Binter, J. et al. (2023). "Ouch!" or "Aah!": Are Vocalizations of 'Laugh', 'Neutral', 'Fear', 'Pain' or 'Pleasure' Reliably Rated? Human Ethology, 38, 17–47. https://doi.org/10.22330/he/38/017-047

Human Ethology 38(2023): 017–047
Research Article

"Ouch!" or "Aah!": Are Vocalizations of 'Laugh', 'Neutral', 'Fear', 'Pain' or 'Pleasure' Reliably Rated?

Jakub Binter[1], Silvia Boschetti[1,2], Tomáš Hladký[1], Hermann Prossinger[3]
[1]Faculty of Humanities, Charles University, Prague, Czech Republic
[2]Faculty of Science, Charles University, Prague, Czech Republic
[3]Department of Evolutionary Biology, University of Vienna, Vienna, Austria

jakub.binter@fhs.cuni.cz

ABSTRACT

Our research consisted of two studies focusing on the probability of humans being able to perceive the difference between the valence of human vocalizations of high (pain, pleasure, and fear) and low (laugh and neutral speech) intensity. The first study was conducted online and used a large sample ($n = 902$) of respondents. The second study was conducted in a laboratory setting and involved a stress induction procedure (target group: $n = 15$; control group: $n = 13$). For both, the task was to categorize whether the human vocalization of affects was rated positive, neutral, or negative. Stimuli were audio records extracted from freely downloadable online videos and can be considered semi-naturalistic. Each rating participant (rater) was presented with five audio records (stimuli) of five females and of five males. All raters were presented with the stimuli twice (so as to statistically estimate the consistency of the ratings). We could test for consistencies and due-to-chance probabilities using a Bayesian statistical approach. The outcomes support the prediction that the results (ratings) are repeatable (not due to chance) but incorrectly attributed, decreasing the communication value of the expressions of fear, pain, and pleasure. Stress induction (in study two conducted on 28 participants) did have an impact on the ratings of male neutral and laugh – it caused a decrease in correct attribution.

Keywords: *vocalization, emotion, pain and pleasure, Dirichlet distribution, Bayesian statistical approach, Cold Pressor Task*

---

submitted: 07. Jan. 2023; revised: 07. June 2023; accepted: 30. June 2023

INTRODUCTION

Previous studies that deal with acoustic channels

Complex language-based communication is a quite remarkable human characteristic. Thus, paraverbal communication is highly important in social species, such as humans. Much can be deduced from the vocalization: biological sex (Puts et al., 2012), attractiveness (Feinberg et al., 2005), and body size (Pisanski & Reby, 2021).

Furthermore, prosody plays a vital role in the communication of sexual interest (Hughes & Puts, 2021), of dominance towards listeners (Leongómez et al., 2021), and — importantly — also of affective states (Pisanski et al., 2018).

The communication of affective and emotional states is fundamental for our everyday lives. The interaction among and between humans mainly involves the visual and acoustic channels (Kibrik & Molchanova, 2013). The *concerto* of the information occupying the senses is what creates the final assessment of the communicator's state, being based on context, (linguistic) content, postures, facial expressions, and prosody (Leongómez et al., 2022). If the linguistic content is not taken into account, the communicative value of one single signal (i.e. facial expression or vocalization alone) is difficult to extract from the overall multimodal perception.

When only emotional vocalization was rated, the specificity and universality of the vocal production were supported in the case of a negative emotion (Sauter et al., 2010; Gendron et al., 2014) by a cross-cultural study of emotional prosody in both speech (Pell et al., 2009) and emotional vocalization (Gendron et al., 2014). In other studies, the specific role of emotional categories had important implications, with these belonging to the "basic emotions" (anger, fear, disgust, happiness, sadness, and surprise) — presumably because these were more easily recognized in cross-cultural contexts (Bryant & Barrett, 2008; Sauter et al., 2010).

Indeed, publications that focused on the communicative value of one single component of the complex expression process showed outcomes that were not in accordance with both major theories of emotions — one conceptualizing emotion as discrete-categorical (Izard, 1994; Ekman & Cordaro, 2011) and one which focused on emotion dimensions (i.e. valence and arousal, Russell, 1980; Posner at al., 2005). Both theories predicted that emotions that are categorically different (discrete category theory) or on opposite sides in terms of valence (emotion dimension theory), are not misunderstood because they are related to very specific and distinct psychophysiological activation.

Based on these theories, we could expect that, even when taken out of context, emotional expressions would be distinguishable from one another; but data-driven research provides us with counterintuitive results.

Recent results on emotional vocalizations have discovered a novel phenomenon — the misattribution of very intense emotions, called "emotion intensity paradox" (a name given by Holz et al., 2021, even though earlier publications on the topic exist): when the intensity of the emotion is very high, it is more difficult to extract the valence (positive or negative) (Atias et al., 2019).

This misattribution not only occurs for acoustic stimuli; it was previously found in facial expressions of emotion as well (Aviezer et al., 2012; Hughes & Nicholson, 2008; Wenzler et al., 2016; Boschetti et al., 2022). Facial expressions of emotions of high arousal are not only very difficult to rate correctly, but are oftentimes rated due to chance ('guessing') and inconsistently (Boschetti et al., 2022). The due to chance probability is rarely studied but is a very important metric that allows, based on the distribution of the ratings, for interpretation related to repeatability of the outcome. It is calculated as the probability by integrating the likelihood function of the Beta distributions over the integral from 0 to $\frac{1}{2}$ or from $\frac{1}{2}$ to 1, depending on which side the mode is; these areas are the probability of the observed distribution of the ratings being due to chance.

A further metric related to reproducibility is consistency. The stimuli are presented twice in randomized order; the probability of the same rating being repeated (irrespective of the correctness of the rating) is reported to evaluate how consistent the rated phenomenon is for the raters. It should be pointed out that the two are not dependent on each other, and are not necessarily related to correctness. The ratings may be incorrect, consistent, and due to chance or any other combination of the three.

We note that guessing can be inconsistent; therefore, any study dealing with rating issues must test for both guessing and inconsistency. Artificial intelligence (AI) analysis conducted on the facial stimuli identified a further interesting phenomenon. While humans are unable to rate correctly due to their inability to extrapolate sufficient cues present in the facial expression, AI can correctly categorize the facial expressions with high accuracy (Binter et al., 2021; Prossinger et al., 2022). The rating inconsistencies by humans is due to their inability to grasp such subtle cues; consequently, the facial expressions of intense emotions are guessed.

This destroys the foundations of the communicative value of the extreme affective state as previously discussed by Aviezer et al. (2012). Since the intensity of the experience (by the expresser) is very high and the situation that evokes it is rich in contextual information regarding the valence, the aim of the behavior may be mainly to capture and orient the attention of those observing (i.e. the receivers). In particular, acoustic stimuli function primarily to gain attention; this has been previously found when comparing screams (intense emotional vocalizations) with regular speech with regard to accuracy and rapidity of localization (Arnal et al., 2015). The results showed that screams were more rapidly and better localized and no differences between natural and synthetic screams (the latter constructed by adding roughness to neutral vocalizations using dedicated software) were found. A recent publication extends this finding and suggests that the ratings of the perceived affects are shifted towards the negative end of the valence scale (Anikin et al., 2020).

Another variable that can have an impact on the success of correct rating is the sex of the rater as well as the sex of the expresser. Previous studies showed that women are better at correct attribution, especially in case of negative emotion evaluation (Thompson & Voyer, 2014). Belin and colleagues (2008) found that male participants rated the expressions as more intense. In that same study, the sex of the expresser was also found to have a significant impact on the ratings of valence and arousal — with a greater arousal but smaller valence attributed to the vocalizations produced by women. The (statistical) association between the sex of the rater and the sex of the expresser was not significant and two groups of authors recommended further studies to clarify this point (Belin et al., 2008; Thompson & Voyer, 2014). In a third study (Vasconcelos et al., 2017), the effect of the sex of the rater was specific for the emotional category (i.e. better recognition of anger and sadness by females in contrast to surprise by males); thus, this (third) study confirms the better recognition of negative expression of vocalizations by women.

There are limitations in the experimental design of these aforementioned studies.

First, the vocalizations were staged (i.e. often performed by 'neutral' actors who do not elicit the acted emotion, because they did so in a laboratory or in a recording studio; Belin et al., 2008; Vasconcelos et al., 2017). The role of ecological validity clearly emerged as limiting in the case of real versus artificial laugh. The two are not only perceived differently but also cause the activation of different brain regions (McGettigan et al., 2015; Kamiloğlu et al., 2022).

Second, the sex of the raters and the expressers is not always taken into consideration. Previous studies (Thompson & Voyer, 2014; Belin et al., 2008; Vasconcelos et al., 2017) showed that this characteristic may systematically affect the correctness of the ratings; it is important to explore the effect of this characteristic — especially the interaction between the sex of the raters and sex of the expressers. Extending on this point, the expressivity of the specific expresser should also be taken into account: some individuals may be more expressive

or more stereotypic (or both) in their emotional expression and, therefore, the stimuli produced by such an individual may be easier to correctly identify.

Third, a further limitation of the previous studies (Thompson & Voyer, 2014), which also contributes to the differing and oftentimes contradicting findings, is the presenting of pre-identified emotion categories and/or the use of a rating scale (Bryant, 2021). The pre-identification of categories (i.e. requesting the respondent to choose whether the emotion displayed is anger, fear or surprise) is more likely to capture the complex psychological representation of the emotion and increase the variance, because it could then be more affected by culture and linguistic differences — as criticized by Boschetti at al. (2022). The focus of previous studies was often on emotions and affects as categories (without attention for the how intense these emotions are) or on the dimensions of the emotions (high vs. low arousal or positive vs. negative), without categorizing the emotions. Consequently, outcomes of a study avoiding these limitations (such as the present one) are very difficult to compare with previous research that focused on the universality of specific categories (such as basic emotions) but not on others (the secondary emotions — the affects).

Fourth, a limitation that should be mentioned is the rater's state (Thompson & Voyer, 2014; Belin et al., 2008; Vasconcelos et al., 2017). In the real-world scenario, it is almost impossible to be calm and remain in a neutral state while being exposed to a high-affect situation that includes the presentation of extreme vocalizations of pain, pleasure or fear. It was previously shown that the vocalization itself changes (Tolkmitt et al., 1986; Sherer, 2003; Cowie & Cornelius, 2003). Others are also able to differentiate whenever the sender is stressed (Kreiman & Sidtis, 2011; Piskanski et al., 2019) as well as when it can be detected by computer algorithms (Han et al., 2018; Praseito et al., 2019).

There are researches that focus on affect rating while under conditions of stress. In all of the previous cases, it was facial expression that was rated. One study has identified the shift towards the negative valence of surprised faces (Brown et al., 2017); another study has found this shift in unambiguous faces (neutral and smile) but not in ambiguous faces (pain and pleasure expression) while the physiological changes were found only in response to fear expressions by the same researchers (Boschetti at al., 2022, Binter et al., 2022).

Studies presented in this paper

Our two studies are designed to overcome some of the aforementioned methodological limitations.

Aim of Study I: to quantify the consistency of the ratings of such vocalizations, focusing on the (biological) sex of the rater as well as the (biological) sexes of the rated (vocal expressers). The stimuli used were audio records of five emotional vocalizations, with high (pain, pleasure and fear) and low (neutral and laugh) intensity. As in previous studies, the stimuli were produced by five *different* male and five *different* female expressers (e.g., Belin et al., 2008). Each of the five emotions was expressed by the same five male and *the same* five female expressers (so there are 25 stimuli for each sex and a total of 50 stimuli), and each rater rated each of these stimuli twice, presented in random order. We thereby control for the sex and the expressivity of the rated individual (the stimulus expresser).

We generate our set of stimuli by using audio records from videos that depict consensual acts of extreme sexual activities. We adopted a categorical methodology in which there are three ratings: positive, negative, or neutral. We predicted that, if the individual has been exposed to a negative stimulus (pain, for instance), the vocalization on the audio record (with the expected high intensity), will be rated as negative. In a manifestly opposite stimulus (pleasure, for instance), the rating should be positive. We expect female raters to perform better than male raters in recognizing negative emotions.

Aim of Study II: to evaluate the impact of the physiological state of the rater on his/her ratings. The procedure followed that of Study I with the addition of a stress inducing procedure (Cold Pressure Task; described in more detail below).

METHODS

Sample

Expressers' vocalizations: In order to be consistent with the published terminology, we use the terms "expressers" and "vocalizations" to describe what was presented in the 50 audio records as stimuli. We specify the biological sex of the raters with the terms male and female. The biological sex of the expressers is documented in the audio records, so we (the authors of this paper) could rely on this information.

Raters: In order maintain consistency with the published terminology, we use the terms "expression raters" and "respondents" to describe the individuals who were presented with the stimuli and who provided their ratings. We specify the biological sex of the expresser with the terms male and female. The biological sex we list is the respondent's self-reported one. We deleted all ratings ($n = 4$) of respondents who did not report their biological sex.

In Study I: A total of 902 individuals (aged 18–50; M_{age} = 32 years, SD = 8.9 years) completed the questionnaires; 526 women (M_{age} = 30.9 years, SD = 8.3 years) and 376 men (M_{age} = 33.6 years, SD = 9.5 years). The data were collected in the Czech Republic in 2021 via the agency Czech National Panel (narodnipanel.cz) and a science-oriented online portal (pokusnikralici.cz) using the online platform for data collection Qualtrics®. Participants submitted responses either via the computer keyboard or touchscreens of mobile devices (smartphones or tablets).

Study II: A total of 28 individuals (aged 19–30; M_{age} = 22.3 years, SD = 2.3 years) participated in Study II; 13 women (M_{age} = 22.7 years, SD = 2.8 years) and 15 men (M_{age} = 21.9 years, SD = 1.8 years). The data were collected in a laboratory in Prague, Czech Republic. 28 participants were presented with the same stimuli as in Study I; the lower right legs of target group members ($n = 15$; 7 women and 8 men) were immersed in cold water (2–4 °C) for 1½ minutes, which subsequently increased their stress level (Cold Pressor Task, CPT; Bullinger et al., 1984; Brown et al., 2017). The control group's 13 (6 women and 7 men) participants' lower right legs were immersed in water at room temperature.

Criteria for inclusion were: (a) age of respondents between 18 and 50 years, and (b) at least a minimal experience with adult media, since the vocalizations used in this study were extracted from such materials.

Stimuli generation

From the numerous audio-visual materials viewed, ten audio records (five with female vocalizations and five with male vocalizations) were chosen. Based on the plot in each of the audio-visual materials, five vocalizations were selected (one of neutrality, one of fear, one of pleasure, one of pain, and one with laugh). Three of the authors (S.B., J.B., and T.H.) are researchers in the field of human sexuality with more than 10 years of experience, specifically focusing on extreme sexual behavior and on the consumption of erotic materials. All three authors (one female and two male) provided their opinion on all of the chosen stimuli. Based on the contextual information, all agreed on the stimuli chosen and what expression is to be expected. Prior to the agreement, stimuli choices were debated among all three researchers in dedicated meetings.

We point out that a common misconception is that the individuals taking part in such exchanges derive sexual pleasures from pain and the two (pain and pleasure) happen simultaneously. Although this may be possible, we have found no mention of it in the published scientific literature. Rather, it should be noted that sensitivity is increased by the experience of pain by various parts of the body (in the case of our stimuli mainly slapping the buttocks and thighs) and only after the painful procedure is over is climax achieved. All the audio stimuli we chose were derived from the whole context of the video. Specifically, we could rely on the images/scenes to identify the emotions and affects (which the raters could not, as they only heard the vocalizations). There is no doubt, due to the camera perspective, about the occurrence of the climax in male expressers. In the female expressers, no such explicit method of judgment can be used, but all signals of the occurrence of climax were identified by the researchers (involving breathing, contraction of pelvic musculature, twitching of anal sphincter muscles, facial blushing, vocalization, etc.; Dubray et al., 2017), and further supported by expressers' self-reports after the videos had ended.

In each audio record, male/female vocalizations expressed fear, pain, and pleasure during the session, while laugh and neutral vocalization (speech) were recorded during an interview prior to the pain and pleasure experiences. All stimuli (audio records) were adjusted to the same sound level and lasted from 0.5 seconds to 1.5 seconds — depending on the stimulus.

Procedures

In Study I, the set of stimuli was presented twice (Task 1 and Task 2), each time with a different randomization sequence: each stimulus was played for approximately 1.5 seconds at random intervals ranging from 1 to 3 seconds (so as to avoid constant/rhythmic preparedness for the stimulus presentation). Thus, a total of 100 ratings (two for each of the 50 different stimuli) were collected for each rater.

In Study II, each rater was presented with the set of stimuli (25 male and 25 female vocalizations from five male and five female expressers) only once. The reason is that the Cold Pressor Task (CPT) has a limited impact on the cortisol release and this allowed us to finish the procedure within 20 minutes after the effect of the CPT had ended.

Both studies were parts of a larger project during which the participants rated facial expressions, vocalizations (this manuscript), followed by congruent and incongruent presentations of both modalities. The presentation was as follows: Task 1 visual, Task 1 auditory, Task 2 visual, Task 2 auditory, congruent and incongruent stimuli presentation. All presentations were randomized separately.

For the Study 2 a sample independent of Study 1 was collected.

Ratings

Previous literature (Dolan et al., 2001, Bryant, 2021) has noted that it is a challenging task to correctly identify human vocalizations (e.g., to categorize the expression of fear as indeed fear). We, therefore, asked our participants to rate the vocal expression as either positive, neutral, or negative. The correct rating for laugh and pleasure is positive, for fear and pain it is negative, and for neutral it is neutral. We thereby avoided the problem of correct labeling and avoided any intricacies associated with a verbal categorization system. The ratings were communicated either via using keyboard keys or a touchpad with dedicated areas (specified by icons); they were subsequently stored in a dedicated database.

Statistical Analyses

(a) Frequentist statistics deals with probabilities as numbers (for example $p_{Heads} = \frac{1}{2}$ for a fair coin), and point estimates of estimators, (such as the average as an estimate of the expectation of a distribution) as well as the estimators of parameters (such as $slope = 0.107\ldots$ for a linear OLSq regression). Bayesian statistics, on the other hand, deals with likelihood distributions which are updated with the collection of evidence ($\mathcal{L}_{posterior} = evidence \cdot \mathcal{L}_{prior}$, where $\mathcal{L}$ is the likelihood — usually the probability density function of a distribution; as in Boschetti et al., 2022). Furthermore, in Bayesian statistics, probability (usually abbreviated s) is a random variable $0 \leq s \leq 1$; research results include the most likely probability — where the maximum of the likelihood function $\mathcal{L}$(s) occurs; see Fig. App-1 in Appendix I). By this construction (a consequence of Bayes Theorem) there are no restrictions on sample sizes and numbers of samples — in contrast to Frequentist statistics (details in Appendix I).

(b) Confusion matrices: Both female and male ratings are Dirichlet-distributed (in our case: 3-parametric). The (Bayesian) method of determining whether two groups are significantly different (or not) is to calculate the confusion matrix; it is the obligatory method to use when sample sizes are small. One sample (F) has a distribution $dist_F$ and another sample (G) has a distribution $dist_G$. When there is an overlap of the *pdf*s (probability density functions) of these two distributions, a fraction of *F* is TRUE (and a fraction is FALSE); likewise, for *G*. The confusion matrix has four entries:

$$\begin{pmatrix} TRUE_F & FALSE_F \\ FALSE_G & TRUE_G \end{pmatrix}$$

If the off-diagonal elements ({FALSE_F, FALSE_G}) are small, there exists a significant difference between the distributions of *F* and of *G* (the significance level being chosen by the researcher). Observe that the sum of each row in the confusion matrix is $1 = 100\,\%$. The fractions in the confusion matrix can also be calculated using Monte Carlo methods.

(c) Possibility of effects being due to chance: Because, in Bayesian statistics, the probability is a random variable ($0 \leq s \leq 1$), the crucial separator for determining chance is $s = \frac{1}{2}$. The probability is either the integral of the likelihood function $\mathcal{L}$(s) over the interval $0 \leq s \leq \frac{1}{2}$ or the integral over the interval $\frac{1}{2} \leq s \leq 0$, depending on which side of $s = \frac{1}{2}$ the mode is. In either case, the integral determines whether an observation is due to chance. (A graphical description is shown in Appendix I.) We note that the probability due to chance is never greater than 50 %. Since there are positive, neutral, and negative responses, we generate a binary case (the correct responses versus the incorrect responses); then the likelihood function is the probability density function of a Beta distribution (Appendix II). For example, for laugh, the correct response is a positive rating while the neutral rating and the negative rating together are incorrect responses.

RESULTS

Probabilities of Correct Ratings

Of the five affective states displayed in vocalizations by each sex, only two were rated with high accuracy (above 85% of correct responses): laugh and neutral (Table 1a and 1b). The laugh vocalizations were correctly assessed by female raters with a 0.932 probability in the case of female expressers and a probability of 0.966 in the case of male expressers. For the male raters, we observed lower probabilities of correct ratings with stimuli produced either by male or female expressers (0.872 in the case of male expressers and 0.893 in the case of female

expressers); however, the difference between male and female raters was significant only in the case of male expressers (Table 1a and Figure 1a).

The ratings of neutral vocalizations had very high accuracy (above 90%), independent of the sex of the raters or of the sex of the expressers. Female raters had slightly higher accuracy probability (0.950 for male expressers and 0.926 for female expressers), than the male raters (0.938 for female expressers and 0.912 for male expressers); these differences were not significant (Table 1b and Figure 1b).

For the vocalizations of fear, we observed that the probability of correct ratings by both sexes of raters was very low. Indeed, female raters had only a 0.138 probability of correctly rating fear for male expressers and 0.028 for female expressers. For male raters, we observed a probability of 0.184 when expressed by male and 0.224 when expressed by female expressers. These rating probabilities were not significantly different between male and female raters (Table 1c and Figure 1c).

The vocalizations of pleasure also had low probabilities of correct ratings: for female raters the probability of correct rating of pleasure vocalization by female expressers was 0.447 while for male expressers it was 0.442. For male raters, the probability of correctly rating pleasure vocalization by female expressers was 0.426 and 0.551 by male expressers. The differences between the probabilities of the ratings by male and female raters were not significant (Table 1d and Figure 1d). We note that these four probabilities are close to the boundary $s = \frac{1}{2}$, so it is important to calculate the due-to-chance probability (the indicator of guessing).

In the case of vocalizations of pain, we observe that the Dirichlet distributions of the ratings are not significantly different for male versus female stimuli — for both the male and the female raters. We also note that, for the female raters, the modes indicate that they rated the stimulus pain incorrectly. For the male raters, we observe that there is no mode of the Dirichlet distribution inside the domain, both for male and for female stimuli. The non-existence of a mode necessitates an interpretation of the ratings, guided by the mathematical properties of the Dirichlet distribution. Similar to the graph for the stimulus pleasure (Fig. 1c), it happens that, when the modes have coordinates close to $\frac{1}{2}$ for both correct and incorrect ratings, they (the modes) approach the hypotenuse of the domain. In the case of the pain stimulus, the ML Dirichlet distribution 'pushes' the inferred mode beyond the domain diagonal — the mode therefore no longer exists. Contributing to this non-existence of the mode is the fact that the ratings for correct and incorrect are in the vicinity of $\frac{1}{2}$; because $s_A + s_B + s_C = 1$, the probability of s_B would be forced to be close to zero — if the mode exists inside the domain. In terms of interpreting how this situation can occur, we point out that the (male) raters are not guessing. Consequently, they are often rating incorrectly, but they are convinced they are not incorrect; or — phrased differently — their conviction of a correct rating fluctuates. In the case of the stimulus pleasure, this fluctuation is just small enough to ensure the mode remains defined and stays within the domain, but very close to the hypotenuse.

The vocalizations of pain also have a low probability of correct rating. The ratings are actually so incorrect that the mode's *pdf* is forced beyond the hypotenuse and the results are (numerically) invalid. Therefore, the outcome is very similar to the one of the pleasure vocalization ratings where the distribution is almost equally distributed between the extreme poles (Table 1e). There is no significant difference between the ratings provided by the male and female raters nor the ratings of male and female vocalizers.

Table 1: The modes and the confusion matrices for the male and female voice stimuli rated by female and male raters. Only for male pain stimulus rated by the female raters (**c**) are the modes outside the domain defined by $s_A + s_B + s_C = 1$ for the Dirichlet distribution, hence expressing a mode is *nA* (not applicable; further information in the Appendix II). If the off-diagonal entries are less than 10% (Caelen, 2017), then the ratings are significantly different; those confusion matrices are marked with an asterisk.

(a)

Stimulus	Expresser	Task	Raters	Modes			Confusion Matrix
				positive	neutral	negative	
Laugh	Male	1&2	Male	0.872	0.104	0.024	$\begin{pmatrix} 92.2 & 7.8 \\ 6.2 & 93.8 \end{pmatrix}^*$
	Male	1&2	Female	0.932	0.058	0.010	
	Female	1&2	Male	0.893	0.081	0.026	$\begin{pmatrix} 75.5 & 24.5 \\ 49.2 & 50.8 \end{pmatrix}$
	Female	1&2	Female	0.966	0.025	0.009	

(b)

Stimulus	Expresser	Task	Raters	Modes			Confusion Matrix
				positive	neutral	negative	
Neutral	Male	1&2	Male	0.029	0.938	0.033	$\begin{pmatrix} 71.9 & 28.1 \\ 26.5 & 73.5 \end{pmatrix}$
	Male	1&2	Female	0.031	0.950	0.019	
	Female	1&2	Male	0.061	0.912	0.027	$\begin{pmatrix} 66.5 & 33.5 \\ 30.0 & 70.0 \end{pmatrix}$
	Female	1&2	Female	0.048	0.926	0.026	

(c)

Stimulus	Expresser	Task	Raters	Modes			Confusion Matrix
				positive	neutral	negative	
Fear	Male	1&2	Male	0.653	0.163	0.184	$\begin{pmatrix} 58.0 & 42.0 \\ 28.3 & 71.7 \end{pmatrix}$
	Male	1&2	Female	0.690	0.172	0.138	
	Female	1&2	Male	0.696	0.080	0.224	$\begin{pmatrix} 74.4 & 25.6 \\ 32.9 & 67.1 \end{pmatrix}$
	Female	1&2	Female	0.600	0.073	0.028	

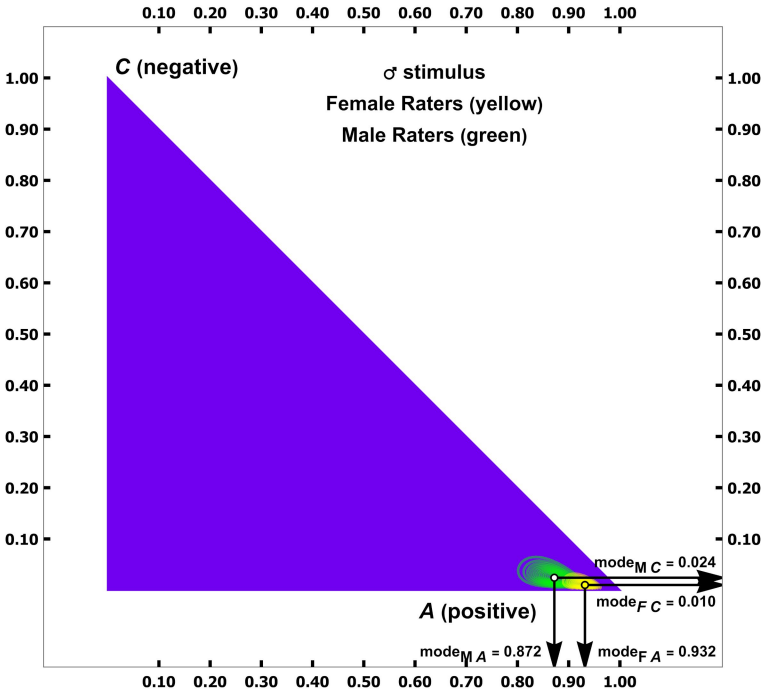
(d)

Stimulus	Expresser	Task	Raters	Modes			Confusion Matrix
				positive	neutral	negative	
Pleasure	Male	1&2	Male	0.404	0.045	0.551	$\begin{pmatrix} 68.4 & 31.6 \\ 57.9 & 42.1 \end{pmatrix}$
	Male	1&2	Female	0.442	0.023	0.535	
	Female	1&2	Male	0.536	0.038	0.426	$\begin{pmatrix} 73.0 & 27.0 \\ 46.3 & 53.7 \end{pmatrix}$
	Female	1&2	Female	0.447	0.005	0.548	

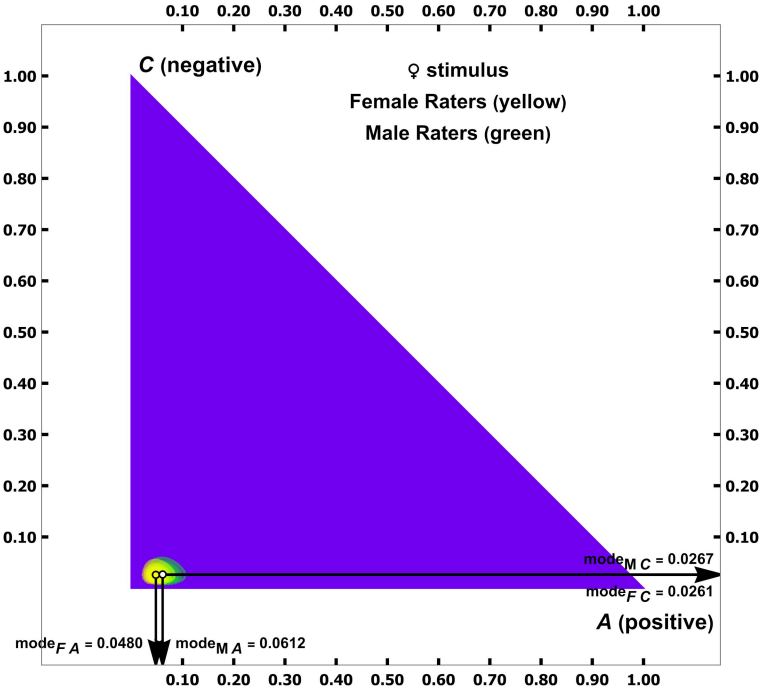
(e)

Stimulus	Expresser	Task	Raters	Modes			Confusion Matrix
				positive	neutral	negative	
Pain	Male	1&2	Male	*nA*	*nA*	*nA*	$\begin{pmatrix} 66.4 & 33.6 \\ 59.2 & 40.8 \end{pmatrix}$
	Male	1&2	Female	*nA*	*nA*	*nA*	
	Female	1&2	Male	0.599	0.099	0.302	$\begin{pmatrix} 80.4 & 19.6 \\ 38.4 & 61.6 \end{pmatrix}$
	Female	1&2	Female	0.521	0.069	0.411	

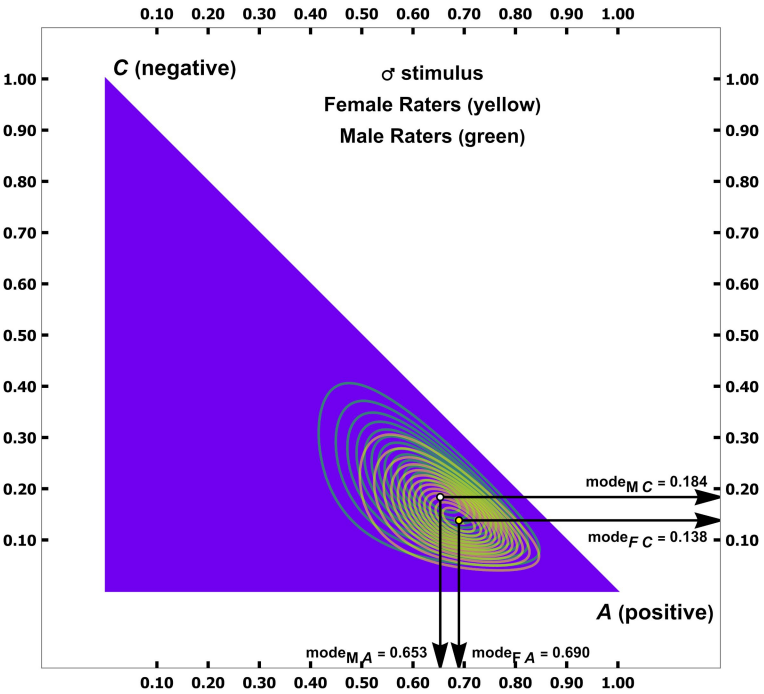
1(a) Laugh



1(b) Neutral



1(c) Fear



1(d) Pleasure

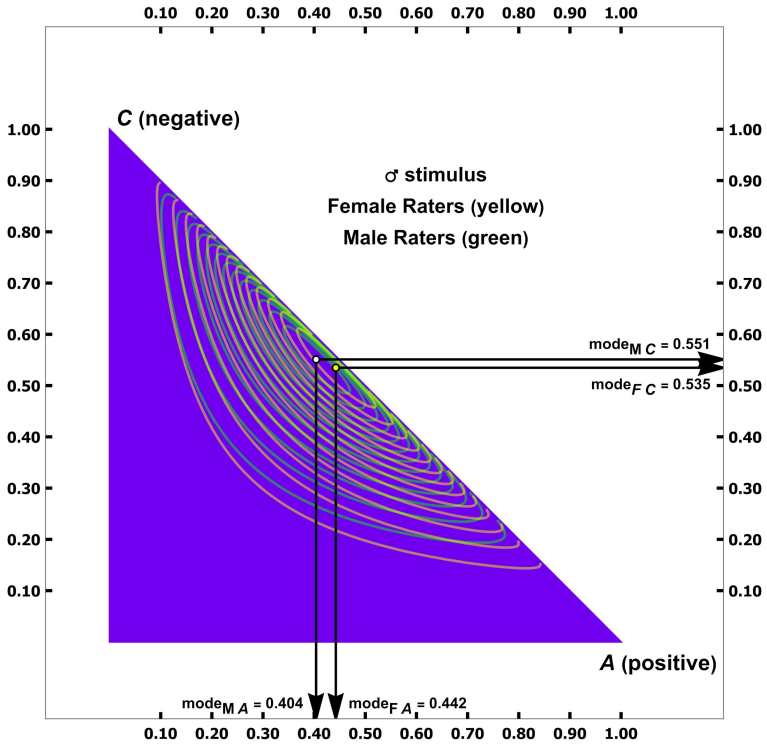


Figure 1: A selection of contour plots of the *pdf*s of the Dirichlet distributions of the ratings of stimuli by both male and female raters. **(a)** Laugh, male stimulus; **(b)** Neutral, female stimulus; **(c)** Fear, male stimulus; **(d)** Pleasure, male stimulus. In all cases, the *pdf*s of the Dirichlet distributions are defined over the domain (rendered as a purple triangle), because $s_A + s_B + s_C = 1$. Contours are rendered in $\frac{1}{14}$ of the maximum likelihood of the *pdf*. We observe that, the farther the modes are from either $s_A = 1$ or $s_C = 1$, the closer the mode is to the hypotenuse of the domain triangle .

Differences in Ratings by male and female raters

The analyses of male and female differences in correct attributions of an individual expresser (or all expressers) displaying one stimulus (e.g. fear) revealed that all results are not significant. In other words, we confirmed the finding published previously that there is no systematic advantage of one sex correctly rating the presented stimulus over the other. Thus, as is deducible from the further breakdown (Figure 2 and Table 2), there is a high degree of similarity between the ratings. Expectedly, neutral and laugh are rated with high assignment accuracy (by both sexes). Interestingly, two male expressers (mA and mB) were also rated with high accuracy by both sexes while the others were rated with an equally low probability of correct attribution.

Table 2: The confusion matrices testing whether the male versus female correct ratings of each expresser and of each affective state were significantly different (Fig. 1). The ratings of all five affective states for each expresser are Dirichlet distributions, each with five concentration parameters; the entries in the confusion matrix are $\begin{pmatrix} TRUE_F & FALSE_F \\ FALSE_G & TRUE_G \end{pmatrix}$.

Expresser	Confusion Matrix	Affective State	Confusion Matrix
fA	$\begin{pmatrix} 49.0 & 50.9 \\ 43.7 & 56.3 \end{pmatrix}$	Laugh	$\begin{pmatrix} 50.6 & 49.4 \\ 45.2 & 54.7 \end{pmatrix}$
fB	$\begin{pmatrix} 47.4 & 52.5 \\ 40.4 & 59.6 \end{pmatrix}$	Fear	$\begin{pmatrix} 53.2 & 46.8 \\ 43.2 & 56.7 \end{pmatrix}$
fC	$\begin{pmatrix} 50.7 & 49.3 \\ 42.6 & 57.4 \end{pmatrix}$	Pain	$\begin{pmatrix} 51.0 & 48.9 \\ 41.6 & 58.4 \end{pmatrix}$
fD	$\begin{pmatrix} 51.5 & 48.5 \\ 48.6 & 51.4 \end{pmatrix}$	Pleasure	$\begin{pmatrix} 57.3 & 42.7 \\ 49.9 & 50.1 \end{pmatrix}$
fE	$\begin{pmatrix} 53.4 & 46.6 \\ 47.1 & 52.9 \end{pmatrix}$	Neutral	$\begin{pmatrix} 44.7 & 55.3 \\ 44.7 & 55.3 \end{pmatrix}$
mA	$\begin{pmatrix} 49.3 & 50.6 \\ 48.1 & 51.8 \end{pmatrix}$		
mB	$\begin{pmatrix} 50.2 & 49.8 \\ 48.3 & 51.7 \end{pmatrix}$		
mC	$\begin{pmatrix} 51.3 & 48.6 \\ 48.1 & 51.9 \end{pmatrix}$		
mD	$\begin{pmatrix} 51.6 & 48.3 \\ 47.2 & 52.7 \end{pmatrix}$		
mE	$\begin{pmatrix} 55.2 & 44.7 \\ 51.0 & 49.0 \end{pmatrix}$		

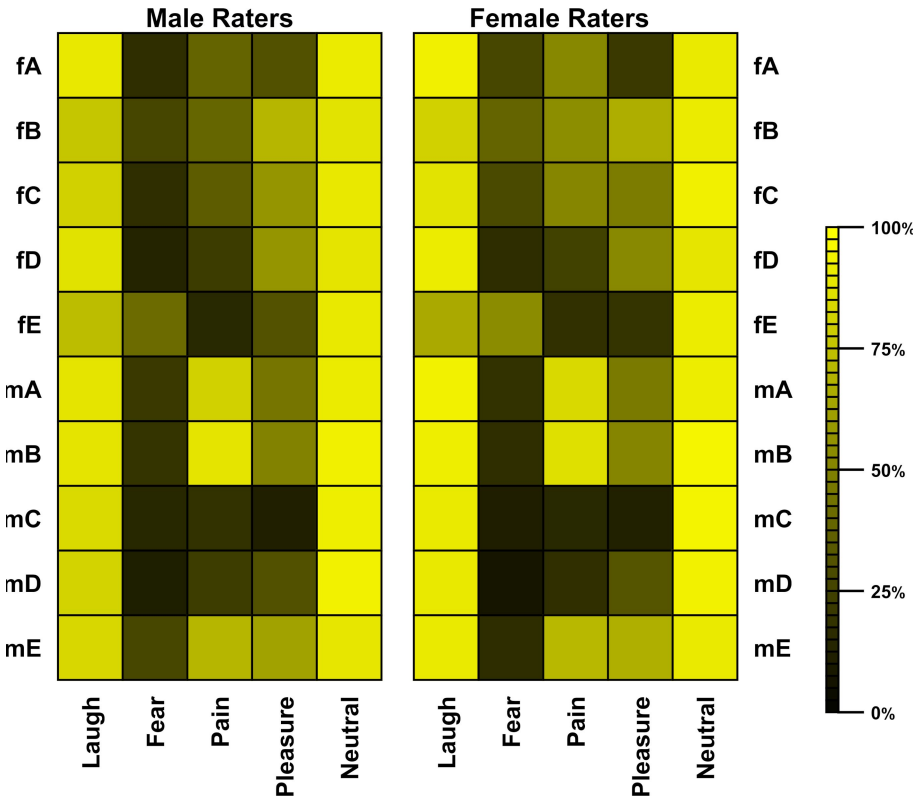


Figure 2: Two heat maps showing the correctness probabilities of ratings by male and female raters of the male vocalizations and the female vocalizations. The male vocalizations and female vocalizations are labeled F_{index} and M_{index} respectively.

Ratings Due to Chance

One of the advantages of the Bayesian statistical approach is the possibility to test whether the result obtained is consistent ('real' in common parlance) or if it has been obtained due to chance. The probability of the result being due to chance ranges between 0 and 50%; the closer to 50%, the more probable that the result is due to chance. The ratings of all of the vocalizations were rated with a chance probability below one percent (not shown). In other words, the rating was not the result of guessing and the result is reproducible. We highlight (again) that this does not mean that the raters are correct or consistent, only that the ratings are not due to guessing and the raters trusted their judgement.

As for patterns that can be deduced and used for further research, it should be pointed out that the female raters were highly consistent when rating neutral vocalizations by men, whereas men were consistent in rating the female fear and male laugh vocalizations. Conversely, an extreme inconsistency was found in the case of men rating male fear, women rating both vocalizations of fear and laugh.

Consistency of ratings

Since we presented all stimuli as two consecutively presented tasks, each in a different randomized order, we have the possibility to test the consistency of the ratings. To do so, we have used a Bayesian probability test; the ratings (correct versus incorrect) are Beta distributed.

The confusion matrices that express how significantly different the ratings of the stimuli were between Task 1 (first rating) and Task 2 (second rating) by the female raters (Table 3a) and the male raters (Table 3b). A significant difference is present if both off-diagonal entries are less than 10 % (Caelen, 2017). For example, for the male raters, the first rating of stimulus f_E for laugh was significantly different from the second rating. On the other hand, the first rating (by male raters) of the stimulus m_A for pleasure was not significantly different for the second rating. For the rating of pleasure by the female raters, their ratings were significantly different for $\frac{8}{10}$ of the stimuli. There is no pattern for significant differences of rating of the stimuli, neither by the female nor by the male raters. Because we have evidenced that the ratings are not due to chance (in other words, the raters are not guessing), the entries in Table 3 show a remarkable result: even if the raters rate the acoustic stimuli wrongly, they are not guessing; that is to say, they are making a different mistake (wrong rating) when rating again in Task 2. This effect is evident in the *pdf*s of the Dirichlet distributions of Task 1 and Task 2. If the modes are far from the correct rating, then one can be close to the maximal incorrect rating, but also only halfway along the incorrect rating, but then — by definition — close to the neutral rating. In such a scenario, the raters are not guessing, but their incorrect ratings are consistently wrong. Consistently wrong does not mean, however, that they gave the same rating for both tasks. This phenomenon seems to be peculiar to acoustic stimuli. In our publication of visual stimuli, we detected that the raters were guessing (Boschetti et al., 2022), and therefore — by definition — were guessing consistently. In either study (visual or acoustic stimuli): only if the raters were guessing during one task and wrongly rating (but not guessing) during the other task, would the off-diagonal elements be very small. In our case of rating acoustic stimuli, we do not observe this phenomenon. To repeat: we observe that the raters make mistakes (albeit not for every stimulus) — but are not guessing. Very often they made a different rating mistake during Task 1 versus Task 2.

Table 3: The confusion matrices (entries in %) expressing how consistent the first versus the second ratings of the stimuli are. Significantly different ones are marked with an asterisk. The symbols fA, mark the female stimulus A, mD the male stimulus D, etc.

(a)

Female Raters					
Expresser	**Laugh**	**Fear**	**Pain**	**Pleasure**	**Neutral**
fA	$\begin{pmatrix}74.0 & 26.0\\24.1 & 75.9\end{pmatrix}$	$\begin{pmatrix}85.8 & 14.2\\15.7 & 84.3\end{pmatrix}$	$\begin{pmatrix}90.9 & 9.1\\10.9 & 89.1\end{pmatrix}$	$\begin{pmatrix}90.4 & 9.6\\10.5 & 89.5\end{pmatrix}$	$\begin{pmatrix}81.0 & 19.0\\21.1 & 78.9\end{pmatrix}$
fB	$\begin{pmatrix}91.1 & 8.9\\9.6 & 90.4\end{pmatrix}^*$	$\begin{pmatrix}91.8 & 8.2\\8.1 & 91.9\end{pmatrix}^*$	$\begin{pmatrix}99.7 & 0.3\\0.3 & 99.7\end{pmatrix}^*$	$\begin{pmatrix}99.5 & 0.5\\0.4 & 99.6\end{pmatrix}^*$	$\begin{pmatrix}52.5 & 47.5\\44.1 & 55.9\end{pmatrix}$
fC	$\begin{pmatrix}95.6 & 4.4\\5.6 & 94.4\end{pmatrix}^*$	$\begin{pmatrix}99.7 & 0.3\\0.2 & 99.8\end{pmatrix}^*$	$\begin{pmatrix}76.1 & 23.9\\23.8 & 76.2\end{pmatrix}$	$\begin{pmatrix}99.7 & 0.3\\0.6 & 99.4\end{pmatrix}^*$	$\begin{pmatrix}81.9 & 18.1\\16.1 & 83.9\end{pmatrix}$
fD	$\begin{pmatrix}92.5 & 7.5\\6.0 & 94.0\end{pmatrix}^*$	$\begin{pmatrix}97.5 & 2.5\\2.1 & 97.9\end{pmatrix}^*$	$\begin{pmatrix}99.4 & 0.6\\0.6 & 99.4\end{pmatrix}^*$	$\begin{pmatrix}99.6 & 0.4\\0.2 & 99.8\end{pmatrix}^*$	$\begin{pmatrix}94.2 & 5.8\\7.3 & 92.7\end{pmatrix}^*$
fE	$\begin{pmatrix}99.2 & 0.8\\0.7 & 99.3\end{pmatrix}^*$	$\begin{pmatrix}96.0 & 4.0\\5.0 & 95.0\end{pmatrix}^*$	$\begin{pmatrix}97.1 & 2.9\\2.5 & 97.5\end{pmatrix}^*$	$\begin{pmatrix}88.7 & 11.3\\12.8 & 87.2\end{pmatrix}$	$\begin{pmatrix}59.1 & 40.9\\39.2 & 60.8\end{pmatrix}$
mA	$\begin{pmatrix}91.1 & 8.9\\7.3 & 92.7\end{pmatrix}^*$	$\begin{pmatrix}92.2 & 7.8\\7.7 & 92.3\end{pmatrix}^*$	$\begin{pmatrix}65.2 & 34.8\\35.5 & 64.5\end{pmatrix}$	$\begin{pmatrix}95.5 & 4.5\\3.9 & 96.1\end{pmatrix}^*$	$\begin{pmatrix}90.0 & 10.0\\11.3 & 88.7\end{pmatrix}$
mB	$\begin{pmatrix}70.2 & 29.8\\31.9 & 68.1\end{pmatrix}$	$\begin{pmatrix}97.1 & 2.9\\3.4 & 96.6\end{pmatrix}^*$	$\begin{pmatrix}78.9 & 21.1\\22.8 & 77.2\end{pmatrix}$	$\begin{pmatrix}99.4 & 0.6\\0.6 & 99.4\end{pmatrix}^*$	$\begin{pmatrix}88.6 & 11.4\\14.6 & 85.4\end{pmatrix}$
mC	$\begin{pmatrix}77.6 & 22.4\\26.3 & 73.7\end{pmatrix}$	$\begin{pmatrix}96.4 & 3.6\\3.1 & 96.9\end{pmatrix}^*$	$\begin{pmatrix}96.0 & 4.0\\3.9 & 96.1\end{pmatrix}^*$	$\begin{pmatrix}96.6 & 3.4\\4.4 & 95.6\end{pmatrix}^*$	$\begin{pmatrix}90.3 & 9.7\\10.3 & 89.7\end{pmatrix}$
mD	$\begin{pmatrix}91.5 & 8.5\\11.1 & 88.9\end{pmatrix}^*$	$\begin{pmatrix}98.5 & 1.5\\1.4 & 98.6\end{pmatrix}^*$	$\begin{pmatrix}98.8 & 1.2\\1.3 & 98.7\end{pmatrix}^*$	$\begin{pmatrix}88.1 & 11.9\\12.8 & 87.2\end{pmatrix}$	$\begin{pmatrix}75.0 & 25.0\\24.6 & 75.4\end{pmatrix}$
mE	$\begin{pmatrix}93.5 & 6.5\\4.8 & 95.2\end{pmatrix}^*$	$\begin{pmatrix}98.2 & 1.2\\1.1 & 98.9\end{pmatrix}^*$	$\begin{pmatrix}91.2 & 8.8\\10.7 & 89.3\end{pmatrix}$	$\begin{pmatrix}99.4 & 0.6\\0.4 & 99.6\end{pmatrix}^*$	$\begin{pmatrix}83.4 & 16.6\\15.4 & 84.6\end{pmatrix}$

(b)

	Male Raters				
Expresser	**Laugh**	**Fear**	**Pain**	**Pleasure**	**Neutral**
fA	$\begin{pmatrix} 89.8 & 10.2 \\ 8.3 & 91.7 \end{pmatrix}$	$\begin{pmatrix} 65.3 & 34.7 \\ 36.7 & 63.3 \end{pmatrix}$	$\begin{pmatrix} 59.7 & 40.3 \\ 42.2 & 57.8 \end{pmatrix}$	$\begin{pmatrix} 80.6 & 19.4 \\ 20.9 & 79.1 \end{pmatrix}$	$\begin{pmatrix} 93.7 & 6.3 \\ 8.4 & 91.6 \end{pmatrix}^*$
fB	$\begin{pmatrix} 54.4 & 45.6 \\ 43.3 & 56.7 \end{pmatrix}$	$\begin{pmatrix} 72.9 & 27.1 \\ 26.8 & 73.2 \end{pmatrix}$	$\begin{pmatrix} 91.5 & 8.5 \\ 8.7 & 91.3 \end{pmatrix}^*$	$\begin{pmatrix} 81.5 & 18.5 \\ 19.2 & 80.8 \end{pmatrix}$	$\begin{pmatrix} 89.0 & 11.0 \\ 11.5 & 88.5 \end{pmatrix}$
fC	$\begin{pmatrix} 83.0 & 17.0 \\ 16.0 & 84.0 \end{pmatrix}$	$\begin{pmatrix} 89.9 & 10.1 \\ 9.4 & 90.6 \end{pmatrix}$	$\begin{pmatrix} 80.7 & 19.3 \\ 20.8 & 79.2 \end{pmatrix}$	$\begin{pmatrix} 80.3 & 19.7 \\ 18.2 & 81.8 \end{pmatrix}$	$\begin{pmatrix} 96.8 & 3.2 \\ 4.5 & 95.5 \end{pmatrix}^*$
fD	$\begin{pmatrix} 89.2 & 10.8 \\ 9.6 & 90.4 \end{pmatrix}$	$\begin{pmatrix} 83.0 & 17.0 \\ 16.4 & 83.6 \end{pmatrix}$	$\begin{pmatrix} 77.4 & 22.6 \\ 23.5 & 76.5 \end{pmatrix}$	$\begin{pmatrix} 93.0 & 7.0 \\ 7.4 & 92.6 \end{pmatrix}^*$	$\begin{pmatrix} 76.7 & 23.3 \\ 24.8 & 75.2 \end{pmatrix}$
fE	$\begin{pmatrix} 97.3 & 2.7 \\ 2.9 & 97.1 \end{pmatrix}^*$	$\begin{pmatrix} 87.7 & 12.3 \\ 13.4 & 86.6 \end{pmatrix}$	$\begin{pmatrix} 73.5 & 26.5 \\ 26.2 & 73.8 \end{pmatrix}$	$\begin{pmatrix} 87.3 & 12.7 \\ 12.6 & 87.4 \end{pmatrix}$	$\begin{pmatrix} 92.4 & 7.6 \\ 9.2 & 90.8 \end{pmatrix}^*$
mA	$\begin{pmatrix} 83.8 & 16.2 \\ 16.9 & 83.1 \end{pmatrix}$	$\begin{pmatrix} 99.9 & 0.1 \\ 0.2 & 99.8 \end{pmatrix}^*$	$\begin{pmatrix} 90.9 & 9.1 \\ 9.7 & 90.3 \end{pmatrix}^*$	$\begin{pmatrix} 78.3 & 21.7 \\ 22.0 & 78.0 \end{pmatrix}$	$\begin{pmatrix} 94.0 & 6.0 \\ 7.4 & 92.6 \end{pmatrix}^*$
mB	$\begin{pmatrix} 90.7 & 9.3 \\ 11.7 & 88.3 \end{pmatrix}$	$\begin{pmatrix} 94.2 & 5.8 \\ 5.5 & 94.5 \end{pmatrix}^*$	$\begin{pmatrix} 83.5 & 16.5 \\ 15.8 & 84.2 \end{pmatrix}$	$\begin{pmatrix} 95.6 & 4.4 \\ 4.5 & 95.5 \end{pmatrix}^*$	$\begin{pmatrix} 81.2 & 18.8 \\ 20.8 & 79.2 \end{pmatrix}$
mC	$\begin{pmatrix} 65.6 & 34.4 \\ 37.8 & 62.2 \end{pmatrix}$	$\begin{pmatrix} 95.9 & 4.1 \\ 3.5 & 96.5 \end{pmatrix}^*$	$\begin{pmatrix} 98.9 & 1.1 \\ 0.9 & 99.1 \end{pmatrix}^*$	$\begin{pmatrix} 76.1 & 23.9 \\ 23.1 & 76.9 \end{pmatrix}$	$\begin{pmatrix} 95.1 & 4.9 \\ 5.4 & 94.6 \end{pmatrix}^*$
mD	$\begin{pmatrix} 81.1 & 18.9 \\ 17.4 & 82.6 \end{pmatrix}$	$\begin{pmatrix} 94.2 & 5.8 \\ 5.5 & 94.5 \end{pmatrix}^*$	$\begin{pmatrix} 98.4 & 1.6 \\ 1.6 & 98.4 \end{pmatrix}^*$	$\begin{pmatrix} 89.2 & 10.8 \\ 11.7 & 88.3 \end{pmatrix}$	$\begin{pmatrix} 99.5 & 0.5 \\ 0.8 & 99.2 \end{pmatrix}^*$
mE	$\begin{pmatrix} 83.2 & 16.8 \\ 16.0 & 84.0 \end{pmatrix}$	$\begin{pmatrix} 95.7 & 4.3 \\ 4.3 & 95.7 \end{pmatrix}^*$	$\begin{pmatrix} 79.5 & 20.5 \\ 20.1 & 79.9 \end{pmatrix}$	$\begin{pmatrix} 83.7 & 16.3 \\ 16.9 & 83.1 \end{pmatrix}$	$\begin{pmatrix} 74.4 & 25.6 \\ 25.3 & 74.7 \end{pmatrix}$

Stress-induced Rating Differences

In Study II, we analyzed the differences in the distributions of ratings in the two groups of participants (control versus stressed group). The confusion matrices (Table 4) display the probabilities of differences of the ratings by the two groups of participants (separately for male vocalizations and female vocalizations). At a 10% significance level (Caelen, 2017), only two results are significantly different: the laugh and neutral for male vocalizations. The probability of correctness of attribution decreased in the stressed group for laugh from 92.5% of the control group to 81.8%, while for neutral the accuracy decreased from 98.5% to 89.6% (these probabilities are not shown). All the remaining ratings are unaffected by the stress induction procedure.

Table 4: The confusion matrices between the distributions of the ratings by the control raters and the stressed raters, male vocalizations and female vocalizations separately. Vocalizations that are significantly differently rated are marked with an asterisk.

Stimulus	Male Vocalizations	Female Vocalizations
Laugh	$\begin{pmatrix} 91.1 & 8.88 \\ 6.82 & 93.2 \end{pmatrix}^*$	$\begin{pmatrix} 53.6 & 46.4 \\ 34.5 & 65.5 \end{pmatrix}$
Fear	$\begin{pmatrix} 67.8 & 32.2 \\ 26.0 & 74.0 \end{pmatrix}$	$\begin{pmatrix} 66.6 & 33.4 \\ 30.1 & 69.9 \end{pmatrix}$
Pain	$\begin{pmatrix} 76.3 & 23.7 \\ 20.9 & 79.1 \end{pmatrix}$	$\begin{pmatrix} 71.7 & 28.3 \\ 24.5 & 75.5 \end{pmatrix}$
Pleasure	$\begin{pmatrix} 68.8 & 31.2 \\ 27.7 & 72.3 \end{pmatrix}$	$\begin{pmatrix} 68.8 & 31.2 \\ 27.7 & 72.3 \end{pmatrix}$
Neutral	$\begin{pmatrix} 95.8 & 4.17 \\ 2.62 & 97.4 \end{pmatrix}^*$	$\begin{pmatrix} 54.6 & 45.4 \\ 33.6 & 66.4 \end{pmatrix}$

DISCUSSION

In this paper, we aimed to investigate how isolated intense affective vocalizations are perceived, without context or other associated cues. With our study design, we could overcome some of the limitations of the previously used methodologies. Specifically: (a) We used vocalizations presented in more 'naturalistic' settings (i.e. occurring within an activity as opposed to in a laboratory). (b) For rating responses, we did not request classification of the emotion as a psychological category (i.e. fear, pain) but rather requested the raters to evaluate them using a valence-based rating (as positive, negative, or neutral). (c) It was therefore possible to integrate both theories of emotion (as discrete categories or as a dimensional phenomenon) in one experimental setting. (d) The sex of the raters and of the vocalizers were taken into consideration in the study.

The outcomes presented in this paper confirm previous published findings about the emotion intensity paradox: vocalizations of high-intensity affective states (pain, pleasure and fear) are misattributed (and therefore not correctly identified) with very high probabilities. In comparison, the low intensity vocalizations (laugh and neutral) were correctly attributed with high probabilities.

Among the high-intensity affective states, we found that the basic emotion — fear — was more often assigned positive valence (Fig 1c) in contrast to the other two intense affective states that were tested (pain and pleasure). This result confirms — to some degree — the basic emotion theory. While we did not find support for the universality of fear perception (the ratings were incorrect, both for male and female raters — albeit not due to guessing), we did find that the processing of this emotion by the raters was different from the other high-intensity affective states.

An alternative reason why fear may be more misattributed and was predominantly rated as positive (Table 1c) is the specificity of the stimuli used in the current study. The experience of fear may be elicited by an unexpected 'scary' or 'surprising' situation — in stark contrast to the stimulus we presented. In our case, the expectation of unpleasant experiences that will happen very soon (among them, spanking) seemed not to overly surprise the expressers. In more conventional cases (i.e. the ones most often studied), the emotion of fear is intermixed with

surprise — in contrast to the case we studied, in which it is mixed with an anticipatory anxiety. The vocalization mainly consists of intense breathing and soft weeping (connected as it is with anxiety) and therefore the stimulus may be perceived as more positive by the raters. This may be especially misleading when other positive but difficult-to-categorize vocalizations are present. Further studies contrasting vocalizations of both these types of fear (scream as a result of a sudden, fear-inducing emotion — such as a scene in a horror video (Prossinger et al., 2021) versus vocalizations of fear due to anticipatory anxiety — as just before a bungee jump) would contribute to clarifying this issue.

The results for pain vocalizations are also in agreement with previous findings; it was the least correctly rated affective state among all the stimuli presented (Anikin et al., 2017; Lima et al., 2013; Belin et al., 2008).

On the other hand, perhaps unexpectedly, the results for pleasure vocalizations contradict several previously presented study outcomes in which pleasure was either well-recognized (Lima et al., 2013; 2014) or at least correctly attributed to the positive valence rating (Belin et al., 2008). However, our results are in agreement with the studies on vocalization of intense affective states and with previous studies on emotional facial expressions, further confirming that for stimuli with high intensity it is more difficult to extrapolate the correct valence from one single, isolated component (i.e. exclusively vocalization or exclusively facial expression). This interpretation is further supported by the very different results we found for low intensity affective states, showing that our participants correctly assessed the valence of these types of stimuli (Table 1a and 1b).

One argument for the phenomena being counterintuitive is that in such highly intense emotional states the context (i.e. what elicits the emotion) very clearly points to the valence of the emotion, and, consequently, the affective expression itself need not convey information regarding its valence but rather bring attention to the stimulus within its context. In most of these close-to-natural scenarios, the context would provide sufficient further information that would then contextualize the (vocal and facial) expressions, enabling ratings as positive or negative.

One natural scenario in which the context may not be of assistance to correctly assess the expression (in terms of facial expression or/and vocalization) is sexual intercourse. Indeed, in the sexual intercourse situation, the stimulus (for example, penetration) may lead to either a positive (pleasure) or a negative (pain) affective state and to the display of such an affective state. One further ramification of the present study is that it highlights how our intuitive interpretation may be in error (as do some others dealing with the emotional intensity paradox; Holz et al., 2021; Atias et al., 2019). As pointed out by Boschetti at al. (2022) for facial cues, this present study brings attention to the possibility of misunderstanding of cues, especially in the above-mentioned situations. The misattribution in these contexts can be avoided when they are accompanied with clarifying verbal communications.

Our novel analytical approach allows us to investigate not only the correct versus incorrect ratings of the stimuli, but also the probability that the ratings could be due to chance: the raters could be guessing, but correctly guessing (or incorrectly guessing). We find that the participant's ratings are not due to chance. When rating the emotional vocalizations, the participants are not guessing the valence (positive, neutral or negative) but they are convinced of the valence of their rating, even when they are incorrectly rating. In another study in which the due-to-chance probability of rating facial expression perception was analyzed (Boschetti et al., 2022), the findings were very different. When rating facial expressions of intense emotions, the participants guessed the valence (specifically — in contrast to the findings presented in this paper) wrongly. When the participants rated the vocalizations (while not being able to see the facial expression), the participants did not guess; they were convinced of the valence of their (wrong) ratings.

A further insight we gained while researching the vocalizations is the consistency of their rating. It is rarely studied even though it should constitute one of the fundamental questions. Is the rating repeatable? We had this expectation for our stimuli; interestingly, only few vocalizations were consistently rated (Table 3). Surprisingly, the female raters were highly consistent in ratings of neutral vocalizations by men but not by women — whereas men were consistent in rating the female fear and male laugh vocalizations. Conversely, an extreme inconsistency was found in the case of men rating male fear, women rating fear and laugh. All three of these outcomes draw into question the classical concept of basic emotion perception — especially the just-so-stories about the female's greater ability to assess positive affects.

As in the case of facial expressions using the same methodology (Boschetti et al., 2022) the due-to-chance analyses provided a novel tool to study affect perceptions. We found that facial expressions (other than laugh and neutral) are rated due to chance. The results for vocalizations are stunningly different. None of the stimuli, no matter how ambiguous, was rated with uncertainty on the part of the raters. This shows that there is a high reliance on the acoustic perception when compared to the visual perception in the case of affect perception.

It was expected that women would be better at correct attribution, especially in case of negative emotion evaluation (Thompson & Voyer, 2014; Belin et al., 2008). In a study by Vasconcelos et al. (2017), the effect of the sex of the rater was specific for the emotional category (i.e. better recognition of anger and sadness vocalization by females in contrast to surprise by males). Our result is in disagreement with both these previous studies; we did not identify any advantage on the side of any sexes in attribution accuracy, nor on the expresser's sex effect. Nor did we find any support for the finding that some vocalization category was better identified.

We conclude that the intrasexual variation was high on the side of the vocalizers. Some male and some female expressers were rated more accurately than the others (Table 3). This should be further studied by including a possible similar effect on the side of the rater. It can be expected that there are individuals with a higher ability to differentiate the vocalizations. While it exceeds the scope of this article, clustering algorithms are a viable way to identify sub-groups of individuals through multiple assessments (accuracy, consistency, and due-to-chance ratings). The influence of stress on rating of vocalizations is a unique feature of our study. None of the results are due-to-chance and the shift only occurred in case of the non-ambiguous vocalizations (neutral and laugh). The direction is always towards the lower accuracy of ratings and only for the male expressers. We do not have an interpretation for this result; it is the first time it is presented, so we cannot compare with published studies. In a similar study focused on the facial expressions, it was pleasure and smile of the male expressers that were rated more accurately by the stressed group. The inner state (stress induction) alters neither the female facial nor vocal ratings. The male vocalizations may be perceived in an altered way as caution for dangers. Again, further studies would be necessary to extend our knowledge on the topic.

Limitations and Future Directions

Bayesian statistics is not susceptible to misinterpretations of a null effect (because the Bayesian methods do not violate Bayes' Theorem). Bayesian methodology specifically includes testing for a null result (when the posterior is not significantly different from the prior). This approach is promising for future research.

We tested for two types of null result. One null result (often observed): the outcome of a statistical test shows that the observed effect is due to chance. The other type we tested for: that the observed difference of a result that is not due to chance but the detected difference is valid with a very small probability.

In both studies presented here, the samples of both stimuli and raters consisted of members of a Caucasian population, since the diversity of population in Czech Republic is minimal. The results, although very strong, may not be directly generalizable to other populations.

Female sexual pleasure is objectively difficult to assess; but this difficulty applies to all related research. There are claims that even self-reports would not be sufficient. Devices used for measuring female sexual arousal are insufficiently reliable (Meston et al., 2004; Cooper et al., 2014; Meston et al., 2019), so we did not use them in this investigation. As in other studies that attempt to relate arousal with female pleasure expression, we use the pragmatic approach: for stimulus creation, it is sufficient to adopt the convention of relying on using already existing, freely downloadable videos with distinctive human vocalizations. Researchers who question this pragmatic approach must then reject the validity of a vast number of studies dealing with vocal expressions of pleasure, not only those using videos. However, it should be pointed out that applying the AI methods to human vocalizations (as was done for facial expressions, Prossinger et al., 2022) have the potential of resolving this impasse.

By the same token, we feel the need to address the possibility that the expression heard does not match the inner feeling of the expresser. This is not a design flaw but involves an inherently biological aspect in the field of research using naturalistic stimuli.

Furthermore, the expression of fear as a reliable stimulus may be considered problematic since the expressers were aware of the fact that, ultimately, the situation is safe: no permanent damage is de facto guaranteed by the plot of the video. Fear, of all expressions considered basic, has the lowest identification reliability rate, and this is especially true in naturalistic expression scenarios. In other words, the results obtained are less unusual than may appear at first glance.

Lastly, it should be pointed out that the situation of sexual play is not transferable to other types of interaction in which such mismatches can be found, e.g., sport, fighting, injury infliction. Therefore, generalization of our findings to include such fields may have to be used with caution.

CONCLUSION

These two studies we presented here bring multiple novel insights to vocalization perception. The first study, with a large sample of participants (exceeding 900) in combination with novel analytical statistical approaches provide us with numerous findings. The low arousal vocalizations (laugh and neutral state) are rated with very high accuracy whereas the fear, pain and pleasure are not.

The ratings of the pain are so scattered (extremely high variability) that it is impossible to assign them a mode whereas the rating of pleasure is almost equally distributed on the extreme poles of the rating distribution — making both of these vocalizations rated with insufficient accuracy. Fear was highly mistaken for positive vocalization; it can be interpreted for this specific type of situation where surprise is not involved.

We found no sex differences between the vocalizers or the raters to have an impact — with one exception. There is a pattern of consistency rating where female raters were consistent in ratings of neutral vocalization by men, but not by women, whereas male raters were consistent in rating the female fear and male laugh. Conversely, an extreme inconsistency was found in the case of men rating male fear, women rating fear and laugh.

None of the ratings were due to chance; actually the probability of the rating being due-to-chance (guessing) was smaller than one percent. The ratings were often incorrect and inconsistent, but they are not the result of guessing.

The second study showed shifts in two male vocalizations — laugh and neutral — after a stress induction procedure. Ratings for both these vocalizations were less accurate in the stressed group.

These many outcomes provide further support for the emotion intensity paradox yet also undermine some of the core concepts of emotional vocalization research. Further studies will be necessary to uncover more about the phenomena we discovered; especially using naturalistic and semi-naturalistic ecologically valid stimuli so as to avoid pre-tested and laboratory obtained stimuli. We recommend that even those studies that rely on stimuli datasets (with known previous results) should be tested using the novel statistical methods provided herein. Lastly, the Cold Pressor Task is an ideal stress induction method; there is a lack of studies in which the arousal of respondents is altered; arguably, a key question related to ecological validity.

ETHICS

Although the materials presented to the participants were not *per se* of a sexual nature (as only audio records were presented), we made precautions to limit any negative impact on our raters.

Informed consent

In Study I: An online information text and consent form was supplied; after reading it, a box was to be ticked by each participant (indicating their informed consent) prior to their participation.

In Study II: Two informed consent forms were to be manually/personally signed. The first was presented to a potential rater prior to participation; it included all the information about procedures (including the CPT), safety measures, kinds of data collected, and risks. The second informed consent form consisted of a full disclosure of the aim(s) of the study, the expected impact of the procedures, and the possible implications for the rater signing this second form. It was to be signed after the debriefing procedure (see below). If the second consent form was not signed, the collected data was discarded (and therefore not used in the analysis).

Post-study Support and Debriefing

All parts of the design and debriefing were conducted in co-operation with a trained psychologist who also supervised all data collection.

For Study I we supplied the participants with a list of contacts: (1) to the principal investigator, (2) to a psychological counseling center, and (3) to an organization that deals with sexuality-related issues.

During the debriefing phase for Study II, every rater participated in a debriefing discussion by a trained psychologist directly after the completion of data collection. The rater then received a written detailed description, with a full explanation of the possible negative aspects of the experiment, especially those related to the stress-induction procedure, and was also supplied with a list of contacts: (1) to the principal investigator, (2) to a psychological counseling center, and (3) to an organization that deals with sexuality-related issues.

ACKNOWLEDGMENTS

The study was preregistered on the OSF portal: https://osf.io/bhk6m/.

FUNDING

This research was funded by the Czech Science Foundation in the project with number GACR 19-12885Y and titled “Behavioral and Psycho-Physiological Response on Ambivalent Visual and Auditory Stimuli Presentation”.

INSTITUTIONAL REVIEW BOARD STATEMENT

This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Faculty of Science, Charles University, Prague, Czech Republic (Protocol Code 2018/08, approval date: 2 April 2018).

DATA AVAILABILITY STATEMENT

The data are available on the OSF portal (link: https://osf.io/5j6zx).
The frames were extracted from commercially available online videos. As the videos are proprietary, we can only make the extracted frames we used available from the corresponding author (upon reasonable requests originating from a serious institutional email address).
The analysis was executed in MATHEMATICA® (version 13.2) from Wolfram Technologies. Interested readers can request sections of code from the corresponding author.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Anikin, A., Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior Research Methods, 49*(2), 758–771. DOI

Anikin, A., Pisanski, K., & Reby, D. (2020). Do nonlinear vocal phenomena signal negative valence or high emotion intensity? *Royal Society open science, 7*(12), 201306. DOI

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology, 25*(15), 2051–22056. DOI

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2019). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General, 148*(10), 1842. DOI

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229. DOI

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*(2), 531–539. DOI

Binter, J., Boschetti, S., Hladký, T., Prossinger, H., Wells, T. J., Jílková, J., & Říha, D. (2022). Quantifying the Rating Performance of Ambiguous and Unambiguous Facial Expression Perceptions Under Conditions of Stress by Using Wearable Sensors. In *International Conference on Human-Computer Interaction* (pp. 519–529). Springer, Cham. DOI

Boschetti, S., Prossinger, H., Hladký, T., Machová, K., & Binter, J. (2022). "Eye can't see the difference": Facial Expressions of Pain, Pleasure, and Fear Are Consistently Rated Due to Chance. *Human Ethology, 37*, 46–72. DOI

Brown, C. C., Raio, C. M., & Neta, M. (2017). Cortisol responses enhance negative valence perception for ambiguous facial expressions. *Scientific Reports, 7*(1), 1–8. DOI

Bryant, G., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8*(1-2), 135-148. DOI

Bryant, G. A. (2021). The evolution of human vocal emotion. *Emotion Review, 13*(1), 25–33. DOI

Bullinger, M., Naber, D., Pickar, D., Cohen, R. M., Kalin, N. H., Pert, A., & Bunney Jr, W. E. (1984). Endocrine effects of the cold pressor test: relationships to subjective pain appraisal and coping. *Psychiatry Research, 12*(3), 227–233. DOI

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence, 81*(3), 429–450. DOI

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech communication, 40*(1–2), 5–32. DOI

Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences, 98*(17), 10006–10010. DOI

Dubray, S., Gérard, M., Beaulieu-Prévost, D., & Courtois, F. (2017). Validation of a self-report questionnaire assessing the bodily and physiological sensations of orgasm. *The Journal of Sexual Medicine, 14*(2), 255–263. DOI

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic? *Emotion Review, 3*, 364–370. DOI

Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Smith, M. J. L., Cornwell, R. E., ... & Perrett, D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior, 26*(5), 398–408. DOI

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*(4), 911–920. DOI

Han, H., Byun, K., & Kang, H. G. (2018). A deep learning-based stress detection algorithm with speech signal. In *Proceedings of the 2018 workshop on audio-visual scene understanding for immersive multimedia* (pp. 11–15). DOI

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific reports, 11*(1), 1–10. DOI

Hughes, S. M. & Nicholson, S. E. (2008). Sex differences in the assessment of pain versus sexual pleasure facial expressions. *Journal of Social, Evolutionary, and Cultural Psychology, 2*(4), 289. DOI

Hughes, S. M., & Puts, D. A. (2021). Vocal modulation in human mating and competition. *Philosophical Transactions of the Royal Society B, 376*(1840), 20200388. DOI

Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*(2), 288–299. DOI

Kamiloğlu, R. G., Tanaka, A., Scott, S. K., & Sauter, D. A. (2022). Perception of group membership from spontaneous and volitional laughter. *Philosophical Transactions of the Royal Society B, 377*(1841), 20200404. DOI

Kibrik, A., & Molchanova, N. (2013). Channels of multimodal communication: Relative contributions to discourse understanding. In P*roceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons. DOI

Leongómez, J. D., Havlíček, J., & Roberts, S. C. (2022). Musicality in human vocal communication: an evolutionary perspective. *Philosophical Transactions of the Royal Society B, 377*(1841), 20200391. DOI

Leongómez, J. D., Pisanski, K., Reby, D., Sauter, D., Lavan, N., Perlman, M., & Varella Valentova, J. (2021). Voice modulation: from origin and mechanism to social impact. *Philosophical Transactions of the Royal Society B, 376*(1840), 20200386. DOI

Lima, C. F., Alves, T., Scott, S. K., & Castro, S. L. (2014). In the ear of the beholder: how age shapes emotion processing in nonverbal vocalizations. *Emotion, 14*(1), 145. DOI

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*(4), 1234–1245. DOI

McGettigan, C., Walsh, E., Jessop, R., Agnew, Z. K., Sauter, D. A., Warren, J. E., & Scott, S. K. (2015). Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity. *Cerebral Cortex, 25*(1), 246–257. DOI

Pisanski, K., Kobylarek, A., Jakubowska, L., Nowak, J., Walter, A., Błaszczyński, K., ... & Sorokowski, P. (2018). Multimodal stress detection: Testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test. *Hormones and Behavior, 106*, 52–61. DOI

Pisanski, K., & Reby, D. (2021). Efficacy in deceptive vocal exaggeration of human body size. *Nature Communications, 12*(1), 1–9. DOI

Pisanski, K., & Sorokowski, P. (2021). Human stress detection: cortisol levels in stressed speakers predict voice-based judgments of stress. *Perception, 50*(1), 80–87. DOI

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*(4), 417–435. DOI

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology, 17*(3), 715–734. DOI

Prasetio, B. H., Tamura, H., & Tanno, K. (2019, October). A deep time-delay embedded algorithm for unsupervised stress speech clustering. In: *2019 IEEE international conference on systems, man and cybernetics* (SMC) (pp. 1193–1198). DOI

Prossinger, H., Hladky, T., Binter, J., Boschetti, S., & Riha, D. (2021). Visual Analysis of Emotions Using AI Image-Processing Software: Possible Male/Female Differences between the Emotion Pairs "Neutral"–"Fear" and "Pleasure"–"Pain". *The 14th Pervasive Technologies Related to Assistive Environments Conference* (pp. 342–346). DOI

Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences, 279*(1728), 601–609. DOI

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161. DOI

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*(6), 2408–2412. DOI

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication, 40*(1-2), 227–256. DOI

Thompson, A. E.; Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion, 28*(7), 1164–1195. DOI

Tolkmitt, F. J., & Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance, 12*(3), 302. DOI

Vasconcelos, M., Dias, M., Soares, A. P., & Pinheiro, A. P. (2017). What is the melody of that voice? Probing unbiased recognition accuracy with the Montreal affective voices. J*ournal of Nonverbal Behavior, 41*(3), 239–267. DOI

Wenzler, S., Levine, S., van Dick, R., Oertel-Knöchel, V., & Aviezer, H. (2016). Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion, 16*(6), 807. DOI

APPENDIX I

The Necessity of Bayesian Statistics

Based on a request by a reviewer of this paper, we supply a brief description of some aspects of Bayesian statistics methodology. There are numerous introductory textbooks on Bayesian statistics (MacKay, 2015; Bishop, 2011; Lambert, 2018), which we do not emulate here.

In Frequentist statistics, estimators supply a numerical value — a number, aptly called an estimate. For example, given a data set of scalars (such as age), an estimate of the expectation value is the average (mean). How close this estimate is to the expectation value is usually given by a rule of thumb, oftentimes called the 'standard error of the mean'. In Bayesian statistics, estimators are likelihood functions. These are updated with supplied evidence. In Frequentist statistics, estimates of estimators may improve with repeated sampling (if not, the sample has been drawn from more than one statistical population); the estimates derived from a Frequentist analysis actually require an infinite number of *samples* drawn from the *same* distribution. In practically all uses of statistics in ethology, psychology, sociology, political science, and related endeavors bordering the humanities, this requirement of infinite sampling cannot be achieved. Consider the prediction of the outcome of an election. It is impossible to collect more than one sample, so Frequentist statistics does not apply, despite many publications that, sadly, do.

An example: consider three data sets of women (XX carriers) and men (XY carriers) with fixed ratios: 2 women, 1 man; 20 women, 10 men; 200 women, 100 men. If the question is "What is the probability of meeting a woman in the population from which this data set was drawn?", the answer in Frequentist statistics is always $\frac{2}{3}$ — the answer is always a numerical value (a number).

In Bayesian statistics, the probability is a random variable (conventionally written s) with $0 \leq s \leq 1$, and not a number. The answer to the above question about the probability is therefore not a numerical value, but a function called the likelihood function (written $\mathcal{L}(s)$). In this example of the three data sets with constant ratios, there are three likelihood functions, namely

$$\mathcal{L}_{\mathrm{A}} = const_A \times s^2 \times (1-s)^1$$
$$\mathcal{L}_{\mathrm{B}} = const_B \times s^{20} \times (1-s)^{10}$$
$$\mathcal{L}_{\mathrm{C}} = const_C \times s^{200} \times (1-s)^{100}$$

because, if s is the probability of meeting a woman, $(1-s)$ is the probability of meeting a man.

The constants $const_A$, $const_B$, and $const_C$ are determined by integration, because the likelihood function is normalized, meaning that $\int_0^1 \mathcal{L}(s)ds = 1$.

The graphs of these three likelihood functions are in Fig. App-1.

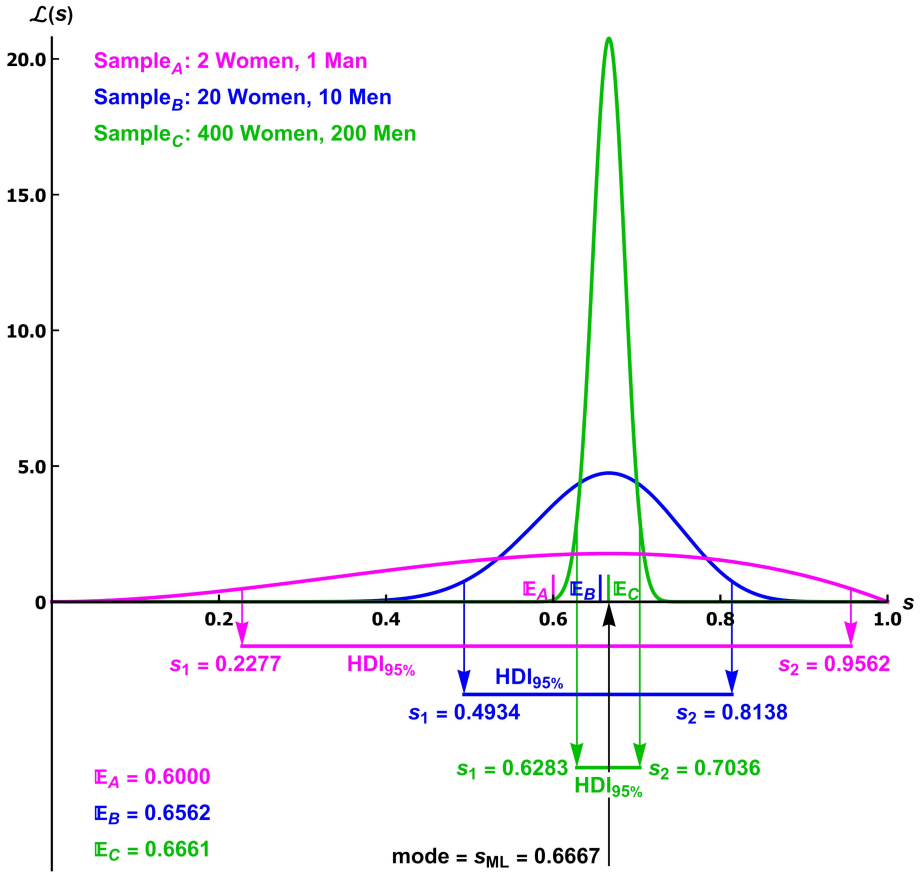


Figure A-1: The likelihood functions for three samples of binary data points (women and men), showing how the credibility interval decreases for larger sample sizes (not: more samples). By construction of this example, the ML probability does not change, but the credibility interval ($HDI_{95\%}$; Kruschke, 2014) does decrease. $\mathbb{E}_A$, $\mathbb{E}_B$, and $\mathbb{E}_C$ are the expectation values of the distributions. The credibility interval is not to be equated with the confidence interval of Frequentist statistics; the credibility interval rigorously gives the range of probabilities in which the true probability lies (Lambeth, 2018). We note that the bounds of the credibility interval are not symmetrically placed about the mode. Furthermore, the expectation values are far from the mode for small sample sizes.

One observes that the maximum likelihood of s (written s_{ML}) is always $\frac{2}{3}$ (this example has been constructed so as to match the estimate of the Frequentist approach). However, Bayesian statistics is far more rigorous; it supplies the credibility intervals in a rigorous manner, without any 'rule of thumb' (Lambeth, 2018; p. 26), and it allows for a calculation of this interval for any chosen significance level, to be calculated mathematically (granted, often in a numerically nontrivial manner). Another shortcoming of the Frequentist approach is a logical one: because a Frequentist probability is a number, it logically does not have an uncertainty (one example of sleight of hand corrections — specifically pointed out in Krutschke (2011), quoted below). Furthermore, Bayesian methods are usable for all sample sizes (without a sleight of hand) and specifically point to where sensitivity to sample size occurs. (In the Frequentist approach, the Pavlovian ritual is "make the sample size large" — often requested by reviewers. We deal with this issue in more detail below.)

This Appendix describes Bayesian approaches to data analysis. There are numerous references to why Frequentist approaches are severely inadequate. We only mentioned two with the intention of showing how the Bayesian approach does not use hand-waving as does

the Frequentist approach to deal with inadequacies (such as Bonferroni (1936) and/or Šidák (1967) corrections). The likelihood function $\mathcal{L}(s)$ supplies an *analysis* independent of sample size (which the Frequentist approach cannot) and this function also shows how the Bayesian estimators become Frequentist ones as the sample size approaches infinity (known as the Laplace limit). This approach to infinity can be, in many situations, very, very gradual; sample sizes of 100 thousand can oftentimes be necessary.

However, the requirement of close to infinitely many samples not being met leads to a, sadly, oftentimes committed error:

"In the Frequentist paradigm, … [a] 95% confidence interval means that across the infinity of intervals that we calculate, the true value of the parameter will lie in this range 95% of the time.

In reality, we draw only one sample from the population and we have no way of knowing whether the confidence interval we calculate actually contains the true parameter value. This means that … for 5% of the samples, the confidence intervals will be nonsense." (Lambeth, 2018; p. 131)

Note that the presented example was for a binary categorical variable (women versus ¬women = men). The fallacious approach of assigning numerical values (computable numbers) to categorical variables has been eliminated completely.

In this paper, we frequently (no pun intended) use non-binary categorical variables, such as $A = positive, B = neutral,$ and $C = negative$. In this case, the likelihood function is

$$\mathcal{L}(s_A, s_B, s_C) = constant \cdot {s_A}^{n_A} \cdot {s_B}^{n_B} \cdot {s_C}^{n_C}$$

with $n_A + n_B + n_C = n$ (total sample size) and $s_A + s_B + s_C = 1$ (there is only 1 probability variable for each category).

Technical notes: (a) The likelihood function for the binary case is the *pdf* (probability density function) of the Beta distribution, and the likelihood function for categorical variables with more than two categories is the *pdf* of the Dirichlet distribution (which can be defined for an arbitrarily large, finite number of categories).
(b) Furthermore, in order to simplify this presentation, we have assumed the prior likelihood $\mathcal{L}_{prior} = 1$ and $\mathcal{L}_{posterior} = evidence \cdot \mathcal{L}_{prior} = \mathcal{L}(s)$. (Further technical details can be found in the references.)

Because, in Bayesian statistics, every parameter is a random variable, it must have its own likelihood function (and therefore is not a number). Finding these likelihood functions can lead to some seemingly unwieldy formulae.

For example, if a researcher is interested in the likelihood functions of the parameters β_i (with $i = 1...k$ in an ordinary least-squares regression $y_i = x_i^T \cdot \beta + \epsilon_i = x_{i1} \cdot \beta_1 + \ldots + x_{ik} \cdot \beta_k$ $i = 1...n$, (in a more compact notation: $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon}$, with the classical, Frequentist solution $\hat{\beta} = (X^T X)^{-1} X^T y$) the solution in Bayesian statistics is

$$\mathcal{L}_{posterior}(\beta, \sigma^2 \,|\, y, X) \propto \mathcal{L}_{normal}(\beta, \sigma^2 \,|\, y, X) \cdot \mathcal{L}_{inverseGamma}(\sigma^2 \,|\, y, X)$$

for the prior

$$\mathcal{L}_{prior}(\beta, \sigma^2 \,|\, y, X) \propto \mathcal{L}_{normal}(\beta, \sigma^2) \cdot \mathcal{L}_{inverseGamma}(\sigma^2).$$

(This solution is readily available in several — but by no means all — software packages, most notably STAN; further information can be obtained from the corresponding author.) As a

consequence, estimating the uncertainties using Bayesian statistics (credibility intervals) does not agree with those obtained by the Frequentist approach (confidence intervals). The uncertainties in classical, Frequentist statistics textbooks always attempt to improve the solutions with a host of ad hoc 'correction factors', which never really do the job (see the Kruschke (2011), quote below). Another method of finding the uncertainty is to repeatedly regress with one sample point left out — the so-called jack-knife (or LOO — leave one out; James et al., 2021) approach. Again, these methods only apply to very large data sets (Laplace limit strikes again), causing a further problem: how much do the (Frequentist) estimates of the coefficients vary, if one data point of many thousands is left out?

In a publication primarily addressing psychologists and practitioners in related fields (presumably including readers of *Human Ethology*), Kruschke (2011), addressing why Frequentist statistics cannot be used when only one *sample* (such as an election), rather than close to infinitely many, is available, writes:

" … The remainder of this article highlights the complementary strengths of the two Bayesian approaches and emphasizes that both are better than NHST. Either Bayesian approach is superior to NHST. As was emphasized earlier in the article, in NHST it is impossible to decide whether $p < 0.05$ because p itself is ill-defined and *cannot be uniquely calculated* [our emphasis with italics]. NHST yields no measure of the relative credibility of null and alternative models, and NHST yields no measure of the credibility of different candidate parameter values. NHST suffers from sampling to a foregone conclusion. For multiple comparisons, NHST uses intention-based corrections whereas Bayesian analysis uses rationally informed shrinkage."

To summarize: By now, reviewers — and potential readers — should have realized that a submitted manuscript should always incorporate Bayesian methods; only in rare cases would a Frequentist approach apply. In studies that involve only one sample, Frequentist statistics is never justified. The author(s) of a manuscript that presents results based on Frequentist statistics need to *justify* why he/she/they used Frequentist methods and not Bayesian ones. Bayesian methods are *always* applicable; whereas Frequentist ones are not. It should not be the case that a reviewer requests that manuscript authors justify why they used Bayesian methods.

REFERENCES

Bishop, C.M. (2011). *Pattern Recognition and Machine Learning*. Springer. NY, USA.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8, 3–62.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. (Springer Texts in Statistics), New York, USA.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. DOI

Kruschke, J.K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, New York, USA.

Lambert, B. (2018). *A Student Guide to Bayesian Statistics*. Sage Publications. London, UK.

MacKay, D.C. (2015). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, UK.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*(318), 626–633. DOI

APPENDIX II

Dirichlet Distribution

The ratings by female raters are Dirichlet distributed (in our case with three concentration parameters $\{\alpha_A, \alpha_B, \alpha_C\}$), as are those of the male raters. We predicted the repeats (Trial I versus Trial II in Study I) to be the same, and we tested for that. We therefore have, for female raters rating five female fear vocalizations, ten registration sets with triples $\{n_A, n_B, n_C\}$ in each set, with $n_A + n_B + n_C = 10$. The *pdf* (probability density function) of the Dirichlet distribution Dir, called the likelihood function $\mathcal{L}(s_1, s_2, s_3) = pdf(Dir(\alpha_A, \alpha_B, \alpha_C), s_1, s_2, s_3)$, with concentration parameters $\{\alpha_A, \alpha_B, \alpha_C\}$ and probabilities s_1, s_2, s_3 of observing the variables var_1, var_2, var_3 is

$$\mathcal{L}(s_1, s_2, s_3) = pdf(Dir(\alpha_A, \alpha_B, \alpha_C), s_1, s_2, s_3) = \frac{\Gamma(\alpha_A + \alpha_B + \alpha_C)}{\Gamma(\alpha_A)\Gamma(\alpha_B)\Gamma(\alpha_C)} s_1^{\alpha_A - 1} s_2^{\alpha_B - 1} s_3^{\alpha_C - 1}$$

with $s_3 = 1 - s_1 - s_2$ and $0 \le s_i \le 1 \; \forall i \; i = 1 \ldots 3$; $\Gamma(\cdots)$ is the Gamma function.

The two modes for A and C are $mode_A = \dfrac{\alpha_A - 1}{\alpha_A + \alpha_B + \alpha_C - 3}$ and $mode_C = \dfrac{\alpha_C - 1}{\alpha_A + \alpha_B + \alpha_C - 3}$.

If we are interested in axes A and B, rather than A and C, then the formulae are cycled. In the text, we justify why we use which axes and when. Note that the formulae for the modes are straightforward, suggesting we need not use the (somewhat complicated) formula for the probability density function *pdf*.

If, as is the case in this project, there are 5 modes for the female stimulus (vocalization), and each has been rated twice, we have 10 modes in the domain triangle (as defined above). We use the scores (ratings) to estimate the Dirichlet distribution in order to obtain the mode for female raters of male stimulus for two rating tasks.

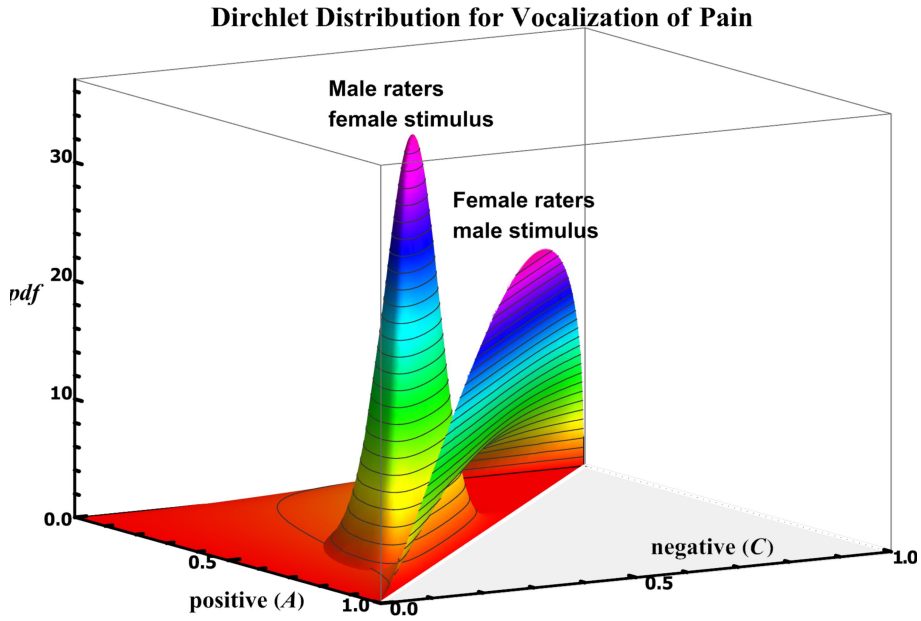


Figure App-2: The *pdf*s of two Dirichlet distributions, one of which has no mode within the domain. Note that the *pdf* of one of the Dirichlet distributions approaches infinity beyond the boundary of the domain (the hypotenuse in this example). Both Dirichlet distributions were determined using the ratings of 10 stimuli (either 5 females or 5 males,

each rated by the raters twice). Further implications of the (mathematical) divergence of the *pdf* of one of the Dirichlet distributions are discussed in the text. The surface of one *pdf* can be seen inside the other *pdf* surface at low likelihood levels. This visibility is intentional, in order to aid in reading the graph.

The ratings do not ensure, of course, that a mode will exist within the (triangular) domain. It can — and does — happen that there exists no such mode (Figure App-2). The *pdf* of the Dirichlet distribution may diverge along or beyond one of the boundaries of the domain. In such a case, the coordinates for the mode are undefined and an interpretation of the statistical properties of the ratings becomes subtle. Such interpretations are to be found in the text for the case of pain vocalizations.

Bayesian estimation of guessing

Each stimulus is rated as exhibiting one of the five vocal expressions. We do not expect, but do postulate — as a test — that the vocal expression laugh (for example) will be rated positive, while the vocal expression pain will be rated negative. We use a Bayesian approach to determine the maximum likelihood of a correct probability(!). For each stimulus of each facial expression rated by the females (say), let n_1 be the number of ratings that agree with the postulated rating, while n_2 is the number of ratings that disagree with the postulated rating (then $n_1 + n_2 = n$; $n = 526$ for female raters; $n = 376$ for male raters). In Bayesian statistics, in which the probability s is a random variable, the likelihood function, for this situation, $\mathcal{L}(s)$ of s is the *pdf* of a Beta Distribution

$$\mathcal{L}(s) = pdf(Be(\alpha,\beta),s) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha-1}(1-s)^{\beta-1} = \frac{\Gamma(n_1+n_2+2)}{\Gamma(n_1+1)\Gamma(n_2+1)} s^{n_1}(1-s)^{n_2}$$

The probability (in Bayesian statistics) of observing a result disagreeing with the postulate is then,

$$\int_0^{1/2} \mathcal{L}(Be(\alpha,\beta),s)ds$$

The most likely probability s_{ML} is the mode. s_{ML} = mode = $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$. We note that the postulate is always s, even if the postulated rating is negative (as in the case of pain).