specific valence (n = 11)
High arousal (n = 62)	Pant hoots when discovering a large food source (n = 12)	Waa barks while threatening an aggressive chimp or predator (n = 16)	
		Victim screams when attacked by another chimpanzee (n = 21)	
		Alarm calls when discovering something scary (n = 13)	
Medium arousal (n = 71)	Rough grunts while eating high value food (n = 19)	Tantrum screams when refused access to food (n = 15)	*Copulation calls while having sex (n = 11)
	Laughter while being tickled (n = 16)	Whimpers by juveniles when separated from mother (n = 10)	
Low arousal (n = 22)	Rough grunts while eating low value food (n = 22)		

Note. For each context, vocalizations were obtained from between 4 and 21 individual chimpanzees.
*copulation calls may be associated with either positive (pleasure) or negative (fear/pain) valence,
thus no specific valence is attributed to the vocalizations produced during copulation.

Experimental procedure. On arrival at the laboratory, each participant was led to a

silent individual cubicle. After completing two practice trials, participants listened to the 155

chimpanzee vocalizations and for each were asked to 1) make a forced-choice context categorisation, selecting from 10 categories, 2) indicate the level of arousal on a 5-point scale (1 = very low, 5 = very high, and 3) indicate the emotional valence on a 5-point scale (1 = very negative, 5 = very positive). Finally, participants reported their familiarity with each behavioural context (How familiar are you with the chimpanzees in the context of X from zoo settings or media?), and a representative vocalization from each context (How familiar are you with this chimpanzee vocalization from zoo settings or media?) on a 5-point scale (1 = not at all, 5 = extremely).

The presentation order of vocalizations, scales, and context categories were randomized separately for each participant. Participants could replay each vocalization as many times as needed to make their judgments. The stimuli were presented through headphones (Monacor MD-5000DR) connected to a computer, and the sound level was held constant across participants. The experimental interface was created using PsychoPy [21].

Statistical Analyses

Before analysis, the data set was checked for outliers, defined as performance of three Sd or more below the mean on the categorisation task. No participants had to be excluded.

To test whether human listeners would perform better than chance in the categorisation of contexts, the proportion of correct responses was calculated for each participant for each context category. Unbiased hit rates (Hu scores, [22]) were calculated to control for individual biases in the use of particular context categories. These were arcsine transformed before the analysis to stabilize variance [22]. Following this transformation, all variables were checked for normality using Shapiro-Wilk test, which indicated that they were not normally distributed ($ps < 0.001$). We therefore employed paired sample Wilcoxon Signed-Rank tests. Chance levels were calculated for each context per individual following Wagner's formula [22]: the product of the column and row marginals, divided by the squared

number of observations. The corrected chance level takes the number of stimuli for each context category into account. Arcsine transformed Hu scores were then compared to chance using a paired sample Wilcoxon-Signed Rank test for each category, Bonferroni corrected for multiple comparisons (0.05/10).

To assess how accurately human listeners judged the arousal level and valence of the vocalizations, ratings on the 5-point scales were transformed into -2 (very low), -1 (low), 0 (medium), 1 (high), 2 (very high); and -2 (very negative), -1 (negative), 0 (neutral), 1 (positive), 2 (very positive). A response was considered correct if 1) arousal ratings were significantly higher (lower) than zero for high (low) arousal vocalizations, 2) arousal ratings were not significantly different from zero for medium arousal vocalizations, 3) valence rating were significantly higher (lower) than zero for emotionally positive (negative) vocalizations. Based on these criteria we calculated arcsine-transformed Hu scores for statistical tests. Using separate Wilcoxon-Signed Rank tests, we tested whether performance was better than chance level for high, medium, and low arousal and for positive and negative vocalizations. Finally, we tested separately for context categorisation, arousal and valence ratings whether perception accuracy was higher for vocalizations produced in negative as compared to positive contexts using paired sample Wilcoxon Signed-Rank tests with the arcsine-transformed Hu scores.

Results

Confusion matrices for average recognition percentages are shown in Figure 1. The results show that participants were not able to accurately categorise any of the behavioural contexts ($ps > 0.005$, Bonferroni corrected; Figure 1a). However, judgments of core affect were significantly better than chance for high ($z = 14.734$, $p < 0.001$), low ($z = 13.567$, $p < 0.001$), and medium ($z = 8.745$, $p < 0.001$) arousal levels (Figure 1b), as well as positive ($z = 14.805$, $p < 0.001$) and negative ($z = 14.713$, $p < 0.001$) valence (Figure 1c).

In the analysis comparing listeners' performance for vocalizations produced in negative vs. positive contexts, the judgements of behavioural contexts were not employed, since participants were unable to identify any of the contexts at better-than-chance levels. The results showed that, consistent with our prediction, participants were more accurate at identifying high arousal from vocalizations produced in negative contexts than in positive contexts ($z = 14.374$, $p < 0.001$). However, listeners were more accurate at inferring medium arousal levels from positive as compared to negative vocalizations ($z = 14.852$, $p < 0.001$). Since the low arousal context category consisted of only one context, vocalizations from negative and positive contexts could not be compared. In terms of valence, listeners were better at judging the valence of vocalizations produced in negative as compared to positive contexts ($z = 15.146$, $p < 0.001$). To assess whether participants tend to perceive vocalizations as more negative or positive in general, we calculated the average of valence ratings across positive and negative vocalizations per participant ($M = -0.29$, $Sd = 1.11$). When the valence ratings were compared against zero, they show a bias towards judging the vocalizations as negative ($z = -14.787$, $p < 0.001$). Human listeners tend to perceive chimpanzee vocalizations as more negative in general.

On average, participants rated both behavioural contexts ($M = 1.86$, $Sd = 0.89$) and representative vocalizations ($M = 2.14$ $Sd = 0.98$) as unfamiliar. Because of the large number of stimuli and judgements, we checked for evidence of fatigue by comparing the accuracy in early (the first 30) and late (the last 30) trials. Pairwise comparisons showed that participants' performance on the arousal judgement task was high in both the early ($M = 46.74$, $Sd = 0.10$) and late trials ($M = 44.92$, $Sd = 0.11$), although participants performed better in the early trials ($z = 2.552$, $p = 0.011$). No difference in accuracy was found for early and late judgments of context categorisation and valence (see Supplementary Materials Table 1S for

details). It is therefore unlikely that participants' performance was adversely affected by possible fatigue.

A.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
c1	**8.4**	9	9.2	9.4	8.4	10.5	9.5	10.3	9.7	10.4
c2	9.6	**11**	10.2	11	9.6	10.6	11.6	9.9	9	9.6
c3	9.9	9.9	**9.5**	8.9	9.9	10.3	9.3	10.1	10.9	10.7
c4	10.5	9.1	11.3	**11.4**	10.5	10.1	10.3	9.8	10.4	11.4
c5	9.2	10.6	11.8	11.1	**9.2**	9.4	10	9.1	9.9	10
c6	10	11.6	10.3	10.4	10	**9**	9.8	9.4	10.5	10.2
c7	11.5	9.9	8.7	9.6	11.5	10.2	**10.6**	10.4	11.6	8.7
c8	10.3	10.4	10.2	9	10.3	10	10	**10.5**	9.4	9.5
c9	9.8	8.7	10.3	9.3	9.8	10.9	9	10.6	**9.2**	9.1
c10	10.7	9.8	8.4	9.9	10.7	9.1	10	9.8	9.5	**10.5**

Percentage: 60, 40, 20, 0

B.

	High	Medium	Low
High	**65.8**	50.2	25.7
Medium	22.5	**31.3**	38.4
Low	11.7	18.5	**35.9**

C.

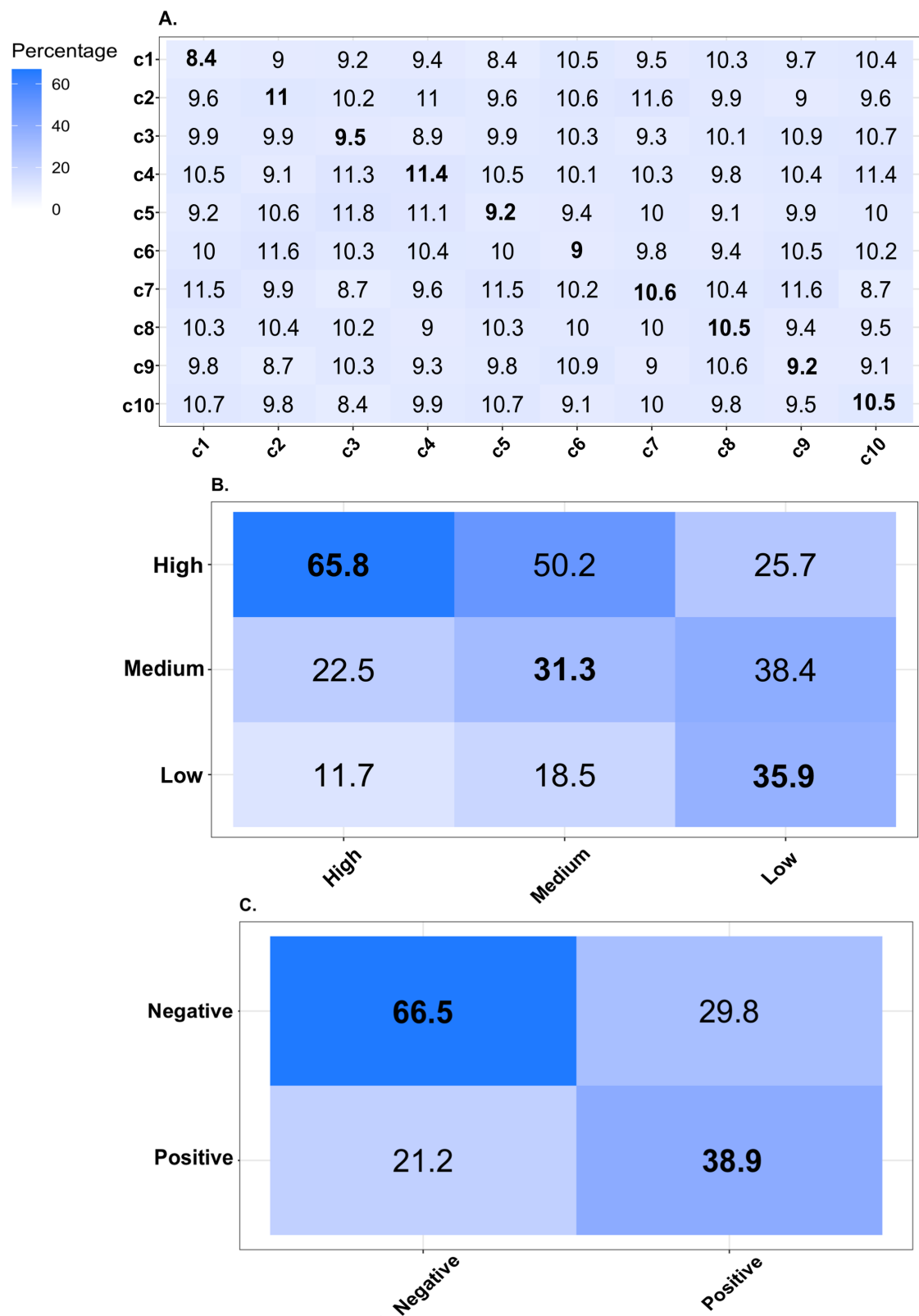|	Negative	Positive
Negative	**66.5**	29.8
Positive	21.2	**38.9**

Figure 1. Heatmap of confusion matrices (%) for behavioural context categorisation data (A) and

arousal (B) and valence (C) judgments. The x-axes represent stimulus types and the y-axes indicate responses. c1 = Eating high value food, c2 = Eating low value food, c3 = Copulating (having sex), c4 = Being separated from mother, c5 = Discovering a large food source, c6 = Being refused access to food, c7 = Being tickled, c8 = Being attacked by another chimp, c9 = Threatening an aggressive chimp, c10 = Discovering something scary.

Experiment 2: Matching Chimpanzee Vocalizations to a Single Behavioural Context

The ten-way context categorisation task used in Experiment 1 was challenging for participants because they were asked to choose from a substantial number of unfamiliar behavioural context categories. In Experiment 2, we therefore sought to test whether listeners would be able to match vocalizations to behavioural contexts in a simpler task involving a single behavioural context for each participant.

Participants

Each participant was given a context matching task for a single context with Yes/No response options on each trial. A power analysis (G*Power 3.1; [20]) based on a t-test given $d = 0.2$, power $= 0.80$, $\alpha = 0.05$ showed that 156 judgments per stimulus were needed. To reduce the risk of learning effects, each participant heard half of the stimuli from the target behavioural context category. This yielded a total number of 312 participants per context category. Since we tested 10 behavioural contexts, the total sample size was set to 3120. Consequently, a total of 3120 participants (1570 females[1], $M_{age} = 34.03$[2], $Sd_{age} = 10.34$, range = 18-75 years old) were recruited from Amazon Mechanical Turk to take part in the experiment. All reported having no hearing impairments or experience of working with or studying chimpanzees. Each session lasted around 10 minutes and participation in the experiment was compensated with monetary reward.

---

[1] Eight participants preferred not to indicate their gender.
[2] Three participants are excluded in the descriptive statistics on age because of birth date errors.

Materials and Procedure

Stimuli. Experiment 2 used the same stimuli as those in Experiment 1 for both the practice trials and the main task.

Experimental procedure. The study was run online using the Qualtrics survey tool (Qualtrics, Provo, UT). Before commencing, participants were instructed to complete the experiment in a silent environment and use headphones. Participants were given two screening questions. On one they were played a doorbell and on the other a car horn sound. They were asked to indicate what they heard, with 'doorbell' and 'car horn' as response options. Participants who failed one or more screening questions were not able to continue to the main experiment.

After the practice and screening trials, each participant was randomly assigned to one of the ten conditions, each focusing on a specific behavioural context. In each condition, participants were asked to give a match-to-context judgment ("Does this vocalization match context X?"), selecting from Yes and No options. The matching vocalizations were a randomly selected subset (half) of the vocalizations from that behavioural context. This constituted 1/4 of the stimuli heard by that participant; the other 3/4 were the non-matching stimuli randomly drawn from all of the other context categories. Only a quarter of the stimuli heard by a given participant were thus from the relevant behavioural context, again to reduce the risk of learning effects. The presentation order of vocalizations was randomized for each subject.

Statistical Analyses

The data set was checked for participants whose performance was three Sd or more below the context-specific mean, but none were and so all data were retained.

We quantified participants' ability to match behavioural contexts using the sensitivity index d-prime. D-prime controls for individual biases in the use of a particular response, and

is calculated as z-transformed hit rates minus false alarm rates [23]. Hit and false alarm rates with extreme values (i.e., 0 or 1) return an error when z-transformed. Those cases are commonly adjusted by replacing rates of zero with 0.5/n (0.5/m) and rates of 1 with (n-0.5)/n ([m-0.5]/m) where n (m) is the number of signal (noise) trials [24]. We calculated hit rates as the proportion of Yes trials to which participants responded Yes, false alarm rates as the proportion of No trials responded to as Yes. In order to test our hypothesis that human listeners would perform better than chance in matching vocalizations to context types, d-prime scores for each participant were tested against chance (random guessing, reflected by a d-prime score of zero) using separate one sample t-tests for each context type at the Bonferroni corrected level α level (α = .005).

Furthermore, we tested whether performance would be better for negative than for positive vocalizations. This was tested using an ANOVA comparing the mean accuracy from negative vs. positive behavioural contexts using d-prime scores.

Results

Mean accuracy levels (d primes) per behavioural context are shown in Figure 2. The statistical tests showed that participants were able to accurately match most of the vocalizations to behavioural contexts. Specifically, performance was significantly better than chance for eating high value food (t = 5.04, p < 0.001, d = 0.28, 95% CIs [0.15, 0.34]), eating low value food (t = 9.59, p < 0.001, d = 0.55, 95% CIs [0.49, 0.75]), discovering a large food source (t = 5.52, p < 0.001, d = 0.31, 95% CIs [0.17, 0.37]), being refused access to food (t = 13.09, p < 0.001, d = 0.74, 95% CIs [0.63, 0.85]), being attacked by another chimpanzee (t = 22.99, p < 0.001, d = 1.29, 95% CIs [1.06, 1.26]), and threatening an aggressive chimp or predator (t = 11.19, p < 0.001, d = 0.64, 95% CIs [0.37, 0.53]). Performance was not better than chance, however, for vocalizations of copulation (having sex) (t = -2.99, p = 0.003, d = 0.17, 95% CIs [-0.04, 0.003]), being separated from mother (t = 2.81, p = 0.005, d = 0.16,

95% CIs [0.04, 0.23]), being tickled ($t = 1.77$, $p = 0.19$, $d = 0.10$, 95% CIs [-0.01, 0.19]), and

discovering something scary ($t = -16.49$, $p = 0.003$, $d = 0.93$, 95% CIs [-0.68, -0.53]).

Accuracy levels for matching vocalizations produced in negative contexts were significantly

better than those from positive contexts (negative: $M = 0.43$, $Sd = 0.37$, positive: $M = 0.22$,

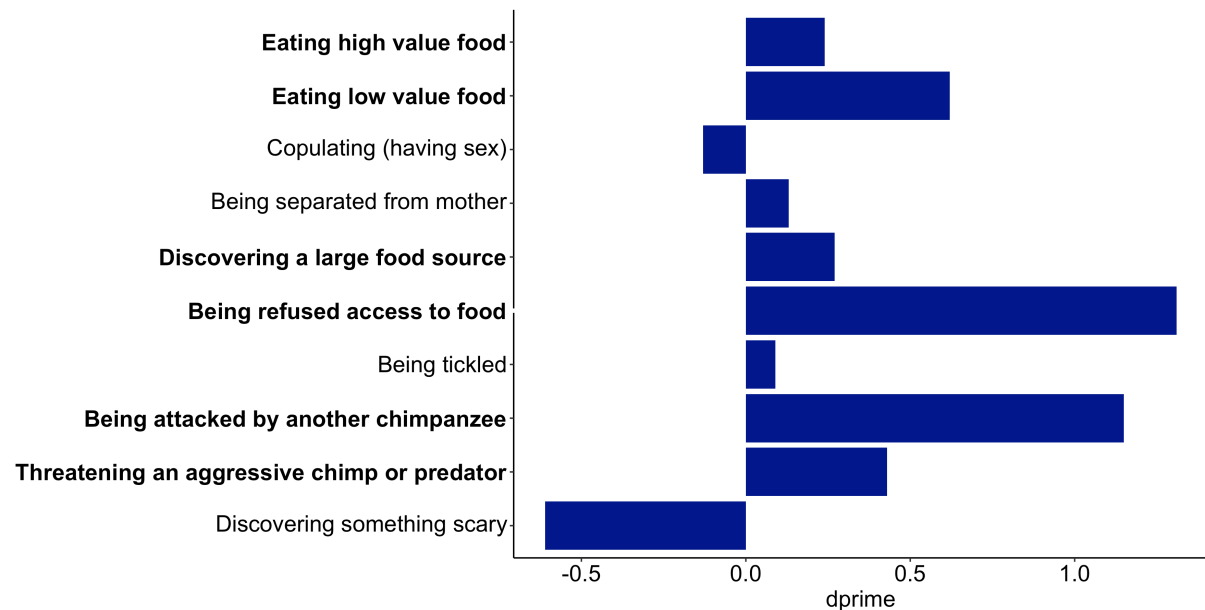$Sd = 0.47$, $F(1,631) = 38.938$, $p < .001$).

Figure 2. d-prime scores per behavioural context showing human listeners' performance in matching

vocalizations to production contexts. Bold indicates better than the chance level performance.

Acoustic Analysis

We performed an acoustic analysis to explore acoustic features shaping human

perception of affective information in chimpanzee vocalizations. First, independently of

perceptual responses of listeners, a classification analysis was conducted to test whether

chimpanzee vocalizations differ by context, arousal level, and valence, in terms of acoustic

features. We then examined which acoustic features, if any, would predict humans' ability to

accurately infer affective information from chimpanzee vocalizations in terms of correctly

judging the arousal level and valence of the vocalizations (Experiment 1) and accurately

matching the vocalizations to the corresponding behavioural context (Experiment 2).

Method

Extraction of acoustic features from chimpanzee vocalizations. We measured

acoustic features of 155 vocalizations produced by 66 individual chimpanzees using PRAAT

[27]. For each vocalization, we measured the following acoustic features: *number of calls in*

a bout, *duration of each call*, *time of the maximum peak frequency*, *relative position of the*

peak frequency within a call, *percentage of voiced frames*, *jitter*, *shimmer*, *spectral centre of*

gravity (*SCoG*) as well as *minimum*, *maximum*, *mean,* and *Sd of fundamental frequency* (f_0)

and *harmonics-to-noise ratio* (*HNR*). We based the choice of parameters on previous findings

on production and perception of affective mammalian vocalizations: *duration*, f_0 and *HNR* are

linked to the affective state of the caller across many animal species [4,14,26,27]. *Peak*

frequency has been found to differ across dog barks recorded in different contexts [17]. In

addition, the *percentage of voiced frames* was added as a tonality measure because in

nonverbal human vocalizations such as laughter, voiced frames are typically more periodic,

while unvoiced frames are noisier and more aperiodic [28]. *Jitter* and *shimmer* are important

parameters for analysis of arousal in animal vocalizations [29], while *spectral centre of*

*grav*ity is associated with the perception of arousal in humans [6,7,30]. Means and Sds of all

acoustic parameters can be found in Supplementary Table 2S.

Further selection of acoustic features. To avoid multicollinearity, we performed a

Principal Component Analysis (PCA) with varimax rotation on the 15 acoustic parameters to

attempt to reduce the number of acoustic parameters. Based on the examination of the scree

plot and selecting components that explain more than 10% of the variance, the first three

components, together explaining 63% of the variance, were retained. Factor loadings on the

three acoustic dimensions can be found in Supplementary Table 3S. Online interactive map

showing the distribution of the 10 behavioural contexts is available in

https://emotionwaves.github.io/context/, arousal levels in

https://emotionwaves.github.io/arousal/, and valence in

https://emotionwaves.github.io/valence/ on the first three acoustic dimensions. These

visualisations demonstrate that the behavioural contexts, arousal levels, and valence are

reflected in the acoustic structure of vocalizations. The first dimension mainly relates to

HNR, which is a measure of clear vs. noisy components in the signal. The second dimension

is primarily related to *pitch*, while the third mainly relates to temporal measures. The

variance of inflation factor (VIF) was substantially greater than 1 for acoustic features

loading on the second and third dimensions (*duration*: 4.62; f_0 *min*: 30.42; f_0 *max*: 64.16; f_0

mean: 37.86; f_0 *Sd*: 21.05; *time of the maximum peak frequency*: 4.10) indicating that there

was a collinearity problem [32]. We therefore selected the features with highest

interpretability based on the previous literature in addition to the factor loadings on the first

three components. This selection allowed us to choose acoustic features with low VIF and

high factor loadings on the first three dimensions. The selected acoustic features for statistical

analyses were thus: *SCoG, duration*, f_0 *mean*, f_0 *Sd, HNR mean*, and *HNR max*. Collinearity

was not a problem for these features (VIF: *SCoG*: 1.67; *duration*: 1.47; f_0 *mean*: 1.62; f_0 *Sd:*

1.25; *HNR mean:* 1.56; *HNR max:* 1.62).

Statistical Analyses

We sought to test whether behavioural contexts, arousal levels and valence could be

differentiated based on the selected acoustic features. Multinomial Logistic Regressions

(MLR) were performed in SPSS (Version 23, IBM Statistics) on the acoustic features to

determine whether the acoustic parameters provide sufficient information to predict the actual

behavioural contexts and arousal levels, and Binomial Logistic Regression (BLR) for

valence.

To assess which, if any, acoustic parameters of the vocalizations would map onto

listeners' ability to accurately perceive 1) behavioural context, 2) arousal levels, and 3)

valence, we conducted three generalized linear mixed models (GLMMs). The dependent variable was a binary response (i.e., correct or incorrect response). Participant and chimpanzee IDs were entered as random factors, accounting for participant and chimpanzee variability. The selected acoustic parameters were set as fixed factors. We used Akaike's information criterion (AIC) to select the most parsimonious model [34]. ΔAICs are calculated as the difference between the AICc of the fitting model and the best model to identify the models with the highest power to explain the variation in the dependent variable. Lower AIC values indicate improved support for each model [32,33], and each added variable is considered to improve the fit only if it increases the AIC value by more than two units [34]. GLMMs were implemented using lme4 package [35] with optimizer 'bobyqa' [36]. Binomial data and estimated odds were plotted as forest plots for fixed effects 'sjplot' package in R [37].

Results

      Classification of behavioural contexts, arousal levels, and valence based on acoustic parameters. MLR on behavioural contexts showed that the overall model was significant $\chi2$ (54) = 595.618, p < 0.001. All acoustic parameters, *SCoG* ($\chi2$ (9) = 92.919, p < 0.001), *duration* ($\chi2$ (9) = 114.154, p < 0.001), *f_0 mean* ($\chi2$ (9) = 92.283, p < 0.001), *f_0 Sd* ($\chi2$ (9) = 50.906, p < 0.001), *HNR mean* ($\chi2$ (9) = 112.324), p < 0.001, and *HNR max* ($\chi2$ (9) = 22.620, p < 0.01) made significant unique contributions and the overall model showed 85.7% classification agreement on behavioural context classification.

      The results from the MLR on arousal levels revealed that the overall model was significant ($\chi2$ (2) = 191.391, p < 0.001). Significant contributions were made by *SCoG* ($\chi2$ (2) = 28.990, p < 0.001), *duration* ($\chi2$ (2) = 72.489, p < 0.001), and *HNR mean* ($\chi2$ (2) = 25.352, p < 0.001). Vocalizations with higher arousal levels were longer in *duration* compared to vocalizations with lower arousal levels. *HNR mean* was higher for high and

medium arousal and lower for low arousal vocalizations, while the *SCoG* of low arousal vocalizations was lower than that of medium and high arousal vocalizations. The final model showed a classification agreement of 83.1%.

Third, the binomial logistic regression on valence showed that the overall model was significant ($\chi2$ (6) = 60,433, $p < 0.001$). *Duration* ($\chi2$ (6) = 8.789, $p < 0.01$), f_0 *mean* ($\chi2$ (6) = 19.797, $p < 0.01$), and *HNR mean* ($\chi2$ (6) = 5.381, $p < 0.05$) made significant unique contributions. *Duration* was longer for negative vocalizations and f_0 *mean* and *HNR mean* were higher for negative vocalizations than positive vocalizations. The final model had a classification agreement of 73.4%.

Prediction of human listeners' perceptual judgments from acoustic parameters. GLMMs revealed that *SCoG* ($z = 6.59$, $p < 0.001$), *duration* ($z = 2.83$, $p < 0.01$), f_0 *Sd* ($z = -2.73$, $p < 0.01$), *HNR mean* ($z = -6.03$, $p < 0.001$), and *HNR max* ($z = 3.31$, $p < 0.001$) significantly predicted accurate match-to-context responses in Experiment 2. *SCOG* is a measure of how high the frequencies in a spectrum are, which is perceptually connected with the impression of *brightness* of a vocalization. *Duration* refers to the total duration of calls in whole stimulus, while f_0 is the lowest periodic cycle of the acoustic signal, which has the perceptual correlate of *pitch*. *HNR* is the degree of acoustic periodicity, which relates to human perception of *noisiness*. The model selection procedure based on the AIC identified the model excluding f_0 *mean* as the strongest model for explaining variation in human listeners' accurate responses in the match-to-context task. The best predictor of performance was *SCoG*, which was linked to participants' ability to correctly match vocalizations to behavioural contexts (see Figure 3).

GLMM predicting accurate arousal level judgments in Experiment 1 revealed significant effects of *SCoG* ($z = 5.33$, $p < 0.001$), *duration* ($z = 2.91$, $p < 0.05$), f_0 *mean* ($z = 13.25$, $p < 0.001$), and f_0 *Sd* ($z = 13.90$, $p < 0.001$). Increases in those acoustic parameters

predicted higher accuracy in listeners' judgments of arousal level. The best predictor of arousal level judgments was f_0 Sd. Specifically, decreases in this parameter (corresponding approximately to less pitch variability) predicted better listener accuracy in identification of arousal levels. For valence judgments, $SCoG$ (z = 11.96, p < 0.001), $duration$ (z = 8.24, p < 0.001), f_0 $mean$ (z = 15.48, p < 0.001), and f_0 Sd (z = -5.78, p < 0.001), showed significant effects on the prediction of listeners' performance. Specifically, increases in $SCoG$, $duration$, and f_0 $mean$ predicted more accurate valence judgments, while increases in f_0 Sd predicted lower accuracy. The best predictor of valence judgements was f_0 $mean$, which predicted better accuracy in the identification of valence. In explaining variation in human listeners' accuracy in identifying both arousal levels and valence from chimpanzee vocalizations, the model excluding HNR $mean$ as well as the model without HNR max were the strongest models. The effects of each acoustic features on the accurate perception of behavioural context, arousal levels and valence are visualised in Figure 3. Full details of the GLMMs and model selection procedures are provided in Supplementary Tables 4S and 5S.
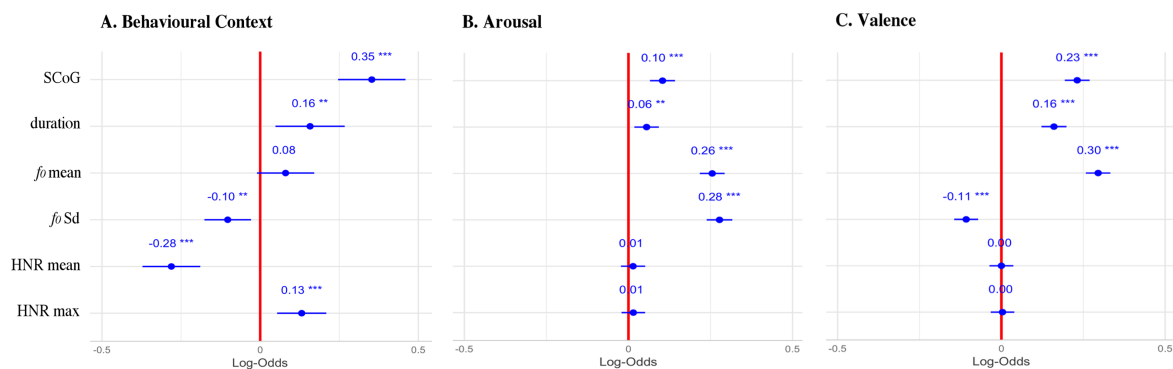

Figure 3. Forest plots of estimates of the generalized linear mixed models. Estimates for fixed effects are given as log-odds. The vertical intercept indicates no effect. A) Behavioural context based on match-to-context task in Experiment 2, B) Arousal level judgment task in Experiment 1, C) Valence judgement task in Experiment 1.

Discussion

Two experiments tested human listeners' ability to accurately 1) perceive behavioural contexts in which chimpanzee vocalizations were produced, using a 10-way context categorisation task and a Yes/No match-to-context task; and judge 2) arousal and valence from chimpanzee vocalizations. Human listeners failed to categorise production contexts of vocalizations when a 10-way forced-choice task was used. However, they were able to match vocalizations to most behavioural contexts in the simpler Yes/No match-to-context task. In addition, the arousal levels (high, medium, low) and valence (positive, negative) of the chimpanzee vocalizations were accurately inferred by human listeners. Overall, participants performed better with negative, as compared to positive, vocalizations.

In Experiment 1, participants were asked to select the best matching context from 10 unfamiliar behavioural context categories. Such tasks are difficult for listeners as it is more challenging to evaluate and compare contexts [4]. Moreover, 10 is a large number of options for a categorisation task. It has been suggested that even though increasing the number of alternatives in forced-choice tasks has advantages (e.g., reducing the guessing rate), for a given task, there is a point at which the number of options becomes too large for participants [38]. The present results suggest that for human listeners to be able to accurately map chimpanzee vocalizations to 10 unfamiliar behavioural contexts, participants may require additional information about the contexts, and/or information carried by other channels such as facial expressions.

In Experiment 2, when a Yes/No match-to-context task was used, listeners accurately matched the vocalizations produced while eating high and low value food, discovering a large food source, being refused access to food, being attacked by another chimpanzee, and threatening an aggressive chimp or predator. Given that listeners in our experiment had minimal prior exposure to chimpanzees, they are unlikely to have learned to decode

chimpanzee vocalizations. Rather, accurately mapping heterospecific vocalizations to behavioural contexts linked to affective states may draw on acoustic regularities that are conserved across related species. For instance, African elephants can differentiate between threatening and non-threatening human vocalizations [39], and Japanese sika deer uses the vocalizations produced by Japanese macaques when they discover a food source to locate fruit [40]. In these contexts, understanding heterospecific vocalizations clearly benefits the perceiver, and thus may confer a fitness advantage. To assess the effect of different degrees of acoustic regularities in vocalizations on perception of behavioural contexts from heterospecific vocalizations, future studies should aim at including vocalizations from multiple species differing in phylogenetic closeness.

Listeners failed to match vocalizations of copulation, being separated from mother, being tickled, and discovering something scary. A possible explanation is that there may be a great deal of variability in the vocalizations produced in these contexts, depending on factors such as who potential perceivers are (e.g., kin vs. non-kin, allies vs. competitors). For instance, female chimpanzee copulation calls have been found to differ when copulating with high ranking males compared to low ranking males [41]. Thus, listeners might need additional contextual information to be able to specify vocalizations produced in certain type of contexts, or might not be able to identify certain contexts from vocalizations at all.

In general, listeners' judgments of negative behavioural contexts were more accurate than judgments of positive contexts. Similarly, high arousal vocalizations and valence were more accurately inferred from vocalizations produced in negative contexts. In particular, accuracy was especially high for highly aroused negative vocalizations, which might signal immediate, potentially dangerous situations. It has been proposed that stronger phylogenetic continuity for negative affective signals may be a result of a homologous signalling system that benefits species in dangerous contexts [8]. From this perspective, the acoustic structure

of vocalizations produced in negative contexts may be more likely to have been conserved, since negative contexts involve risks. Survival might be facilitated by the ability to recognise vocalizations produced in negative contexts not only by conspecifics, but also by members of other species [42]. Indeed, cross-species "eavesdropping" on alarm calls has been suggested to increase chances of survival [43]. Thus, acoustic structure may have been preserved to a greater degree for negative as compared to positive vocalizations.

Independently of listeners' perceptual responses, acoustic features of chimpanzee vocalizations varied systematically across different behavioural contexts, arousal levels, and valence. Listeners utilized *brightness*, *duration*, *pitch variation*, *noisiness*, and *maximum level of noisiness* to make accurate classifications of vocalizations into behavioural contexts. *Brightness*, *duration*, *pitch*, and *pitch variability* predicted listeners' ability to correctly infer both arousal levels and valence. *Noisiness* of vocalizations was a more useful acoustic feature in matching production contexts compared to other features, while more simple acoustic features like *pitch mean* and *pitch variation* were more effective in identification of arousal and valence. In line with our findings, Maruščáková and colleagues [11] found that simple acoustic features such as *pitch* were more useful in human judgments of valence than *noisiness* in piglet vocalizations. Similarly, Filippi and colleagues [7] have shown that humans mainly rely on *pitch* to identify high arousal vocalizations across nine vertebrate species. Furthermore, consistently with our findings, *duration* and *brightness* have also been suggested to be effective acoustic features in humans' ability to identify arousal level from vocalizations [5, 7, 30]. In sum, acoustic analysis revealed that chimpanzee vocalizations differ by context, arousal, and valence based on acoustic features and allowed us to identify specific features contributing to human listeners' perceptual judgments.

In conclusion, the present study demonstrates that human listeners can accurately perceive affective information beyond core affect dimensions from the vocalizations of a

closely related species, chimpanzees. These findings suggest phylogenetic preservation of acoustic features mapping onto specific behavioural contexts, as well as features characterising arousal levels and valence.

Acknowledgements

Ethics

The School of Psychology Ethics committee, University of St Andrews gave ethical clearance for the non-invasive, behavioural studies that included the recording of chimpanzee vocalizations. Experiment 1 (project number 2018-SP-9517) and Experiment 2 (project number 2019-SP-10653) were approved by the Ethics Review Board of the Faculty of Social and Behavioural Sciences, University of Amsterdam. All participants provided informed consent before participation.

References

1.Darwin C. 1872 *The expression of the emotions in man and animals.* London, UK: John Murray. (doi:10.1037/10001-000)

2. Briefer EF. 2012 Vocal expression of emotions in mammals: mechanisms of production and evidence. *J. Zool.* **288,** 1–20. (doi:10.1111/j.1469-7998.2012.00920.x)

3. Russell JA. 1980 A circumplex model of affect. *J. Pers. Soc. Psychol.* **39,** 1161–1178. (doi:10.1037/h0077714)

4. Andics A, Faragó T. 2018 Voice perception across species. In *Oxford handbook of voice perception* (eds S. Frühholz & P. Belin), pp. 363–39. Oxford, UK: Oxford University Press. (doi: 10.1093/oxfordhb/9780198743187.013.16)

5. Congdon JV, Hahn AH, Filippi P, Campbell KA, Hoang J, Scully EN, Bowling DL, Reber SA, Sturdy CB. 2019 Hear them roar: a comparison of black-capped chickadee (Poecile atricapillus) and human (Homo sapiens) perception of arousal in vocalizations across all classes of terrestrial vertebrates. *J. Comp. Psychol.* **133,** 520–541. (doi:10.1037/com0000187)

6. Faragó T, Andics A, Devecseri V, Kis A, Gácsi M, Miklósi A. 2014 Humans rely on the same rules to assess emotional valence and intensity in conspecific and dog vocalizations. *Biol. Lett.* **10,** 20130926. (doi:10.1098/rsbl.2013.0926)

7. Filippi P et al. 2017 Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proc. R. Soc. Lond. B. Biol. Sci.* **284,** 20170990. (doi:10.1098/rspb.2017.0990)

8. Filippi P, Gogoleva SS, Volodina EV, Volodin IA, Boer B de. 2017 Humans identify negative (but not positive) arousal in silver fox vocalizations: implications for the adaptive value of interspecific eavesdropping. *Curr. Zool.* **63,** 445–456. (doi:10.1093/cz/zox035)

9. Fritz T, Mueller K, Guha A, Gouws A, Levita L, Andrews TJ, Slocombe KE. 2018 Human behavioural discrimination of human, chimpanzee and macaque affective vocalizations is reflected by the neural response in the superior temporal sulcus. *Neuropsychologia* **111,** 145–150. (doi:10.1016/j.neuropsychologia.2018.01.026)

10. McComb K, Taylor AM, Wilson C, Charlton BD. 2009 The cry embedded within the purr. *Curr. Biol.* **19,** 507–508. (doi:10.1016/j.cub.2009.05.033)

11. Maruščáková IL, Linhart P, Ratcliffe VF, Tallet C, Reby D, Špinka M. 2015 Humans (Homo sapiens) judge the emotional content of piglet (Sus scrofa domestica) calls based on simple acoustic parameters, not personality, empathy, nor attitude toward animals. *J. Comp. Psychol.* **129,** 121–131. (doi:10.1037/a0038870)

12. Scheumann M, Hasting AS, Kotz SA, Zimmermann E. 2014 The voice of emotion across species: how do human listeners recognize animals' affective states? *PLoS ONE* **9,** e91192. (doi:10.1371/journal.pone.0091192)

13. Tallet C, Špinka M, Maruščáková I, Šimeček P. 2010 Human perception of vocalizations of domestic piglets and modulation by experience with domestic pigs (Sus scrofa). *J. Comp. Psychol.* **124,** 81–91. ([doi:10.1037/a0017354](doi:10.1037/a0017354))

14. Morton ES. 1977 On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *Am. Nat.* **111,** 855–869. ([doi:10.1086/283219](doi:10.1086/283219))

15. Rendall D, Owren MJ, Ryan MJ. 2009 What do animal signals mean? *Anim. Behav.* **78,** 233–240. (doi:10.1016/j.anbehav.2009.06.007)

16. Adolphs R, Anderson DJ. 2018 *The neuroscience of emotion: a new synthesis*. Princeton, NJ: Princeton University Press. ([doi: 10.2307/j.ctvc77b1j](doi:10.2307/j.ctvc77b1j))

17. Pongrácz P, Molnár C, Miklósi Á, Csányi V. 2005 Human listeners are able to classify dog (Canis familiaris) barks recorded in different situations. *J. Comp. Psychol.* **119,** 136–144. ([doi:10.1037/0735-7036.119.2.136](doi:10.1037/0735-7036.119.2.136))

18. Nicastro N, Owren MJ. 2003 Classification of domestic cat (Felis catus) vocalizations by naive and experienced human listeners. *J. Comp. Psychol.* **117,** 44–52. ([doi:10.1037/0735-7036.117.1.44](doi:10.1037/0735-7036.117.1.44))

19. Scheumann M, Hasting AS, Zimmermann E, Kotz SA. 2017 Human novelty response to emotional animal vocalizations: effects of phylogeny and familiarity. *Front. Behav. Neurosci.* **11,** 204. (doi:10.3389/fnbeh.2017.00204)

20. Faul F, Erdfelder E, Lang A-G, Buchner A. 2007 G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods.* **39,** 175–191. ([doi:10.3758/bf03193146](doi:10.3758/bf03193146))

21. Peirce JW. 2007 PsychoPy—Psychophysics software in Python. *J. Neurosci. Methods.* **162,** 8–13. (doi:10.1016/j.jneumeth.2006.11.017)

22. Wagner HL. 1993 On measuring performance in category judgment studies of nonverbal behavior. *J. Nonverbal. Behav.* **17,** 3–28. (doi:10.1007/bf00987006)

23. Macmillan, NA. 1993 Signal detection theory as data analysis method and psychological decision model. In *A handbook for data analysis in the behavioural sciences: methodological issues* (eds G. Keren & C. Lewis), pp. 21–57. Hillsdale, NJ: Lawrence Erlbaum.

24. Macmillan NA, Kaplan HL. 1985 Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychol. Bull.* **98,** 185–199. (doi:10.1037/0033-2909.98.1.185)

25. Boersma P, Weenink D. 2017 Praat: Doing phonetics by computer (version 6.0.30). Computer program. See http://www.fon.hum.uva.nl/praat/.

26. Taylor AM, Reby D. 2009 The contribution of source-filter theory to mammal vocal communication research. *J. Zool.* **280,** 221–236. (doi:10.1111/j.1469-7998.2009.00661.x)

27. Zimmermann E, Leliveld L, Schehka S. 2013 Toward the evolutionary roots of affective prosody in human acoustic communication: A comparative approach to mammalian voices. In *Evolution of emotional communication* (eds E. Altenmüller, S. Schmidt & E. Zimmermann), pp. 116–132. Oxford, UK: Oxford University Press. (doi:10.1093/acprof:oso/9780199583560.003.0008)

28. Bachorowski J-A, Owren MJ. 2001 Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychol. Sci.* **12,** 252–257. (doi:10.1111/1467-9280.00346)

29. Li X, Tao J, Johnson MT, Soltis J, Savage A, Leong KM, Newman JD. 2007 Stress and emotion classification using jitter and shimmer features. In *Proc. IEEE Int. Conf. Acoust.*

Speech Signal Process, Vol. IV, pp. 1081-1084, Honolulu, HI, USA, April 2017. IEEE, ISBN 1-4244-0727-3. (doi:10.1109/icassp.2007.367261)

30. Sauter DA, Eisner F, Calder AJ, Scott SK. 2010 Perceptual cues in nonverbal vocal expressions of emotion. *Q. J. Exp. Psychol.* **63,** 2251–2272.

(doi:10.1080/17470211003721642)

31. Bowerman BL, O'connell RT. 1994 *Linear statistical models: An applied approach.* Monterey, CA: Brooks/Cole.

32. Akaike H. 1974 A new look at the statistical model identification. In *Selected papers of Hirotugu Akaike* (eds E. Parzen, K. Tanabe & G. Kitagaw), pp. 215–222. New York, NY: Springer. (doi:10.1007/978-1-4612-1694-0_16)

33. Golabek KA, Ridley AR, Radford AN. 2012 Food availability affects strength of seasonal territorial behaviour in a cooperatively breeding bird. *Anim. Behav.* **83,** 613–619.

(doi:10.1016/j.anbehav.2011.11.034)

34. Burnham KP, Anderson DR. 2004 Multimodel inference. *Sociol. Methods. Res.* **33,** 261–304. (doi:10.1177/0049124104268644)

35. Bates D, Mächler M, Bolker B. 2014 Walker S. Fitting linear mixed-effects models using lme4. *arXiv preprint* See https://arxiv.org/pdf/1406.5823.pdf.

36. Powell MJ. 2009 The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA report,* Cambridge, UK: Cambridge University Press.

37. Lüdecke D. 2018 *sjPlot: Data visualization for statistics in social science.* R package version. See https://cran.r-project.org/web/packages/sjPlot/index.html

38. Vancleef K, Read JCA, Herbert W, Goodship N, Woodhouse M, Serrano-Pedraza I. 2018 Two choices good, four choices better: For measuring stereoacuity in children, a four-alternative forced-choice paradigm is more efficient than two. *PLoS ONE* **13,** e0201366.

(doi:10.1371/journal.pone.0201366)

39. McComb K, Shannon G, Sayialel KN, Moss C. 2014 Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. *Proc. Natl Acad. Sci.* **111,** 5433–5438. (doi:10.1073/pnas.1321543111)

40. Koda H. 2012 Possible use of heterospecific food-associated calls of macaques by sika deer for foraging efficiency. *Behav Process.* **91,** 30–34. (doi:10.1016/j.beproc.2012.05.006)

41. Townsend SW, Deschner T, Zuberbühler K. 2008 Female chimpanzees use copulation calls flexibly to prevent social competition. *PLoS ONE* **3,** e2431. (doi:10.1371/journal.pone.0002431)

42. Nesse RM. 1990 Evolutionary explanations of emotions. *Hum. Nat.* **1,** 261–289. (doi:10.1007/bf02733986)

43. Magrath RD, Haff TM, Fallow PM, Radford AN. 2014 Eavesdropping on heterospecific alarm calls: from mechanisms to consequences. *Biol. Rev. Camb. Philos. Soc.* **90,** 560–586. (doi:10.1111/brv.12122)

Supplementary Materials

Human Listeners' Perception of Behavioural Context and Core Affect Dimensions in

Chimpanzee Vocalizations

Text S1: Recording of chimpanzee vocalizations

The behavioural contexts as well as classifications based on arousal level (high, medium, low) and valence (positive, negative) were determined by author K.E.S., who is an expert on chimpanzee vocal communication. K.E.S. recorded the stimuli from individuals in the Sonso community of wild chimpanzees, Budongo Forest, Uganda, and captive chimpanzees housed at Edinburgh Zoo and the Wolfgang Kohler Primate Research Centre, Leipzig, Germany. Audio recordings were made with a Sennheiser K6/ME67 directional microphone using either (i) a Sony TCD-D8 portable digital audiotape (DAT) recorder (Budongo/Edinburgh), and the recordings were then digitized at a sampling rate of 44.1 kHz, 16 bits accuracy with Cool Edit Pro LE (1999), or (ii) a MARANTZ PMD660 solid state recorder (sampling rate of 44.1 kHz, 16 bits accuracy) (Leipzig). The behaviour of the caller, the response of individuals in the group or party, and the general behavioural context were noted for each vocalization. Only vocalizations from single individuals whose identities were known used in this study.

Audio 1S: Examples of chimpanzee vocalizations produced in 10 behavioural contexts

Rough grunts while eating high value food:

Rough grunts while eating low value food:

Copulation calls while having sex:

Whimpers by juveniles when separated from mother:

Pant hoots when discovering a large food source:

Tantrum screams when refused access to food:

Laughter while being tickled:

Victim screams when attacked by another chimpanzee:

Waa barks while threatening an aggressive chimp or predator:

Alarm calls when discovering something scary:

Examples of chimpanzee vocalizations produced in 10 behavioural contexts can also be listened from https://emotionwaves.github.io/chimp/

Table 1S

Table 1S. *Comparison of accuracy in early and late trials in Experiment 1*

	Percentage accuracy (Mean and SD)		Pairwise comparison (first 30 trials vs last 30 trials)	
	First 30 trials	**Last 30 trials**	***Z***	***p-value***
Behavioural Context	10.30 (0.06)	10.22 (0.07)	0.560	0.576
Arousal level	46.74 (0.10)	44.92 (0.11)	2.552	0.011
Valence	51.84 (0.12)	53.50 (0.14)	-1.898*	0.059

Note. * t value. Wilcoxon signed-rank test is used for pairwise comparisons of accuracy in context categorisation and arousal judgements since data were not normally distributed while pair sampled t-test is used for comparisons of valence judgements since data was normally distributed.

Table 2S

Table 2S. *Acoustic analysis of vocalizations of each behavioural context, arousal level and valence, as Means per stimulus category (separate rows), with Standard Deviations (separated by comma)*

	SCoG	No. Call	Dur.	f_0 min	f_0 max	f_0 mean	f_0 Sd	Time Max	Peak Pos.	%Voi	Jitter	Shimmer	HNR min	HNR max	HNR mean	HNR Sd
Eating high value food	1161.640, 355.521	2.737, 0.452	0.142, 0.050	633.064, 169.215	800.896, 197.138	734.491, 186.646	56.475, 40.446	0.155, 0.054	114.753, 30.428	97.668, 5.599	0.000, 0.000	1.442, 0.179	-225.78, 0.524	28.907, 6.067	11.499, 3.149	5.579, 0.895
Eating low value food	482.756, 400.873	3.545, 1.011	0.94, 0.030	462.106, 153.080	551.556, 192.803	501.649, 164.476	36.217, 31.461	0.143, 0.029	171.512, 53.826	55.094, 27.668	0.002, 0.005	2.064, 0.428	-224.97, 0.363	21.095, 5.874	1.462, 1.697	4.490, 0.908
Copulating (having sex)	951.698, 511.915	2.273, 0.905	0.261, 0.180	711.059, 210.167	806.838, 236.922	757.436, 217.221	27.300, 15.048	0.184, 0.165	73.575, 31.598	100.000, 0.000	0.000, 0.000	1.356, 0.241	-226.02, 0.465	29.355, 5.330	13.803, 3.073	5.730, 1.062
Being separated from mother	1148.748, 740.622	3.700, 1.06	0.242, 0.087	583.082, 169.378	693.122, 256.831	632.062, 191.137	32.205, 28.531	0.185, 0.052	96.676, 42.150	96.928, 6.314	0.000, 0.000	1.465, 0.236	-225.58, 0.384	23.626, 6.196	9.010, 5.097	4.328, 0.958

Discovering a large food source	1653.112, 617.530	3.833, 0.937	0.568, 0.237	282.587, 116.743	771.236, 157.211	561.161, 142.385	154.175, 85.360	0.444, 0.193	90.054, 17.448	95.582, 4.925	0.001, 0.001	1.323 0.248	-226.068, 0.409	33.819, 4.702	12.974, 4.726	6.645, 1.114
Being refused access to food	3000.796, 493.814	2.200, 0.414	0.375, 0.099	877.821, 355.987	1408.327, 242.432	1200.024, 234.628	185.684, 143.729	0.308, 0.115	81.464, 16.294	87.578, 22.024	0.000, 0.000	1.626, 0.248	-225.446, 0.319	25.390, 7.112	5.824, 2.806	3.780, 0.816
Being tickled	1437.125, 1344.207	4.750, 1.915	0.172, 0.113	374.973, 250.662	610.146, 298.147	456.149, 242.061	92.842, 117.532	0.213, 0.096	136.606, 47.763	60.033, 33.584	0.002, 0.002	1.778, 0.150	-225.527, 0.481	25.515, 5.684	3.882, 3.165	3.384, 0.895
Being attacked by another chimpanzee	2226.703, 604.872	1.762, 0.625	0.535, 0.237	979.130, 383.364	14430.974, 172.694	1229.112, 247.389	149.64, 183.116	0.371, 0.156	74.056, 20.606	99.350, 1.317	0.000, 0.001	1.497, 0.252	-225.686, 0.368	27.364, 6.464	9.559, 3.670	4.407, 1.423
Threatening an aggressive chimp or predator	982.113, 306.565	1.375, 0.500	0.314, 0.70	476.619, 197.657	893.497, 220.929	716.981, 199.027	127.861, 68.784	0.237, 0.051	77.019, 17.915	84.816, 17.448	0.001, 0.001	1.795, 0.237	-225.720, 0.468	29.947, 6.239	6.386, 2.250	4.754, 1.197
Discovering something scary	758.891, 146.810	1.385, 0.506	0.441, 0.197	429.200, 139.269	630.338, 158.077	535.946, 130.238	59.533 36.618	0.289, 0.186	67.681, 23.510	100.000, 0.000	0.000, 0.001	1.140, 0.324	-225.681, 0.613	33.100, 7.566	16.294, 4.198	6.339, 1.674

High Arousal	1486.734, 767.927	1.98, 1.123	0.465, 0.217	599.327, 378.920	996.704, 370.779	822.327, 357.173	126.006, 122.581	0.333, 0.167	76.580, 20.895	95.006, 10.905	0.001, 0.001	1.465, 0.349	-225.768, 0.474	30.483, 6.752	10.813, 5.115	5.335, 1.652
Medium Arousal	1577.933, 1070.613	3.140, 1.457	0.231, 0.136	635.323, 291.384	875.722, 375.458	763.576, 329.320	83.904, 104.569	0.209, 0.113	103.249, 40.778	87.312, 24.282	0.001, 0.001	1.546, 0.252	-225.665, 0.480	26.725, 6.322	8.590, 4.925	4.551, 1.320
Low Arousal	482.756, 400.873	3.550, 1.011	0.941, 0.030	462.106, 153.080	551.556, 192.803	501.649, 164.476	36.217, 31.461	0.143, 0.029	171.512, 53.826	55.094, 27.668	0.002, 0.005	2.064, 0.428	-224.974, 0.363	21.095, 5.874	1.462, 1.697	4.490, 0.908
Positive Valence	1094.538, 862.970	3.650, 1.359	0.208, 0.204	458.973, 214.835	672.916, 238.367	567.173, 213.366	75.184, 81.349	0.215, 0.146	133.578, 50.969	75.004, 29.944	0.001, 0.003	1.699, 0.410	-225.517, 0.600	26.484, 7.201	6.789, 5.784	4.908, 1.451
Negative Valence	1717.861, 931.710	1.960, .965	0.400, 0.189	703.538, 363.699	1074.62556, 401.490	914.284, 362.193	120.928, 130.640	0.291, 0.140	78.081, 24.670	93.685, 14.174	0.000, 0.001	1.520, 0.330	-225.632, 0.435	28.016, 7.223	9.229, 5.008	4.680, 1.490

Note. Mean and standard deviations of the acoustic parameters. *No. Call* = number of calls in the stimulus (mean), *Dur.* = total duration of calls in whole stimulus (seconds), f_0 = fundamental frequency, *Time Max.* = time of the maximum peak frequency within a call; *Peak Pos.* = relative position of the peak frequency in the call; *% Voi* = percentage of voiced frames in the stimulus, *HNR* = harmonics-to-noise ratio, *SCoG* = spectral centre of gravity

Table 3S

Table 3S. *Rotated component-loadings of each acoustic feature for the three-dimensional*

principal components' solution

	PCA1 (% of variance = 23.51)	PCA2 (% of variance = 22.60)	PCA3 (% of variance = 16.42)
ScoG	-0.67	0.599	0.388
duration	0.262	0.246	0.790
f_0 *min*	-0.006	0.901	-0.187
f_0 *max*	0.094	0.870	0.392
f_0 *mean*	0.078	0.956	0.152
f_0 *Sd*	0.062	0.092	0.735
Time Max	0.094	0.087	0.887
Peak Pos	-0.514	-0.434	-0.242
% Voi	0.658	-0.255	0.072
Jitter	-0.018	-0.250	0.025
Shimmer	-0.641	0.033	-0.162
HNR min	-0.780	0.094	-0.031
HNR max	0.781	-0.035	0.153
HNR mean	0.887	-0.024	0.054
HNR Sd	0.640	-0.296	0.068

Note. Bold type indicates parameters with highest loadings

Table 4S_A

Table 4S_A. GLMMs testing the prediction of behavioural context, arousal and valence recognition from acoustic features

Fixed Effect	Behavioural Context				Arousal				Valence			
	Estimate	SE	Z Value	Pr (>|z|)	Estimate	SE	Z Value	Pr (>|z|)	Estimate	SE	Z Value	Pr (>|z|)
Intercept	0.669	0.89	7.495	<0.001	-0.08	0.170	-0.518	0.604	0.251	0.91	2.738	0.006
SCoG	0.353	0.54	6.586	**<0.001****	0.104	0.019	5.325	**<0.001**	0.232	0.019	11.958	**<0.001**
duration	0.158	0.056	2.825	**<0.01***	0.055	0.019	2.912	**<0.05**	0.161	0.019	8.242	**<0.001**
f₀ mean	0.080	0.462	1.740	0.082	0.255	0.019	13.252	**<0.001**	0.295	0.019	15.483	**<0.001**
f₀ Sd	-0.103	0.38	-2.729	**<0.01***	0.277	0.020	13.902	**<0.001**	-0.108	0.019	-5.777	**<0.001**
HNR mean	-0.281	0.047	-6.026	**<0.001****	0.014	0.019	0.745	0.456	0	0.019	0.001	0.999
HNR max	0.131	0.040	3.310	**<0.001****	0.015	0.018	0.811	0.417	0.003	0.018	0.186	0.852

Note. *p* < 0.05; **p* < 0.001. Recognition scores for arousal level and valence are obtained in Experiment 1 and for behavioural context in Experiment 2. All coefficients represent changes in the log-odds of recognised or not recognised responses as a function of the acoustic predictors. Bold type indicates statistically significant p values (*p* < 0.05).

Table 4S_B

Table 4S_B. *Random effects in prediction of behavioural context, arousal and valence recognition*

	Context		Arousal		Valence	
	Variance	Std. Dev.	Variance	Std. Dev.	Variance	Std. Dev.
Part ID	3.284	1.812	0.071	0.267	0.176	0.419
Chimpanzee ID	0.272	0.522	1.879	1.370	0.462	0.680

Note. Context recognition is based on context-matching task (see Experiment 2), arousal level and valence recognition are based on Experiment 1.

Table 5S

Table 5S. *Model selection procedure based on Akaike's Information criterion*

Model	Behavioural Context			Arousal			Valence		
	Intercept	AICs	ΔAICs	Intercept	AICs	ΔAICs	Intercept	AICs	ΔAICs
Excluding *SCoG*	0.669	17932.6	37.7	-0.097	56155.9	27.7	0.229	54299.7	142.8
Excluding *duration*	0.690	17899.8	4.9	-0.082	56136.1	7.9	0.759	54224.4	67.5
Excluding *f₀ mean*	0.665	17894.9	**0.0**	-0.058	56306.1	177.9	0.288	54398.5	241.6
Excluding *f₀ Sd*	0.667	17899.0	4.1	-0.070	56324.2	196	0.241	54190.5	33.6
Excluding *HNR mean*	0.638	17927.7	32.8	-0.086	56128.2	**0.0**	0.251	54156.9	**0.0**
Excluding *HNR max*	0.653	17902.4	7.5	-0.090	56128.3	**0.1**	0.250	54157.0	**0.1**

Note. Context recognition is based on context-matching task (see Experiment 2), arousal level and valance recognition are based on Experiment 2. Bold type indicates models with the highest power to explain variation the dependent variable (i.e., correct or incorrect response0, based on lowest AICc.