Behav Res
DOI 10.3758/s13428-016-0736-y

Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus

Andrey Anikin[1] · **Tomas Persson**[1]

Abstract This study introduces a corpus of 260 naturalistic human nonlinguistic vocalizations representing nine emotions: amusement, anger, disgust, effort, fear, joy, pain, pleasure, and sadness. The recognition accuracy in a rating task varied greatly per emotion, from <40% for joy and pain, to >70% for amusement, pleasure, fear, and sadness. In contrast, the raters' linguistic–cultural group had no effect on recognition accuracy: The predominantly English-language corpus was classified with similar accuracies by participants from Brazil, Russia, Sweden, and the UK/USA. Supervised random forest models classified the sounds as accurately as the human raters. The best acoustic predictors of emotion were pitch, harmonicity, and the spacing and regularity of syllables. This corpus of ecologically valid emotional vocalizations can be filtered to include only sounds with high recognition rates, in order to study reactions to emotional stimuli of known perceptual types (reception side), or can be used in its entirety to study the association between affective states and vocal expressions (production side).

Keywords Emotion · Nonlinguistic vocalizations · Naturalistic vocalizations · Acoustic analysis

Language use is so central in our characterization of what it means to communicate as a human that it is easy to overlook the roles of nonspeech vocal sounds, such as laughs, screams, grunts, and so forth. Nevertheless, such sounds indeed play their part in daily communication alongside language. But where do they come from?

✉ Andrey Anikin
andrey.anikin@lucs.lu.se

[1] Division of Cognitive Science, Department of Philosophy, Lund University, Box 192, SE-221 00 Lund, Sweden

Neurological evidence suggests that nonlinguistic vocalizations are controlled by neural circuitry that is common to all mammals and distinct from the evolutionarily younger structures responsible for the production of language (U. Jürgens, 2009). These two systems may be thought of as two separate pathways connecting the cortex with the laryngeal motor neurons that control the vocal cords. The limbic pathway goes from the anterior cingulate cortex, via the periaqueductal gray area, to the reticular formation. It is found in all mammals and triggers species-specific emotional vocalizations. The basic acoustic structure of such vocalizations is predetermined by the pattern-generating neurons in medullary reticular formation and cannot be modified voluntarily (Hage, Gavrilov, & Nieder, 2013; U. Jürgens, 2009). The second path, with direct projections from primary motor cortex (M1) to motor neurons in the reticular formation, enables fine voluntary control of laryngeal muscles, which is necessary for the production of complex learned vocalizations. Direct cortical projections from M1 to laryngeal motor neurons are thought to be absent in nonhuman primates (U. Jürgens, 2009) and weak in other mammals (Arriaga, 2014; Petkov & Jarvis, 2012). In general, monosynaptic projections from motor cortex to laryngeal motor neurons appear to enable precise voluntary control of vocalizations in vocally gifted mammals, including humans (Schusterman, 2008).

Individuals with a lesion in the area of motor cortex projecting to laryngeal motor neurons suffer from a complete loss of fine voluntary control over the vocal cords and cannot speak, while spontaneous vocalizations such as moaning, laughing, and crying are preserved. On the contrary, individuals with bilateral lesions of the anterior cingulate have reduced motivation to speak, although the motor control is preserved (U. Jürgens, 2009). Moreover, congenitally deaf human infants produce species-specific calls without any auditory feedback, whereas language-like babbling in such infants

Published online: 29 April 2016

is delayed or absent (Scheiner, Hammerschmidt, Jürgens, & Zwirner, 2006). The limbic and cortical pathways are thus to some extent functionally distinct. This is consistent with the interpretation that the mammalian vocalization system is *production-first*—that is, largely "hardwired," affectively triggered, feedback-independent, and not heavily dependent on social learning. In contrast, learning plays a central role in *reception-first* vocal systems characterized by flexible acoustics, such as human language (Owren, Amoss, & Rendall, 2011).

The distinction between these two vocal systems is therefore not that of voluntary versus involuntary vocalizations (Simonyan & Horwitz, 2011). The point, rather, is that the basic acoustic structure of nonlinguistic, emotionally triggered vocalizations appears to be "hard-coded" in the brain stem. Within this basic template, dialectal variation mediated by social learning is certainly possible. This learning may even happen in utero, so that newborn babies already cry with the typical prosody of their mother's native language (Mampe, Friederici, Christophe, & Wermke, 2009). However, such variation by no means prevents the basic acoustic type from being recognized as crying (Newman, 2007), just as the pant-hoots of wild chimpanzees are readily recognizable as pant-hoots, although subtle group-specific acoustic differences have been described (Crockford, Herbinger, Vigilant, & Boesch, 2004).

Even if laughs and other nonspeech sounds are neurologically distinct from language, both systems must coexist harmoniously in order to enable successful communication. When we speak, the intonation is shaped by language-specific prosodic rules, such as rising intonation in questions or prosodic marking of emphasized words (Scott, Sauter, & McGettigan, 2009). Language-internal factors thus constrain prosodic markers of emotion in speech. But if the mammalian vocalization system is relatively inflexible, it is also plausible that the more malleable language would have to adapt to the structure of innate vocalizations, allowing both systems to achieve their full communicative potential without a clash. If this is true, the prosody of verbal utterances can be expected to follow, or at least not to contradict, the vocal patterns associated with nonspeech vocalizing.

A number of cross-cultural studies have investigated the universality of emotional prosody in speech (Banse & Scherer, 1996; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). The overall conclusion from this research is that strong regularities in the acoustic characteristics of speech depend on the emotion portrayed. Furthermore, listeners are successful at guessing expressed emotions even in languages they are not familiar with, or in meaningless pseudophrases, although accuracy tends to be lower in cross-cultural comparisons than in material collected and tested within one culture (Bryant & Barrett, 2008; Neiberg, Laukka, & Elfenbein, 2011; Scherer, Banse, & Wallbott, 2001). The universality of certain prosodic features suggests the presence of something species-typical, making the mammalian vocalization system the prime suspect. It is therefore of interest to investigate this system directly, and not only in its interaction with language.

Until recently, vocal markers of emotion in humans have primarily been studied by linguists, who have focused on prosody in verbal utterances rather than on purely nonverbal sounds. Over the last few years, however, researchers have begun to look into nonlinguistic (also referred to as *nonverbal* or *nonspeech*) emotional vocalizations—sounds with little or no phonemic structure (Belin, Fillion-Bilodeau, & Gosselin, 2008; Hawk, Van Kleef, Fischer, & Van der Schalk, 2009; Lima, Castro, & Scott, 2013; Schröder, 2003; Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009).

Most studies report some in-group advantage: Emotion is recognized more accurately when the producer and the receiver belong to the same sociocultural group (Elfenbein & Ambady, 2002; Koeda et al., 2013; Laukka et al., 2013; Sauter & Scott, 2007). Nevertheless, cross-cultural recognition typically remains better than would be expected by chance. These studies vary widely with respect to the chosen emotional categories, number of callers, and elicitation techniques, as well as in the presence of pseudoverbal utterances with some phonemic structure (*Yuck! Wow!*), experimental design, and tested linguistic groups. As a result, it is difficult to draw firm conclusions about the universality of these vocalizations. The tentative consensus appears to be that displays of negative emotions are more universal, whereas positive emotions show more cultural variation (Gendron, Roberson, van der Vyver, & Barrett, 2014; Sauter, Eisner, Ekman, & Scott, 2010).

Despite the differences mentioned above, one thing most published studies of nonlinguistic vocalizations have in common is that the stimuli are produced on demand by volunteers or actors, the justification being that such portrayals are intended to be widely understood, if also exaggeratedly stereotypical (Banse & Scherer, 1996). Concerns about the validity of playacted vocalizations, however, are often voiced in the literature (Batliner, Fischer, Huber, Spilker, & Nöth, 2000; Douglas-Cowie, Campbell, Cowie, & Roach, 2003; Gendron et al., 2014; Parsons, Young, Stein, Craske, & Kringelbach, 2014). By definition, playacted vocalizations are produced voluntarily, with the explicit intent to communicate a particular message. They are therefore likely to be dominated by so-called *pull* effects, such as cultural conventions and self-conscious impression management (Scherer & Bänziger, 2010). In contrast, someone vocalizing in real-life situations is often reacting to an unexpected and dramatic situation, such as a sudden fright. There is presumably little time or incentive to deliberately fine-tune such vocalizations, making the internal *push* effects more prominent. Furthermore, the typical contexts evoking a particular emotion, as well as the display rules governing what is

appropriate, may differ across cultures, whereas the structure of the display itself may well be invariant (Ekman & Friesen, 1969).

Despite these concerns over the validity of playacted vocalizations, their real-life counterparts appear to be almost unexplored. Fortunately, this is beginning to change, as a series of recent studies have analyzed spontaneous laughter and found some acoustic differences between the spontaneous and volitional varieties (Bryant & Aktipis, 2014; Lavan, Scott, & McGettigan, 2015). In addition to laughter, Parsons and coauthors (2014) also collected authentic sounds of crying and neutral vocalizations from online videos. There is thus a growing interest in collecting real-life emotional displays and comparing them with actor portrayals.

The bad news is that working with naturalistic emotional vocalizations raises a number of complex methodological issues. For one thing, it is harder to ensure their acoustic quality than for sounds recorded with professional equipment in a soundproof chamber. There is also limited control over potential confounds, such as the speaker's sex, age, and native language, background noise, the presence of other people, type of interaction, and so forth (Douglas-Cowie et al., 2003; Scherer, 2013). In addition, where playacted vocalizations are elicited by asking the person to portray a particular emotion, in the case of naturalistic vocalizations, the "true" underlying affective state of the caller may be hard to ascertain.

On the other hand, studying vocalizations in natural contexts may offer benefits that cannot be achieved with artificially elicited stimuli. Given the evolutionary arms race between signalers and receivers, authentic emotional markers can be expected to be honest, hard-to-fake signals (Searcy & Nowicki, 2005). The documented differences between authentic and volitional smiles (Ekman, Davidson, & Friesen, 1990) and laughs (Bryant & Aktipis, 2014; Lavan et al., 2015) raise the possibility that systematic differences may also exist for other emotional signals. Furthermore, overcoming the methodological complications associated with naturalistic data has practical applications. Several engineering projects have approached the task of developing acoustic and statistical techniques for the rapid and reliable online classification of emotional vocalizations in the context of human–machine interaction (Batliner et al., 2000; Breazeal & Aryananda, 2002). Social robots and voice recognition software implement computational models for classifying real-life materials and have to deal with all the problems described above. Machine learning also requires validated databases of realistic emotional stimuli, such as the corpus of naturalistic nonlinguistic vocalizations presented in this study.

To recapitulate, there is neurological evidence that nonlinguistic vocalizations predate language and are controlled by distinct circuitry in the brain. We therefore hypothesized that their natural form should be more apparent in spontaneous, emotionally triggered nonlinguistic vocalizations. To collect suitable material, we chose to work with real-life vocalizations, hypothesizing that they might be more spontaneous and free from voluntary modulation than are actor portrayals. If this is true, naturalistic sounds may also be less culture-specific and more useful for phylogenetic reconstruction of the evolutionary roots of human vocalizations.

Method

Compilation of the corpus

Vocalizations (N = 260) were obtained from online videos (www.youtube.com). We were aiming to find real-life examples of sounds from the emotional categories previously investigated with actor portrayals. Whenever possible, we attempted to find situations similar to scenarios that had been used to elicit emotional vocalizations in previous studies. For example, sounds of disgust are typically elicited with such imaginary scenarios as eating rotten food (Sauter, Eisner, Ekman, & Scott, 2010) or inadvertently putting a hand in vomit (Lima et al., 2013). Different "food challenges" or videos of people declogging a toilet provided comparable real-life examples (for a full list, see Table 1). In practice, however, the availability of materials was the main criterion that determined which contexts were used as examples of each emotion.

Priority was given to eliciting contexts that were:

(1) Sudden. Less time available for deliberation presumably minimized the likelihood of impression management and posing for the camera.
(2) Unambiguous. This minimized the risk of misunderstanding the emotional content of vocalizations in the corpus at the collection stage.
(3) Powerful. Authentic calls, seen as "expensive" honest signals (Searcy & Nowicki, 2005), are more likely to occur in situations associated with genuinely high arousal, such as sudden fright, acute pain, supreme physical effort, and so on. Moreover, high-arousal vocalizations are known to appear more authentic to listeners (Lavan et al., 2015).

All chosen video clips were unambiguous and associated with medium to high arousal, but not all of the eliciting contexts were sudden. The degree of our certainty about the authentic, spontaneous, and nonposed nature of the vocalizations varied accordingly: Certainty was highest for sounds of fear, pain, amusement, and joy, and lowest for the sounds of disgust and pleasure.

We labeled each vocalization, on the basis of the caller's facial expression, verbal comments, and other contextual cues, as *amusement*, *anger*, *disgust*, *effort*, *fear*, *joy*, *pain*, *pleasure*,

Table 1 Numbers of vocalizations in the corpus ($N = 260$) per emotion and gender–age group

Emotion	Contexts	Qualitative Acoustic Description	Number of Sounds (Adult Man/ Woman or Child)
Amusement	Pranks, failed stunts, distorting web camera, social play	Laughs	25 (13/12)
Anger	Malfunctioning computer, losing a game or detecting a cheater, tantrum	Roars or noisy screams, growls	25 (13/12)
Disgust	Unblocking a clogged toilet, food challenges (Surströmming, baby food)	Grunting, retching noises, "Aah," "Ugh"	25 (12/13)
Effort	Weightlifting, amateur gymnastics (pull-ups, push-ups)	Grunts, roars	25 (12/13)
Fear	Scare pranks, bungee jumping, "haunted house" attraction, spiders	Screams	25 (10/15)
Joy	Opening exam results, "We're pregnant!" videos, sport fans cheering after a score	Screams, laughs, roars, sighs	48 (18/30)
Pain	Men: failed stunts, sport injuries; women: giving birth	Roars, screams, moans	38 (19/19)
Pleasure	Having sex or masturbating (usually without a video track, so that the authenticity of these vocalizations cannot be guaranteed)	Moans, grunts	25 (12/13)
Sadness	Complaining and crying about someone's death, broken relationship, a sad movie, etc.	Crying with tears	24 (11/13)

or *sadness*. Four of them (*anger*, *disgust*, *fear*, and *sadness*) are among Ekman's six basic emotions (1992). They are routinely investigated in studies of emotional expression, and it was straightforward to find suitable video clips with vocalizations related to these emotions. *Amusement* in our corpus corresponds to laughing at something funny. *Pleasure* is less established in the literature, but it has sometimes been investigated as an independent category in studies of nonlinguistic vocalizations (e.g., Belin et al., 2008; Lima et al., 2013; Simon-Thomas et al., 2009). *Pain* has seldom been considered in acoustic research (but see Belin et al., 2008), and *effort* apparently not at all. We do not claim that these two states are necessarily emotions, but both are commonly associated with nonlinguistic vocalizing, and both are straightforward to identify on the basis of the context. We therefore included them in order to explore the entire range of nonlinguistic vocalizations with identifiable contexts. Another emotion intended for the study—surprise—proved impossible to distinguish from either *fear* or *joy* on the basis of the context, and it was therefore dropped. Finally, positive experiences not related to amusement or sensual pleasure were labeled *joy*. A more detailed classification was attempted but proved impractical, since the semantic ambiguity between achievement, relief, pleasant surprise, and general happiness at the stage of selecting and labeling video clips made it more conservative to treat all four as a single emotion.

These nine categories (i.e., *amusement*, *anger*, *disgust*, *effort*, *fear*, *joy*, *pain*, *pleasure*, and *sadness*) in effect span the entire range of naturalistic nonlinguistic vocalizations with unambiguous emotional value that we have been able to discover in online videos.

The choice of emotional categories and contexts was validated in a survey administered to 11 participants in English ($n = 7$), Russian ($n = 2$), and Swedish ($n = 2$). The respondents were asked to (1) list typical examples of situations in which each of the nine emotions is likely to be experienced, and (2) classify the contexts listed in Table 1 into these nine emotional categories. Overall, the results confirmed the experimenters' classification. Notably, a clear semantic distinction was drawn between *amusement*, which was associated with laughing at something funny, and *pleasure*, or sensual enjoyment of food or sex. *Joy* was less clear-cut: Examples of this emotion suggested by the respondents included a wide variety of good events. Several contexts chosen as examples of one emotion were described by the respondents as a mixture of two or more emotions. In particular, giving birth (labeled as *pain* by the experimenters) was described by respondents as a mixture of *pain* and *effort*, and having sex or masturbating (labeled as *pleasure*) as a mixture of *pleasure* and *effort*. These nuances highlight the difficulty of assigning a single emotional label to each context.

For some video clips the country of origin was not identifiable, but overall, an overwhelming majority of the vocalizations in the corpus are from Western, and above all English-speaking, countries. Typically, only one vocalization per caller was taken from each video. Fewer than 5% of the vocalizations in the corpus are from prepubescent children, and preliminary analyses showed that women and children were sufficiently similar acoustically to be placed in the same gender–age category (as opposed to adult men).

Preparation of audio clips

Each audio clip represents a single continuous vocal element surrounded by silence (*syllable*) or a train of syllables that occur within the space of a few seconds, typically within one exhalation (a *call* or a *bout*: Bachorowski, Smoski, & Owren 2001; Bohn, Schmidt-French, Ma, & Pollak, 2008). The durations range from about 400 ms to 9.5 s (mean = 2.1 s, median = 1.5 s, $SD = 1.7$). Published corpora of

nonlinguistic vocalizations usually include sounds of 0.5 to 2 s in duration (Belin et al., 2008; Lima et al., 2013; Sauter, Eisner, Calder, & Scott, 2010). However, both mono- and polysyllabic vocalizations over 3 s in length were common in the video clips, justifying their inclusion. Some vocalizations had to be truncated because of external noises. We took the precaution of not using the sound's duration or the number of syllables as predictors in our supervised classificatory models, and we were careful to ensure that the entire bout was produced in the same emotional state.

All audio clips were converted to 16-bit, mono .wav files with a sample rate of 44100 Hz and normalized for peak amplitude with Audacity (http://audacity.sourceforge.net). The original sampling rates varied greatly, limiting the usefulness of certain spectral variables such as quartiles of energy distribution. Background noise was manually removed using a combination of filtering methods: low- and high-pass filters (e.g., to remove the wind), notch filters (to remove pure tones or sounds with a few strong harmonics), and the "noise profile" feature (to remove constant broadband noise). Clicks were removed by deleting a few milliseconds of audio. Filtering was heaviest for screams of fear and joy, because of the boisterous contexts in which these vocalizations were encountered, but all emotional categories required some filtering. Vocalizations were rigorously selected for the highest attainable audio quality, and hundreds were discarded after processing because of irremovable noise.

Methods of acoustic analysis

Acoustic features were extracted in both PRAAT (version 5.3; www.fon.hum.uva.nl/praat/) and R (R Development Core Team, 2014). The range of the fundamental frequencies (F0) in the corpus was unusually broad: from 75 to over 3000 Hz. Although completely voiceless vocalizations were excluded, many of the sounds were atonal, further complicating pitch measurement. To make measurements consistent, we used the same pitch floor and ceiling for all vocalizations (75 and 3500 Hz, respectively) and checked pitch-related variables (mean, minimum, maximum) manually. *Pitch* in this study thus corresponds to F0 for relatively tonal sounds, or to the lowest dominant frequency band for sounds with no detectable harmonics.

Segmenting vocalizations into syllables based on silences between them or absolute amplitude thresholds produced unsatisfactory results. We therefore developed a custom second-pass algorithm, implemented in R, which searches smoothed-amplitude envelopes for local maximums (vocal bursts) that were high enough relative to the global mean amplitude of the entire call, steep enough, and spaced far enough relative to the median syllable length within the same call. The exact thresholds were optimized to achieve a compromise between detecting too many and too few bursts. This algorithm measured the average spacing of vocal bursts (*mean interburst interval*) and their regularity (*SD of interburst interval*).

In addition, a number of commonly used acoustic variables were extracted automatically in PRAAT, including general descriptives (duration, mean, and *SD* of amplitude). Peak frequency, mean frequency, spectral tilt, and the difference in spectral energy between bands above and below certain thresholds (200, 500, 1000, and 2000 Hz) were extracted using fast Fourier transform of the entire sound clip. The harmonics-to-noise ratio (HNR) was measured for voiced frames using the cross-correlation method, silence threshold 0.1, and pitch floor 75 Hz. Jitter, shimmer, the number of voice breaks, and the proportion of unvoiced frames were extracted from PRAAT's voice report. We also manually encoded the intonation and call type of each vocalization for descriptive purposes.

Statistical modeling based on acoustic measurements

All statistical analyses were performed in R. Given the large number of potential predictors, some nonnormally distributed and others categorical, the main classification algorithm we used was a nonparametric method: random forests (RF; Breiman, 2001). RF models consist of a large number of decision trees, with several predictors used at each branch of each tree. The final classification of an observation is made by combining the votes of individual decision trees.

Estimates of recognition accuracy in RFs are calculated for out-of-bag observations: each individual tree is trained on approximately two thirds of the data, while the remaining third are left for the cross-validation. There is thus no need to split the data into a training set and a testing set manually. The exact model is slightly different every time the algorithm is executed, and therefore confidence intervals (CIs) for recognition accuracies were calculated by refitting the model 1,000 times with the same sample (but with different training and testing sets for each tree).

Error rates in supervised RF models depend on the prior probabilities of group membership, and therefore the sample was stratified by the smallest category. Individual variables, rather than principal components or factor scores, were used for prediction, since this was associated with considerably better performance of the classification models (as was also reported by Wadewitz et al., 2015). The contribution of each variable is estimated internally by RF models by excluding this variable from the pool of predictors and measuring the loss of classification accuracy. Using this metric and attempting to make the model as transparent as possible without sacrificing overall prediction accuracy, we short-listed only a few best predictors (six in the final model).

In addition to simple hit rates and false alarm rates, we report corrected chance levels and the unbiased hit rate (H_u). H_u was calculated as the ratio of the squared number of correct

classifications to a product of the column and row marginals in the confusion matrix (Wagner, 1993). For example, in Table 3a below, 18 out of 25 sounds of amusement were classified correctly, seven sounds of joy were misclassified as amusement (false alarms), and seven sounds of amusement were misclassified as joy (misses). H_u for amusement was therefore 18 × 18/(25 × 25) = 52%. The corrected chance level was calculated as the product of the column and row marginals, divided by the squared total number of observations: For amusement in Table 3a, this is 25 × 25/(260 × 260) = 0.9%. However, the most certain way to avoid bias is to report the actual confusion matrices (Bänziger, Mortillaro, & Scherer, 2012), and therefore they are also presented in full.

Rating experiment

Procedure A rating experiment was written in html/javascript and made available online. Participants were asked to rate approximately 100 sounds presented in random order, which normally took 15–20 min. Sounds could be replayed, but after five repetitions a pop-up alert reminded participants to respond quickly. From zero (skipping the sound) to nine emotional labels could be chosen to describe each sound. The labels were the same as in Table 1: *amusement*, *anger*, *disgust*, *effort*, *fear*, *joy*, *pain*, *pleasure*, and *sadness*. Each emotion was scored by moving a slider on a horizontal visual analog scale marked *None* to *Strong*.

Participants Ninety respondents rated at least ten sounds and were included in the analysis. They performed the test in English (n = 39), Swedish (n = 36), or Russian (n = 15). The sample was further subdivided by location (IP address): Scandinavia (n = 38), UK/USA (n = 16), Russia (n = 15), Brazil (n = 10), and "Other" (n = 11, primarily Europe outside Britain and Sweden). To account for differences in language and location, all available trials were grouped into five linguistic–cultural groups, shown in Table 2. English-speaking participants were recruited as the "in-group" (the same as the callers in the corpus), and the remaining groups were contacted on the basis of the availability of participants while attempting to maximize cultural and linguistic diversity.

Participation was voluntary and anonymous, and beyond the choice of language and IP address, no demographic information was recorded. Swedish participants were primarily recruited on the campus of the University of Lund, whereas international participants were recruited online and through personal contacts. Recruitment was stopped once the planned number of 30 ratings per sound had been achieved.

Statistical analysis of the rating task Since every participant rated a random selection of sounds rather than the entire corpus, recognition accuracy was analyzed both individually and collectively. To calculate individual accuracy, each time a participant rated a sound, the perceived emotion of this sound was defined as the category with the highest score. This classification was considered correct if it was the same as the context-based label of the sound (see the Compilation of the Corpus section above), and incorrect otherwise. The accuracy of recognition by the popular vote was calculated only once for each sound by averaging its scores on nine emotions across all participants who had rated this sound (i.e., typically about 30 people). All but five participants used the entire 0–100 scale, without avoiding extreme values, and therefore no adjustment to the personal range of responses was made when averaging the scores.

All models of individual accuracy presented in the text included fixed effects, such as caller's sex or the rater's linguistic–cultural group, and two random effects: participant and sound. We calculated p values using generalized linear mixed-effects models (GLMM) of the binomial family and 95% confidence intervals using the Markov-chain Monte Carlo method in the Stan computational framework (Stan Development Team, 2014).

It is worth reiterating that "correct classification" in this study meant that human raters, who had heard the sound but had not seen the clip, chose the same emotional label as the researchers, who did have access to contextual information. This is not equivalent to recognition accuracy in studies in which the sounds are produced by actors instructed to portray a particular emotion, which is then recognized (or not) by the raters.

The raw data and PRAAT and R scripts used for the extraction of acoustic features and the statistical analysis are available in the supplementary materials. The audio files can be provided upon request or can be downloaded from www.cogsci.se/personal/results.html.

Results

Rating task

Overall classification accuracy Each of the 260 sounds in the corpus was rated by 30.5 ± 4.6 participants. Recognition accuracy by the popular vote was higher than by individual raters (59% vs. 46%), but in both cases the hit rates varied greatly per emotion. *Amusement*, *effort*, *sadness*, *fear*, and *pleasure* were classified correctly (i.e., consistently with the context), with hit rates by the popular vote above 70%. The second group of emotions, with hit rates between 45% and 55%, included *disgust*, *anger*, and *pain*. Finally, *joy* proved deeply problematic, with hit rates under 25% (Table 3). Some emotions, such as *fear*, were seldom missed by the human raters, but with many false alarms. Others, like *disgust* and *sadness*, were associated with very few false alarms. The

Table 2 Linguistic–cultural groupings of individual responses (~94 responses per participant)

Location (IP Address)	Language			Linguistic–Cultural Group
	Russian	Swedish	English	
Russia	1,384	–	–	Russia (1,384)
Scandinavia	–	3,347	197 (added to "Other" group)	Sweden (3,347)
UK/USA	–	–	1,288	UK/USA (1,288)
Brazil	–	–	916	Brazil (916)
Other	–	–	821	Other (1,018)

confusion matrix of individual decisions in the lower block of Table 3 shows the same trends, but with lower hit rates.

Since recognition accuracy varied significantly per emotion and per sound, we controlled for the amount of disagreement among the raters by analyzing the entropy of popular votes. For some sounds, only one emotion had a high average score (low entropy, high degree of agreement among raters), whereas for other sounds, several emotions had high scores (high entropy, low degree of agreement). When all sounds with above-median entropy were removed from the analysis, recognition accuracy for the remaining half of the corpus became nearly perfect for all emotions except *joy*, *pain*, and *anger*. Sounds in these three categories were thus misclassified in a highly consistent manner by most participants, suggesting either that these emotions are not well represented in the corpus or that they recruit vocalizations more characteristic of other emotions, and thus cannot be recognized reliably from spontaneous vocalizations without a verbal component or contextual information.

The category of *pain* can be subdivided into acute injury and giving birth. Only 35% of the sounds made by a woman in labor were classified by human raters as *pain*, whereas 41%

Table 3 Classification of the corpus and confusion matrices for the popular vote (A) and individual raters (B)

A.		Perceived Emotion Chosen by the Popular Vote									Hit Rate,% [95% CI]	False Alarms,%	H_u,%	Corrected Chance,%
		amsm	angr	dsgs	effr	fear	joy	pain	plsr	sdns				
Context-based emotion	amsm	**18**					7				72 [55.5, 88.2]	3.0	51.8	0.9
	angr		**11**		2	11		1			44 [24.1, 63.1]	3.0	26.9	0.7
	dsgs			**14**	3	1	1	2	4		56 [35.3, 74.7]	0	56.0	0.5
	effr		2		**19**	2		2			76 [59.3, 90.2]	8.9	36.1	1.5
	fear		1		1	**23**					92 [79.2, 97.8]	16.2	34.7	2.3
	joy	7	2		4	14	**11**	3	6		23.4 [11.7, 34.6]	4.2	12.9	1.4
	pain		2		6	10	1	**17**	2		44.7 [28.7, 60.9]	5.0	27.2	1.6
	plsr				4			1	**20**		80 [65.1, 92.6]	5.5	48.5	1.2
	sdns				1			2	1	**21**	84 [69.3, 94.9]	0	84.0	0.8
Total											59.2 [55.4, 62.9]	5.1	42.0	1.2
B.		Perceived Emotion Chosen by Individual Raters									Hit Rate,% [95% CI]	False Alarms,%	H_u,%	Corrected Chance,%
		amsm	angr	dsgs	effr	fear	joy	pain	plsr	sdns				
Context-based emotion	amsm	**429**	1	1	1	6	290		11	31	55.7 [44.5, 65.3]	4.6	31.5	0.9
	angr	15	**272**	27	100	231	9	90	4	6	36.1 [26.6, 46.1]	5	15.6	0.7
	dsgs	16	31	**284**	139	24	26	87	124	30	37.3 [25.4, 45.6]	3.1	21.0	0.6
	effr	6	68	67	**405**	48	1	75	44	6	56.2 [44.2, 65.5]	9.5	20.8	1.2
	fear	21	68	15	28	**520**	35	73	10	13	66.4 [56.2, 75.6]	13.4	23.3	1.8
	joy	227	82	35	105	340	**275**	132	163	66	19.3 [12.7, 22.4]	6.4	7.7	1.6
	pain	22	89	49	164	275	35	**396**	88	41	34.2 [25.9, 41.7]	9	13.4	1.8
	plsr	10	11	20	116	6	12	69	**535**	24	66.6 [56.2, 75.3]	6.6	35.3	1.3
	sdns	12	7	7	36	31	8	85	30	**562**	72.2 [63.4, 80.7]	3	52.1	1.0
Total											46.2 [40.5, 48.4]	6.7	24.5	1.2

Hit rate (correct detection): percentages of sounds classified as category *c* out of all sounds that really belonged to category *c*. False alarm rate (incorrect detection): percentages of sounds classified as category *c* out of all sounds that did not belong to category *c*. Amsm = amusement, angr = anger, dsgs = disgust, effr = effort, plsr = pleasure, sdns = sadness

were classified as *fear* and 11% as *effort*. In contrast, 52% of the sounds corresponding to an acute injury were classified by the popular vote as *pain*. It is hard to say whether the context or gender of the caller was decisive, but it is also helpful to consider the acoustic types of these sounds. It turns out that 8/15 "screams of pain" were classified by the popular vote as *fear*, 1/1 "sigh of pain" as *pleasure*, and 4/11 "roars of pain" as *effort*.

As for *joy*, these 47 sounds come from three contexts: opening exam results, getting good news of a daughter's pregnancy, and witnessing a spectacular score by the favorite sports team. Overall accuracy was low for all three: about 30% for exams and pregnancy videos, and only 8% for sport fans. About half of the sounds of *joy* were correctly described as being positively valenced, so to some extent the problem lay in making subtle distinctions between positive states. However, joy was also commonly confused with *fear*, *pain*, and *anger* (Table 3). Again, a closer look at the acoustic types helps understand the confusion: 4/5 "laughs of joy" were classified by the popular vote as *amusement*, 10/14 "screams of joy" as *fear*, and 5/9 "sighs of joy" (or relief) as *pleasure*. In other words, it seems highly probable that sounds were often (mis)classified according to their acoustic type.

Effects of the rater's linguistic–cultural background Considering that the vocalizations in the corpus were predominantly produced by native English speakers, we analyzed the effect of the raters' linguistic–cultural background on recognition accuracy.

When we calculated the average score of each sound on the context-based emotion separately for each linguistic–cultural group, the correlation between these five groups was considerable (Cronbach's $\alpha = .91$). Overall, individual accuracies in the five linguistic–cultural groups were also similar: UK/USA 46.4%, Sweden 46.7%, Russia 44.9%, Brazil 43.6%, Other 49.0%. The effect of linguistic–cultural group in a model without language–emotion interaction was negligible (likelihood ratio test in a GLMM: $L = 3.4$, $df = 4$, $p = .49$).

Participants from English-speaking countries were thus no better at recognizing the underlying emotions than were those from other cultures, including Russians and Brazilians. A retrospective power analysis showed that the probability of obtaining a statistically significant ($p < .05$) main effect of linguistic–cultural group, given the observed differences between them (*SD* of overall accuracies ~2%) and the sample size of 90 participants, was ~42%. To achieve a power of >80%, 270 participants would be needed. However, if the *SD* of overall recognition accuracy across groups were at least 5%, 90 participants would ensure a power of ~96%. In other words, if there had been a group effect large enough to be of even minimal practical importance, it would probably have been detected.

Despite the similar overall accuracies in all groups, we did find evidence for an interaction between linguistic–cultural group and emotion as predictors of accuracy ($L = 182.2$, $df = 32$, $p < .001$). The only emotion for which there was a major difference between groups was *amusement*, for which Sweden and Brazil stand out as having unexpectedly low hit rates (Fig. 1). This may be related to language-specific semantic nuances: In all groups, laughs were classified as either *amusement* or *joy*, with negligible confusion with any other emotion. Swedes and Brazilians were more likely to refer to laughs as *joy* rather than *amusement*, whereas for other groups *amusement* was the preferred term.

Effects of the caller's sex and the sound's duration The average individual accuracy was 44.2% when the caller was an adult man, and 48.0% when the caller was a woman or a child. This difference was not statistically significant after controlling for the sound's duration (likelihood ratio test: $L = 0.89$, $df = 1$, $p = .35$). However, there was a strong interaction between the caller's sex and emotion as predictors of accuracy ($L = 23.9$, $df = 8$, $p = .002$). *Pleasure*, *fear*, and *joy* tended to be recognized more accurately when the caller was a woman or child, whereas for *anger* and *pain*, adult male callers had some advantage (Fig. 2).

The sound's duration was a strong predictor of recognition accuracy ($L = 12.1$, $df = 1$, $p < .001$). For each extra second of duration, the odds of a correct response increased by 24%. Interestingly, the effect of duration on accuracy seems to have been driven by only two emotions: *joy* and *pleasure*. Although the recognition accuracy of most emotions did not change much with increasing duration, *joy* and *pleasure* were recognized by human raters considerably better if the sound was longer than ~3 s. To verify the effects of the caller's gender and sound's duration on recognition accuracy, more sounds will need to be tested.

Co-occurrence of verbal labels Considering that the verbal labels of emotion were chosen by the researchers, it was important to investigate their use by participants. Correlations of the scores on all emotions were investigated using exploratory factor analysis. When analyzing scores from individual trials, we found no discernible factor structure, and each emotion was best described by its own factor. If we looked at the scores averaged across all participants, however, *amusement* and *joy* formed a single factor with loadings .92 and .93, respectively, and the remaining emotions seemed to be independent. This means that, although individual participants used all nine labels independently and consistently, the distinction between *amusement* and *joy* was consistent for each participant, but inconsistent across participants.

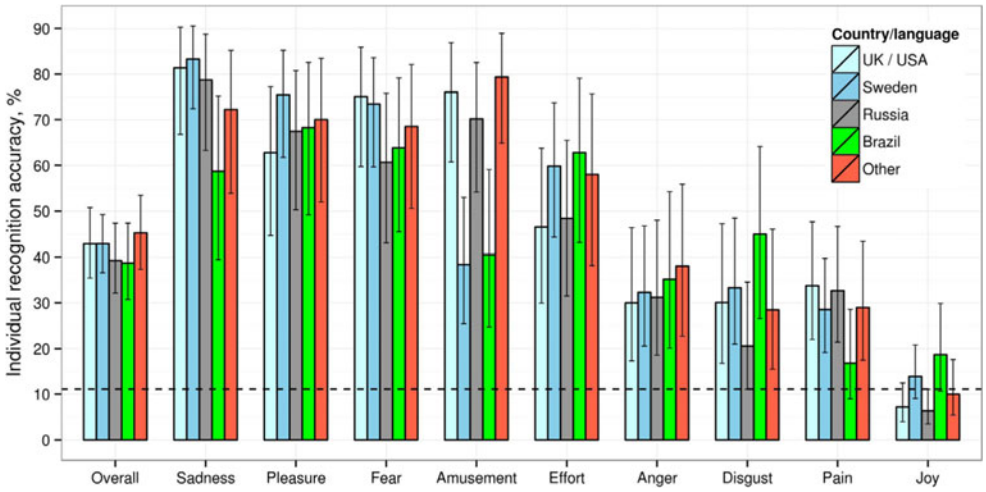

Fig. 1 Individual recognition accuracies by emotion and the rater's linguistic–cultural group: Posterior medians and 95% confidence intervals. The dotted line shows the probability of guessing correctly by chance (11%)

Acoustic analysis

Supervised classification models A series of supervised RF models were explored and streamlined to shortlist a small number of the most important acoustic predictors of emotion. The final model had a cross-validation accuracy of ~46% and included only six variables: mean pitch, mean HNR, mean interburst interval, *SD* of interburst interval, energy above minus energy below 500 Hz, and *SD* of amplitude (see the Methods of Acoustic Analysis section above). The classification of sounds in the corpus by this model is presented in Table 4. With more predictors included in the model, recognition accuracy improved by no more than a few percentage points. Simulations with permuted datasets confirmed that there was no overfitting: The probability of an RF model correctly classifying a sound by chance was only 11.4%, which was close to the expected level of 11.1%.

Just as with ratings by human participants, the most problematic category for the supervised RF models was *joy*. If the sounds assigned to the *joy* category were excluded from the corpus, classification accuracy by the acoustic models improved by 10%, and *amusement* achieved nearly perfect recognition rates. Although for most emotions low-entropy (i.e., certain, unambiguous) classification decisions are usually correct, *joy* and *pain* were the only two emotions for which recognition accuracy dropped as we focused on low-entropy decisions. *Joy* and *pain* thus appear to be a hodgepodge of sounds "borrowed" from other emotions. Sounds of *joy* recruit vocalizations typically associated with *amusement* and *fear*, whereas sounds of *pain* recruit vocalizations of *effort* and *fear*. This is entirely in line with the classification mistakes made by human raters.

Since the RF model operates within a six-dimensional space, it cannot be visualized directly. A simplified classificatory model in Fig. 3 illustrates the effects of two key predictors—pitch and harmonicity. Naturally, other acoustic parameters also contribute to classification. For example, laughs are recognized primarily by their frequent, regularly spaced bursts.

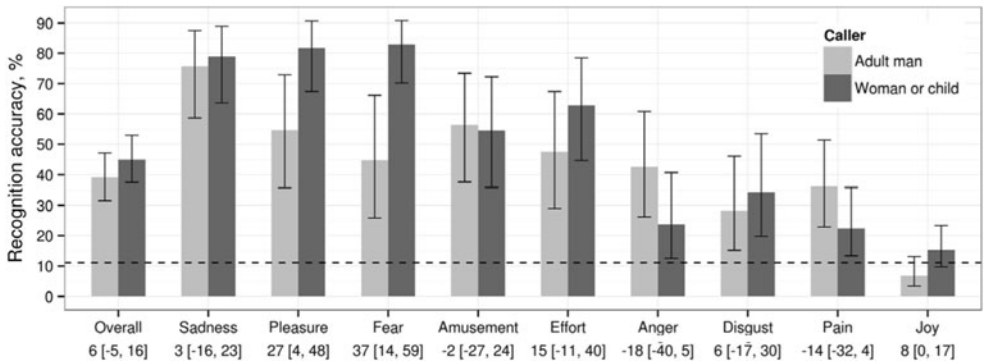

Fig. 2 Individual recognition accuracies by emotion and the caller's gender–age group: Posterior medians and 95% CI's. The most credible difference (%) between gender–age groups is shown underneath each bar, with 95% CI. The dotted line shows chance level (11%)

Table 4 Classification of 260 sounds in the corpus by a random-forest model with six acoustic predictors

		Emotion Predicted by the Model									Hit Rate,% [95% CI]	False Alarms,%	H_u,%	Corrected Chance,%
		amsm	angr	dsgs	effr	fear	joy	pain	plsr	sdns				
Context-based emotion	amsm	**79**		4		4	8.1	0.9	0.2	3.8	79 [76, 80]	4.5	51.5	1.1
	angr		**55.7**	10.2	6.2	15.4	4.6	4	4		55.7 [52, 60]	6.3	27.1	1.1
	dsgs	4	0	**59.9**	10		4	9.9	4.9	7.5	59.9 [56, 68]	11	21.9	1.5
	effr	2.2	14.1	8.1	**59.5**		8.1			8	59.5 [56, 64]	7.4	27.5	1.2
	fear		9.9	4	0.9	**57.6**	11.4	7.1	4	5.1	57.6 [52, 60]	8.9	23.5	1.3
	joy	12.5	14	14.6	12.5	16.6	**8.7**	12.3	2.5	6.2	8.7 [6.2, 10.4]	5.9	2.2	1.2
	pain		5.3	15.7	8	21.1	6.4	**33**	5.3	5.3	33 [31.6, 34.2]	5.7	16.5	1.4
	plsr			18	12.3		4	0.5	**50.8**	14.4	50.8 [44, 56]	5	26.3	0.9
	sdns	12.5		7.8	4.2		0.3	4.8	22.2	**48.3**	48.3 [45.8, 50]	6.2	21.4	0.9
Total											45.6 [44.5, 46.8]	6.8	24.2	1.8

All rows in the confusion matrix sum to 100%, minus rounding error. Measures of classification accuracy are the same as in Table 3. Amsm = amusement, angr = anger, dsgs = disgust, effr = effort, plsr = pleasure, sdns = sadness

Acoustic models versus human raters The overall accuracy and confusion matrices of the acoustic model (Table 4) are similar to those based on individual human ratings (Table 3), and the correlation between the two confusion matrices is high (r = .86). Fear, pleasure, and effort were detected by both human raters and acoustic models more reliably when the caller was a woman or child. But there were also differences between the classifications by acoustic models and human raters. The recognition accuracy of *amusement*, *disgust*, and *sadness* by acoustic models was 25%–40% higher when the caller was an adult male; no such gender effect was evident with human raters. Listeners were particularly adept at detecting *sadness* and (sexual) *pleasure*, whereas acoustic models proved superior at distinguishing between *anger* and *fear*.

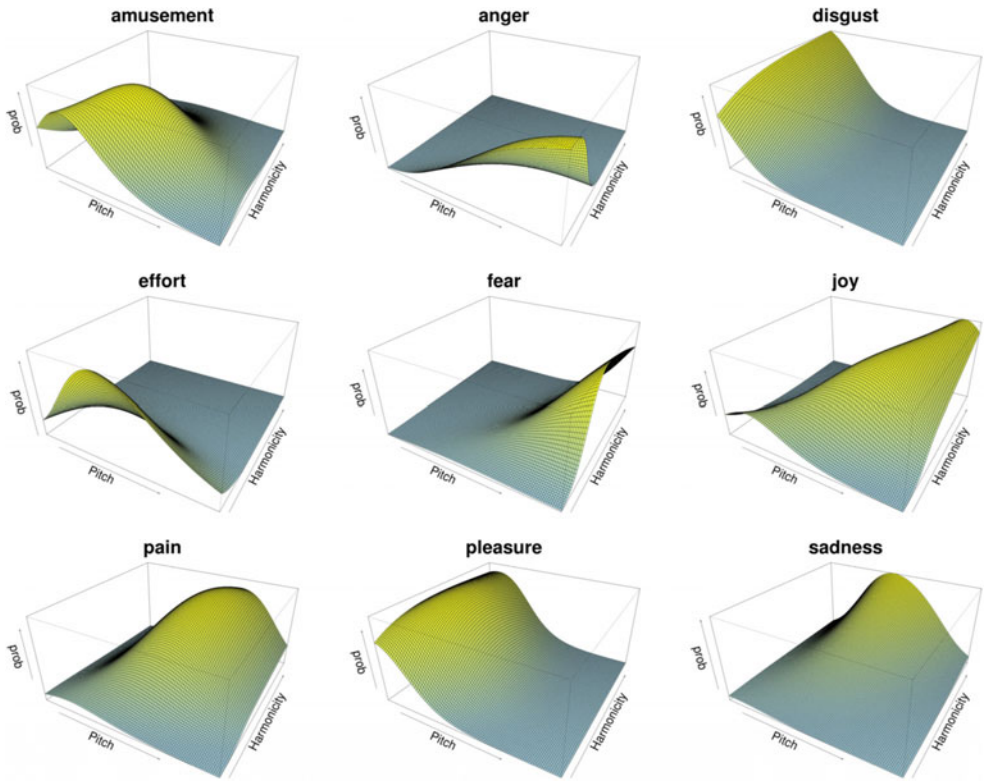

Fig. 3 Perspective plots showing the predicted probability that a sound of particular pitch and harmonicity belongs to each of the nine emotional categories. The probabilities were calculated with multinomial logistic regression. For each point on the pitch–harmonicity plane, the nine probabilities sum to 1. Pitch was log-transformed for this model

Another difference is that human listeners demonstrated clear response biases: *Fear* was overdetected, whereas *disgust* and *sadness* were detected conservatively, with few false alarms. In contrast, the acoustic models had more uniform false alarm rates for all emotions.

The best acoustic predictors of human ratings (perceived emotion) were similar to the predictors of the context-based, "true" emotion. The first and most important predictor of the human classifications of a sound was its mean pitch—having just this one variable, one could still predict ~30% of human decisions. The mean interburst interval and HNR followed pitch as the second and third most important predictors of the emotion perceived by human listeners. An RF model with the same six acoustic variables as above was able to predict the perceived emotion (i.e., the one with the highest score in the rating task) with overall accuracy ~53%, with the following hit rates per emotion: *amusement* 76%, *fear* 67%, *sadness* 62%, *pain* 61%, *anger* 56%, *effort* 45%, *disgust* 36%, *joy* 35%, and *pleasure* 24%.

On the basis of the available acoustic measurements, it is thus easier to predict which sounds will be classified by human listeners as *amusement*, *fear*, *sadness*, *pain*, or *anger*, and less straightforward to predict which sounds will be classified as *disgust*, *pleasure*, or *joy*. Especially for *pleasure*, it seems that its detection by human raters is unexpectedly accurate and based on something that is not captured by the acoustic analysis.

Discussion

This study introduces a corpus of 260 human nonlinguistic emotional vocalizations obtained from online videos. This seems to be the first time a sizable corpus of authentic nonlinguistic vocalizations representing a wide variety of emotions has been compiled and analyzed. An innovative feature of this corpus is that it was built to maximize diversity: Different contexts and types of vocalizations were selected, mostly a single sound per caller. The purpose was to create a database of ecologically valid emotional vocalizations and to train robust acoustic models for their recognition.

One research objective was thus to estimate the extent to which noisy real-life recordings are useful for emotion research and acoustic modeling. Despite the variable quality of the original clips and the presence of background noise, our acoustic models proved capable of classifying the emotions with accuracies similar to those achieved by human raters. It is worth reiterating, however, that all files were manually prefiltered and that certain acoustic measurements were checked by the experimenters: To go from this to automatic classification in real time would require considerably more technical sophistication.

The second objective was to look for the emotions associated with universally recognized nonlinguistic vocalizations. Some emotions were recognized well (*amusement*, *sadness*, *pleasure*, *fear*, and *effort*: 70% or better), others less well but still better than chance (*anger*, *disgust*, *pain*: 40%–50%), and *joy* very poorly, at close to chance levels. Interestingly, the average accuracy increased by 13% if all individual ratings per sound were pooled (popular vote), indicating that group decisions are considerably more accurate than classifications by individual raters. Detailed comparisons of these hit rates with previous research may not be meaningful, since the sounds and emotional categories vary across studies. Furthermore, with playacted sounds, the purpose of rating experiments is to validate the corpus—that is, to show that the intended emotion can be reliably recognized by participants. With naturalistic observations, on the contrary, "validation" would be a misnomer, since the informational content of the sound, rather than its validity, is being investigated.

A serious methodological problem with observational materials is that the "true" emotion of the caller has to be determined by the researcher. As a result, the confusion matrices reported in this study should be treated with some caution. The labeling survey demonstrated that several of the contexts chosen as examples of a particular emotion may have been associated with mixed emotions. For instance, giving birth was judged to evoke a mixture of *pain* and *effort*, and indeed, the raters quite often described the sounds made by women in labor as both *pain* and *effort*. Similarly, sounds produced in sexual contexts may well represent a mixture of *pleasure* and *effort*. Purely semantic ambiguities, especially between the corresponding words for *amusement* and *joy* in the three languages in which participants were tested, also produced classification decisions that were formally errors, but that are not likely to represent a failure to correctly apprehend the emotional state of the caller. It is therefore likely that the informational content of vocalizations in the corpus is considerably richer than the hit rates suggest.

The rating experiment could still be treated as validation if the purpose were to find recognizable naturalistic exemplars of emotional vocalizations. Many sounds were both unambiguous with respect to the underlying emotion and recognized reliably by the human raters. Our analysis of entropy suggests that the sounds with the highest interrater agreement were classified with nearly perfect accuracy in all emotional categories except *joy*, *pain*, and (in the case of human raters but not acoustic models) *anger*. At least six emotions in the dataset thus have strong perceptual anchors—distinct, prototypical, and well-recognized exemplars. These sounds could be used in psychological research as ecologically valid emotional stimuli.

The alternative way to use the corpus would be to view it as a slice of the gamut of nonlinguistic vocalizations that people

produce in a wide variety of situations, whether or not the underlying emotion is obvious. We tested the entire corpus in a rating task and observed two noteworthy differences as compared to earlier studies of playacted emotional vocalizations: the apparent lack of an in-group advantage, and the emergence of call types as apparently natural categories for describing the data.

The rating experiment included participants from several countries, who took the test in three languages: Swedish, English, and Russian. To test for an in-group advantage, ideally the participants from each culture would rate stimuli from each culture. Unfortunately, it proved impractical to find enough non-Western material, and the present corpus is effectively English. This lack of a balanced design means that the absence of group effects must be interpreted with caution. Nevertheless, if emotional vocalizations rely on culture-specific codes, the participants from Brazil, Russia, and perhaps even Sweden should have performed worse than the participants from the English-speaking world, especially with positively valenced vocalizations (as in Sauter, Eisner, Ekman, & Scott, 2010). A noticeable group effect has previously been reported in studies of the same design that compared such relatively proximal groups as British versus Swedish (Sauter & Scott, 2007) and German versus Romanian (R. Jürgens, Drolet, Pirow, Scheiner, & Fischer, 2013) participants. In this study, however, the recognition accuracies were similar in all linguistic–cultural groups. Apart from the seemingly semantic ambiguity of the terms for *amusement* and *joy*, recognition of the sounds in this corpus thus appears to be universal.

This contrasts with previous findings (Elfenbein & Ambady, 2002; Gendron et al., 2014; Koeda et al., 2013; Sauter, Eisner, Ekman, & Scott, 2010; Sauter & Scott, 2007) and raises the speculative but exciting possibility that authentic nonlinguistic vocalizations are less culture-specific than their playacted counterparts. For example, powerful spontaneous bursts of triumph, such as the sounds made by sport fans in our corpus, might bypass cultural conventions and find expression in roars or screams that are species-typical but normally associated with other emotions (fear, anger), and therefore hard to recognize as positively valenced without access to contextual information. Milder sounds of joy in the corpus, such as the happy exclamations of students who have passed an exam, were in fact recognized better than sounds of wild jubilation, but the milder they were, the harder it was to ascertain their spontaneous character. The distinction between mostly-pull (spontaneous, culture-independent) and mostly-push (voluntary, culture-specific) emotional expressions may be at best partial in real life (Scherer & Bänziger, 2010), and particularly problematic in YouTube videos, in which the caller may or may not be posing for the camera. However, this distinction has a sound theoretical foundation. If two distinct neural pathways are responsible for the production of human vocalizations, and spontaneous nonlinguistic sounds are primarily the output of the limbic pathway (U. Jürgens, 2009), then their basic acoustic patterns could be less dependent on social learning. It would be useful to test naturalistic vocalizations further while including culturally and linguistically more remote groups. Ideally, a truly international corpus of spontaneous vocalizations should be compiled and then tested for recognition in several culturally diverse locations, allowing a more formal test of in-group advantage. We predict that this in-group advantage would be attenuated for spontaneous vocalizations, relative to their playacted counterparts.

Another noteworthy result of analyzing the corpus was that acoustically similar calls were used in association with vastly different emotional states, suggesting that that the repertoire of emotional nonlinguistic vocalizations may consist of a small number of species-specific calls. Their mapping to emotions is not random—hence, the better-than-chance overall recognition accuracies—but it is less perfect than has been suggested by controlled studies, in which the actor consciously chose which sound to produce and the listeners may have relied on a culture-specific code to resolve ambiguities.

A shift of focus from emotion to call types may thus offer new insights into the observed confusion patterns, as well as into gender differences in call production. For instance, in the present study women and children were less successful at communicating anger, and better at communicating fear, than men. Why? As it turns out, there was very little acoustic overlap between sounds of anger and fear in men, because men never screamed in contexts suggestive of anger. On the contrary, women produced many screams of anger (acoustically hard to distinguish from screams of fear or pain), and only a few low-pitched, noisy roar- or growl-like sounds of anger (although women made such sounds in association with physical effort). This tendency for women to scream and for men to roar may be related to general aggression levels and the physical ease of producing such vocalizations, given sex differences in the vocal apparatus. Cultural expectations may play a role, as well, but they are unlikely to be the complete story. In fact, the same tendency for males to roar and for females to scream has been reported in monkeys (Leinonen et al., 1991).

Much work remains to be done to investigate human emotional vocalizations from the perspective of call types. In addition, human vocalizations could be traced down to their evolutionary roots by comparing them with the vocalizations of our nearest primate relatives (Brudzynski, 2014; Ross, Owren, & Zimmermann, 2009; Sauter, Eisner, Ekman, & Scott, 2010). Evidence of phylogenetic parallels strongly supports the claim that certain nonlinguistic vocalizations correspond to innate call types, and such evidence is indeed becoming available for both laughter (Ross et al., 2009) and crying (Newman, 2007). Naturalistic human vocalizations recorded in real-life contexts are particularly valuable for phylogenetic reconstruction, because they are more likely to be produced

unintentionally, bypassing language-like and culture-specific modifications.

Conclusions

1. Emotional sounds obtained from noisy online videos were successfully analyzed acoustically to ensure recognition accuracy by statistical models on a par with that of human raters. The key acoustic predictors of emotion were pitch, harmonicity, and measures of temporal structure.
2. The overall recognition accuracy in a rating task was relatively low for some emotions (joy, pain, and anger) and high for other emotions (amusement, fear, pleasure, and sadness). No effect of linguistic–cultural group on recognition accuracy was discovered.
3. The confusion patterns were compatible with the hypothesis that nonlinguistic emotional vocalizations include a small number of call types, which are easily recognized but not specific to one emotion each.

Author note We thank Susanne Schötz and two anonymous reviewers for useful comments. We are also grateful to the many participants who volunteered their time to rate the sounds.

References

Arriaga, G. (2014). Why the caged mouse sings: Studies of the mouse ultrasonic song system and vocal behavior. In G. Witzany (Ed.), *Biocommunication of animals* (pp. 81–101). Germany: Springer. doi:10.1007/978-94-007-7414-8_6

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *Journal of the Acoustical Society of America, 110,* 1581–1597.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception. *Emotion, 12,* 1161–1179. doi:10.1037/a0025827

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 195–200). Belfast: ISCA.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods, 40,* 531–539. doi:10.3758/BRM.40.2.531

Bohn, K. M., Schmidt-French, B., Ma, S. T., & Pollak, G. D. (2008). Syllable acoustics, temporal patterns, and call composition vary with behavioral context in Mexican free-tailed bats. *Journal of the Acoustical Society of America, 124,* 1838–1848.

Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots, 12,* 83–104.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Brudzynski, S. (2014). Social origin of vocal communication in rodents. In G. Witzany (Ed.), *Biocommunication of animals* (pp. 63–80). Berlin: Springer.

Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior, 35,* 327–335.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8,* 135–148. doi:10.1163/156770908X289242

Crockford, C., Herbinger, I., Vigilant, L., & Boesch, C. (2004). Wild chimpanzees produce group-specific calls: A case for vocal learning? *Ethology, 110,* 221–243.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40,* 33–60.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6,* 169–200. doi:10.1080/02699939208411068

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1,* 49–98.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology, 58,* 342–353.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203–235. doi:10.1037/0033-2909.128.2.203

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25,* 911–920.

Hage, S. R., Gavrilov, N., & Nieder, A. (2013). Cognitive control of distinct vocalizations in rhesus monkeys. *Journal of Cognitive Neuroscience, 25,* 1692–1701. doi:10.1162/jocn_a_00428

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of non-linguistic affect vocalizations. *Emotion, 9,* 293–305.

Jürgens, U. (2009). The neural control of vocalization in mammals: A review. *Journal of Voice, 23,* 1–10.

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology, 4,* 111. doi:10.3389/fpsyg.2013.00111

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Frontiers in Psychology, 4,* 105. doi:10.3389/fpsyg.2013.00105

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., . . . Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, 353. doi:10.3389/fpsyg.2013.00353

Lavan, N., Scott, S. K., & McGettigan, C. (2015). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, 1–17. doi:10.1007/s10919-015-0222-8

Leinonen, L., Linnankoski, I., Laakso, M. L., & Aulanko, R. (1991). Vocal communication between species: Man and macaque. *Language and Communication, 11,* 241–262.

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods, 45,* 1234–1245. doi:10.3758/s13428-013-0324-3

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology, 19,* 1994–1997.

Neiberg, D., Laukka, P., & Elfenbein, H. A. (2011). Intra-, inter-, and cross-cultural classification of vocal affect. In *Proceedings of Interspeech 2011* (pp. 1581–1584). Florence: ISCA.

Newman, J. D. (2007). Neural circuits underlying crying and cry responding in mammals. *Behavioural Brain Research, 182,* 155–165.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology, 73,* 530–544.

Parsons, C., Young, K., Stein, A., Craske, M., & Kringelbach, M. L. (2014). Introducing the Oxford Vocal (OxVoc) Sounds database: A validated set of non-acted affective sounds from human infants, adults, and domestic animals. *Frontiers in Psychology, 5,* 562. doi:10.3389/fpsyg.2014.00562

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37,* 417–435.

Petkov, C. I., & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience, 4,* 12. doi:10.3389/fnevo.2012.00012

R Development Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from www.R-project.org

Ross, M., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology, 19,* 1106–1111. doi:10.1016/j.cub.2009.05.028

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31,* 192–199.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63,* 2251–2272.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107,* 2408–2412.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2006). Vocal expression of emotions in normally hearing and hearing-impaired infants. *Journal of Voice, 20,* 585–604.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language, 27,* 40–58.

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *A blueprint for affective computing: A sourcebook and manual* (pp. 166–176). Oxford: Oxford University Press.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32,* 76–92.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40,* 99–116.

Schusterman, R. J. (2008). Vocal learning in mammals with special emphasis on pinnipeds. In D. Oller & U. Gribel (Eds.), *The evolution of communicative flexibility: Complexity, creativity, and adaptability in human and animal communication* (pp. 41–70). Cambridge: MIT Press.

Scott, S., Sauter, D., & McGettigan, C. (2009). Brain mechanisms for processing perceived emotional vocalizations in humans. In S. M. Brudzynski (Ed.), *Handbook of mammalian vocalization: An integrative neuroscience approach* (pp. 187–197). San Diego: Academic Press.

Searcy, W. A., & Nowicki, S. (2005). *The evolution of animal communication: Reliability and deception in signaling systems*. Princeton: Princeton University Press.

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion, 9,* 838–846. doi:10.1037/a0017810

Simonyan, K., & Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist, 17,* 197–208.

Stan Development Team. (2014). *Stan: A C++ library for probability and sampling, Version 2.5.0*. Retrieved from mc-stan.org.

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing vocal repertoires—Hard vs. soft classification approaches. *PLoS ONE, 10*, e125785. doi:10.1371/journal.pone.0125785

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17,* 3–28.