

1 **Sound context modulates perceived vocal emotion**

2 Marco Liuni ^{1*}, Emmanuel Ponsot ^{2,1}, Gregory A. Bryant ^{3,4}, JJ Aucouturier ¹

3

4 ¹ STMS Lab (IRCAM/CNRS/Sorbonne Universités)

5 ² Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale
6 supérieure, Université PSL, CNRS, Paris, France.

7 ³ UCLA Department of Communication

8 ⁴ UCLA Center for Behavior, Evolution, and Culture

9 *Correspondence to: M. Liuni, IRCAM, 1 place Igor Stravinsky, 75004 Paris, FR

10 marco.liuni@ircam.fr

11

12 Abstract

13 Many animal vocalizations contain nonlinear acoustic phenomena as a consequence of
14 physiological arousal. In humans, nonlinear features are processed early in the auditory system,
15 and are used to efficiently detect alarm calls and other urgent signals. Yet, high-level emotional
16 and semantic contextual factors likely guide the perception and evaluation of roughness features
17 in vocal sounds. Here we examined the relationship between perceived vocal arousal and
18 auditory context. We presented listeners with nonverbal vocalizations (yells of a single vowel) at
19 varying levels of portrayed vocal arousal, in two musical contexts (clean guitar, distorted guitar)
20 and one non-musical context (modulated noise). As predicted, vocalizations with higher levels of
21 portrayed vocal arousal were judged as more negative and more emotionally aroused than the
22 same voices produced with low vocal arousal. Moreover, both the perceived valence and
23 emotional arousal of vocalizations were significantly affected by both musical and non-musical
24 contexts. These results show the importance of auditory context in judging emotional arousal and
25 valence in voices and music, and suggest that nonlinear features in music are processed similarly
26 to communicative vocal signals.

27

28 Keywords

29 Vocal arousal, vocalization, sound context, emotional judgments

30 1. Background

31 When animals are highly aroused there can be many effects on their bodies and
32 behaviors. One important behavioral consequence of physiological arousal is the introduction of
33 nonlinear features in the structure of vocalizations (Briefer, 2012; Fitch, Neubauer, & Herzel,
34 2002; Wilden et al., 1998). These acoustic correlates of arousal include deterministic chaos,
35 subharmonics, and other non-tonal characteristics that can give vocalizations a rough, noisy
36 sound quality. Nonlinear phenomena are effective in communicative signals because they are
37 difficult to habituate to (Blumstein & Recapet, 2009), and they successfully penetrate noisy
38 environments (Arnal et al. 2015).

39 Alarm calls and threat displays in many species often have such nonlinear features. In
40 humans, alarm signaling manifests as screaming. Screams are characterized by particularly high
41 signal-to-noise ratios at the most sensitive frequencies of human hearing (Begault, 2008). Recent
42 psychophysical and imaging studies suggest that vocal sounds containing low-level spectro-
43 temporal modulation features (i.e., modulation rate of 30-150 Hz) are perceived as rough,
44 activate threat-response circuits such as the amygdala, and are more efficiently spatially
45 localized (Arnal et al., 2015). But what is the effect of sound context on the perception of vocal
46 screams? Many variable internal and external factors can potentially affect subjective judgments
47 of a stimulus, and cues from multiple modalities can give contradictory information (Abitol, et
48 al., 2015; Driver & Noesselt, 2008). In humans, high-level cognitive processes such as the
49 emotional or semantic evaluation of a context can change the way signals are integrated across
50 sensory areas (Blumstein et al., 2012; Mothes-Lasch et al., 2012). For example, musical pieces
51 with added distortion are judged differently on the two emotional dimensions of valence and
52 arousal (termed here *emotional* arousal) from the same pieces without the noise, but the
53 difference in emotional arousal disappeared when presented in a benign visual context (i.e.,
54 video with very little action; Blumstein et al., 2012).

55 These results suggest that visual information can reduce the arousing impact of nonlinear
56 acoustic features, potentially dampening fear reactions. But very little is known about the effect
57 of sound context on auditory perception: such context effects may be particularly strong in music
58 where indicators of vocal arousal likely take a different semantic or stylistic meaning than in
59 more ecologically valid situations (Neubauer et al., 2004). Belin and Zatorre (2015) described
60 the example of enjoying certain screams in operatic singing. In a more formal experimental
61 setting, fans of “death metal” music reported experiencing a wide range of positive emotions
62 including power, joy, peace, and wonder, despite the aggressive sound textures associated with
63 this genre (Thompson, Geeves & Olsen, 2018). Other examples include the growl singing style
64 of classic rock and jazz singers, or carnival lead voices in samba singing (Sakakibara et al.
65 2004).

66 Here we examined the relationship between manipulated cues to vocal arousal, perceived
67 emotional valence and arousal, and the sonic context in which the vocal signal occurs. We
68 created two musical contexts (clean guitar, distorted guitar) and a non-musical context
69 (modulated noise), in which we integrated nonverbal vocalizations (specifically, screams of the
70 vowel /a/) at varying levels of portrayed vocal arousal. Based on the work described above, we

71 expected that aroused voices in a musically congruent context (distorted guitar) would be judged
72 less emotionally negative and less emotionally arousing than voices presented alone, or
73 presented in musically incongruent contexts (clean guitar, noise).

74 **Methods**

75 *Participants*

76 23 young adults (12 women, mean age = 20.8, $SD = 1.3$; 11 men, mean age = 25.1 $SD =$
77 3.2) with self-reported normal hearing participated in the experiment. All participants provided
78 informed written consent prior to the experiment and were paid 15€ for their participation.
79 Participants were tested at the Centre Multidisciplinaire des Sciences Comportementales
80 Sorbonne Université-Institut Européen d'Administration des Affaires (INSEAD), and the
81 protocol of this experiment was approved by the INSEAD Institutional Review Board.

82 *Stimuli*

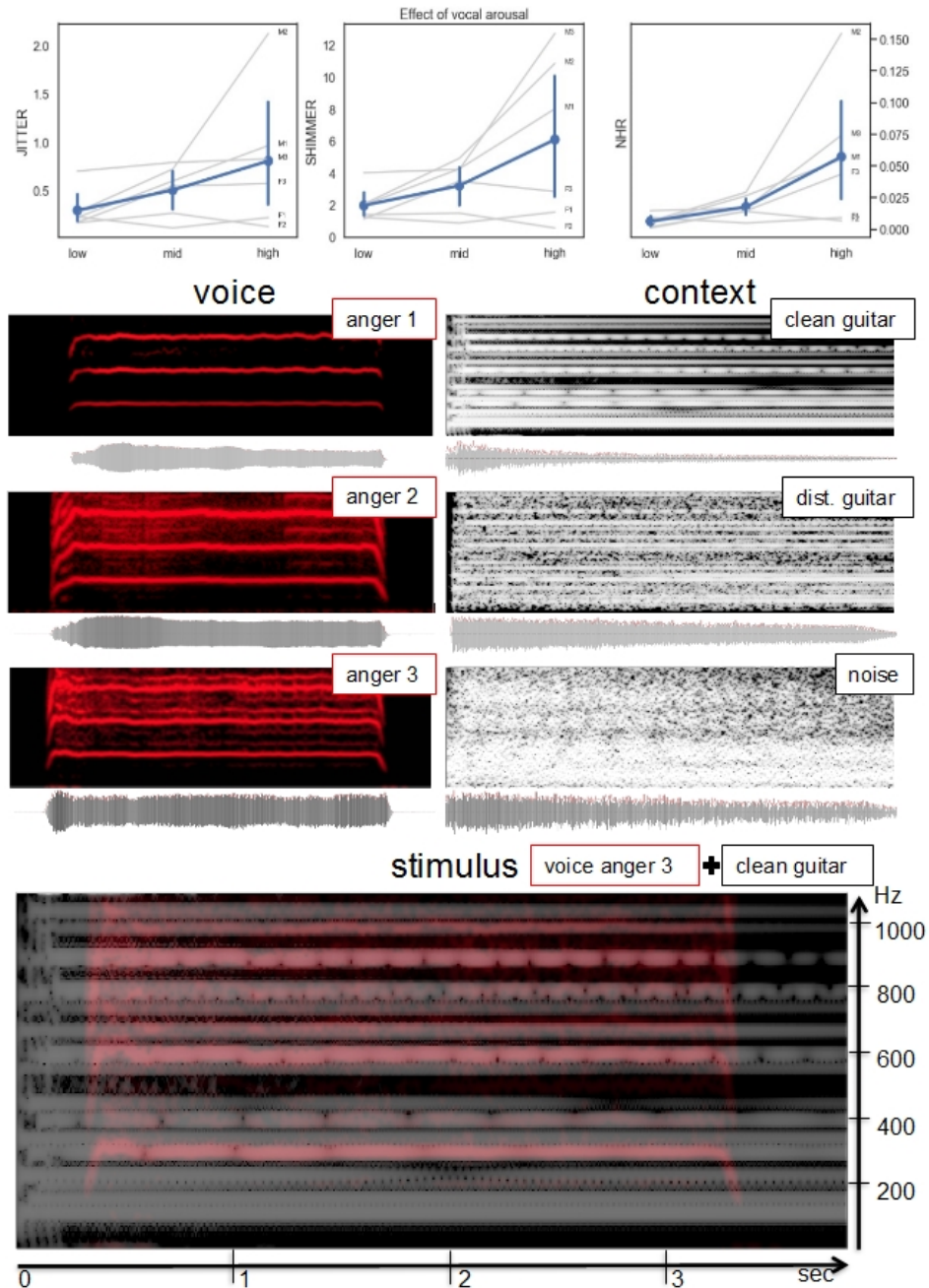
83 Two singers, one man and one woman, each recorded nine utterances of the vowel /a/
84 (normalized in duration, ~1.6 s.), at three different target pitches (A3, C4 and D4 for the man,
85 one octave higher for the woman), with three target levels of portrayed anger / vocal arousal
86 (Figure 1 - *voice* column). The files were recorded in mono at 44.1 kHz/16-bit resolution with
87 the Max software (Cycling '74, version 7).

88 A professional musician recorded nine chords with an electric guitar, on separate tracks
89 (normalized in duration, ~2.5 s.). All the chords were produced as permutations of the same
90 pattern (tonic, perfect fourth, perfect fifth), harmonically consistent with the above vocal pitches.
91 The nine guitar tracks were then distorted using a commercial distortion plugin (Multi by Sonic
92 Academy) with a unique custom preset (.fxb file provided as supporting file, see the section
93 *Data accessibility*) to create nine distorted versions of these chords. Lastly, we generated nine
94 guitar-shaped noise samples by filtering white noise with the spectral envelope estimated frame-
95 by-frame (frame size = 25 ms) on the distorted guitar chords (see Figure 1- *context* column).

96 Both vocalizations and backgrounds (noise, clean guitar, distorted guitar) were
97 normalized in loudness using the maximum value of the long-term loudness given by the
98 Glasberg and Moore's model for time-varying sounds (Glasberg & Moore, 2002), implemented
99 in Matlab in the Pysound3 toolbox. The vocalizations were normalized to 16 sones for the first
100 level of anger, 16.8 (=16*1.05) sones for the second level, and to 17.6 (=16*1.05*1.05) for the
101 third and highest level. All the backgrounds were set to 14 sones. For a more intuitive analysis of
102 the sound levels reached after this normalization, we evaluated the level of the vocalizations and
103 backgrounds composing our stimuli a posteriori. The vocalizations reached very similar levels
104 (with differences < 2 dB among the different levels of anger), as did the different backgrounds
105 (differences < 1.5 dB). Therefore, the overall signal-to-noise ratio (SNR) between voice and
106 backgrounds was similar across the three different contexts (SNR=8.4 dB \pm 1.6).

107 The 18 vocalizations and 4 backgrounds (no context, noise, clean guitar, distortion guitar)
108 were then superimposed to create 72 different stimuli (onset of the vocalization = 30 ms. after
109 the onset of the context, See Figure 1- bottom). While we were primarily interested in context

110 effects, the rationale for varying both cues to vocal arousal and stimulus contexts was to avoid
 111 experimental demand effects induced by situations where context is the only manipulated
 112 variable.



113 **Figure 1:** Top: Acoustical analysis (jitter_loc, vshimmer_loc and noise-harmonic ratio) of the
 114 vocalizations, as a function of level of portrayed vocal arousal. Middle: spectrograms of
 115 selected original recordings. Left column shows a male vocalization, of the note D4 at 3 anger
 116 levels. Right column shows a D-G-A guitar chord in the different sound contexts. Bottom: one
 117 example of the resulting mix.

118 *Procedure*

119 On each trial, participants listened to one stimulus (vocalization + background) and
120 evaluated its perceived emotional arousal and valence on two continuous sliders positioned
121 below their respective self-assessment manikin (SAM) scales (Bradley & Lang, 1994): a series
122 of pictures varying in both affective valence and intensity, that serve as a nonverbal pictorial
123 assessment technique directly measuring a person's affective reaction to a stimulus. Participants
124 were instructed to evaluate the emotional state of the speaker, ignoring background (noise or
125 guitar). A first training block was presented, with 20 trials composed of vocalizations with no
126 background (randomly selected from the subsequent set of stimuli). After this phase, listeners
127 received a score out of 100 (actually a random number between 70 and 90) and were asked to
128 maintain this performance in subsequent trials, despite the sounds being thereafter embedded in
129 background noise. Each of the 18 vocalizations (2 speakers x 3 pitches x 3 anger levels) was then
130 presented five times in four different contexts (none, clean guitar, distortion guitar, noise), with a
131 1s inter-trial interval. The main experiment included 360 randomized trials divided into 6 blocks.
132 To motivate continued effort, participants were asked to maximize accuracy during the practice
133 phase, and at the end of each block they received fake feedback scores, on average slightly below
134 that of the training phase (a random number between 60 and 80). Participants were informed that
135 they could receive a financial bonus if their accuracy was above a certain threshold (all
136 participants received the bonus regardless of their score). The experiment was run in a single
137 session lasting 75 minutes. Sound files were presented through headphones (Beyerdynamic DT
138 270 PRO, 80 ohms), with a fixed sound-level that was judged to be comfortable by all the
139 participants.

140 *Statistical analyses*

141 We used the continuous slider values corresponding to the two SAM scales (coded from
142 0 to 100 between their two extremes) to extract valence and emotional arousal ratings. We
143 analyzed the effect of context on these ratings with two repeated-measures ANOVAs, conducted
144 separately on negativity (100-valence) and emotional arousal, using level of portrayed vocal
145 arousal in the voice (3 levels) and context (4 levels: Voice-alone, clean-guitar, distorted guitar,
146 noise) as independent variables. Post-hoc tests were Tukey HSD. All statistical analyses were
147 conducted using R (R Core Team, 2013). Huynh-Feldt corrections ($\tilde{\epsilon}$) for degrees of freedom
148 were used where appropriate. Effect sizes are reported using partial eta-squared η_p^2 .

149 **2. Results**

150 To control that stimuli with increasing levels of portrayed vocal arousal indeed had more
151 nonlinearities, we subjected the 18 vocal stimuli to acoustic analysis with the Praat software
152 (Boersma, 2011), using three measures of voice quality (jitter, shimmer, noise-harmonic ratio)
153 commonly associated with auditory roughness and noise. All three measures scaled as predicted
154 with portrayed vocal arousal (Figure 1-top).

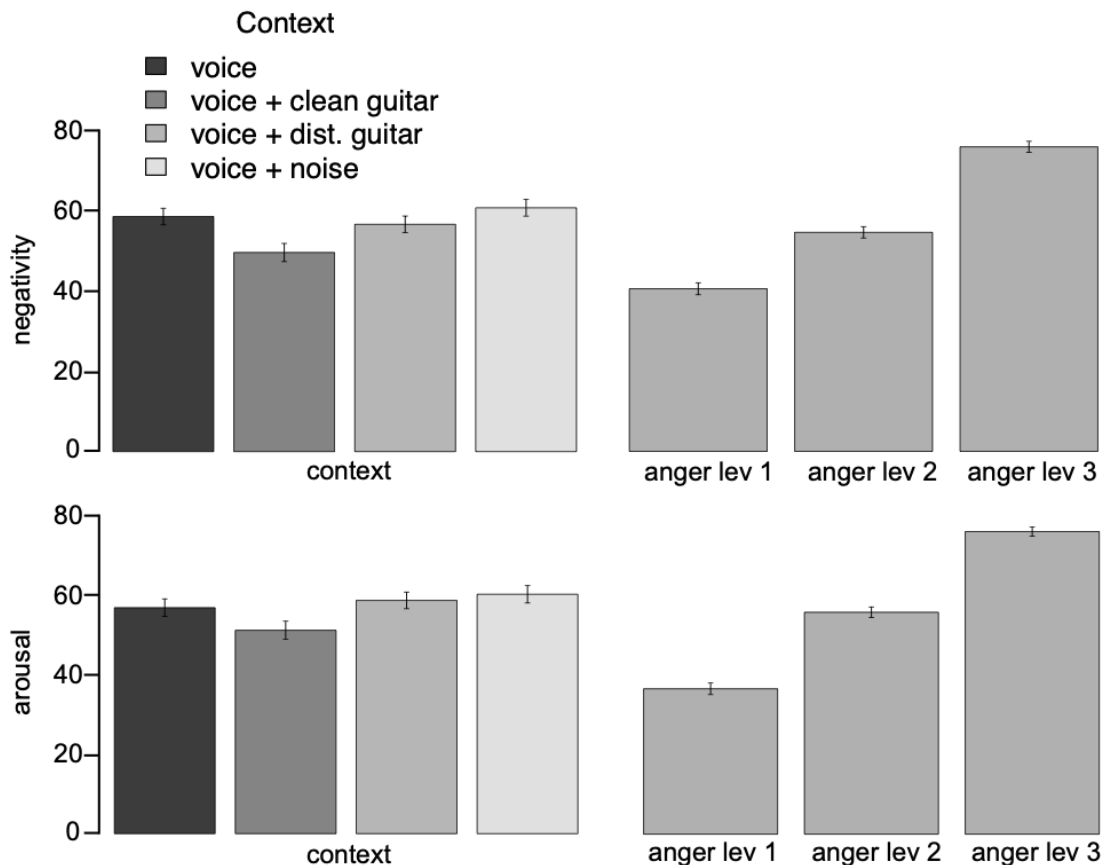
155 Repeated-measures ANOVAs revealed a main effect of portrayed vocal arousal on
156 judgments of emotional valence, $F(2,44) = 97.2, p < .0001, \eta_p^2 = 0.82, \tilde{\epsilon} = 0.61$, with greater
157 levels of speaker anger associated with more negative valence; a main effect of context $F(3,66) =$

158 39.5, $p < .0001$, $\eta_p^2 = 0.64$, $\tilde{\epsilon} = 0.79$, with sounds presented in distorted guitar judged more
159 negative than in clean guitar, and sounds presented in noise judged more negative than in
160 distorted guitar (Figure 2-top, left). Ratings of stimuli presented alone or in distorted guitar did
161 not differ (Tukey HSD). Context did not interact with the effect of anger on perceived valence
162 (Figure 2-top, right).

163 A similar pattern of results was found for judgments of emotional arousal, with a main
164 effect of portrayed vocal arousal, $F(2,44) = 118.2$, $p < .0001$, $\eta_p^2 = 0.84$, $\tilde{\epsilon} = 0.62$, increasing
165 perceived emotional arousal, and a main effect of context $F(3,66) = 17.3$, $p < .0001$, $\eta_p^2 = 0.84$, $\tilde{\epsilon}$
166 = 0.44. Sounds presented with noise and distorted guitar were judged as more emotionally
167 aroused than with clean guitar (Figure 2-bottom, left). Ratings in noise and distorted guitar
168 contexts did not differ (Tukey HSD) and, as above, context did not interact with the effect of
169 anger on perceived emotional arousal (Figure 2-bottom, right).

170

171



172 **Figure 2:** Perceived valence (negativity) and emotional arousal as a function of three levels of
173 portrayed vocal arousal of human vocalizations presented in different sound contexts. Error-
174 bars show 95% confidence intervals on the mean.

175

176 3. **Discussion**

177 As expected, voices with higher levels of portrayed anger were judged as more negative
178 and more emotionally aroused than the same voices produced with less vocal arousal. Both the
179 perceived valence and emotional arousal of voices with high vocal arousal were significantly
180 affected by both musical and non-musical contexts. However, contrary to what would be
181 predicted e.g. by the aesthetic enjoyment of nonlinear vocal sounds in rough musical textures by
182 death metal fans (Thompson, Geeves & Olsen, 2018) or more generally, the suggestion of the
183 top-down deactivation of the effect of nonlinear features in appropriate musical contexts (Zatorre
184 & Belin, 2015), we did not find any incongruency effect between musical contexts and vocal
185 signals, and in particular did not find support for the notion that the presence of a distorted guitar
186 would reduce emotional effects of nonlinearities in voices. Instead, screams with high levels of
187 portrayed vocal arousal were perceived as more negative and more emotionally aroused in the
188 context of background nonlinearities. This suggests that local judgments such as the emotional
189 appraisal of one isolated part of an auditory scene (e.g., a voice) are computed on the basis of the
190 global acoustic features of the scene. It is possible that this effect results from a high-level
191 integration of cues similar to the emotional evaluation of audiovisual signals (de Gelder &
192 Vroomen, 2000), in which both signal and context are treated as coherent communicative signals
193 reinforcing each other. Another possibility is that stream segregation processes cause the
194 nonlinear features of the vocal stream to perceptively fuse with that of the other (i.e., musical or
195 noise) stream (see e.g. Vannier et al., 2018).

196 These results raise several questions for future research. First, to understand the neural
197 time-course of these contextual effects as well as disentangle the respective contributions of
198 subcortical (e.g. amygdala; Arnal et al., 2015) and high-level decision processes, the current
199 study could be complemented with fMRI imaging and MEG analysis. Second, future research
200 could explore psychoacoustical sensitivity to vocal roughness, and how discrimination thresholds
201 might shift as a function of low-level properties of the acoustic context. Additionally, error
202 management principles could be at play as listeners might be biased to over-detect nonlinear-
203 based roughness in cases of possible danger (Blumstein, Whitaker, Kennen, & Bryant, 2017).
204 Finally, research could examine whether context effects exist with signals that are not
205 conspecific vocalizations, such as other animals' vocalizations/calls or synthetic alarm signals.

206 More generally, these findings suggest that musical features are processed similarly to
207 vocal sounds, and that affective information in music is treated as a communicative signal
208 (Bryant, 2013). Instrumental music constitutes a recent cultural innovation that often explicitly
209 imitates properties of the voice (Juslin & Laukka, 2003). The cultural evolution of music
210 technology and sound generation exploits perceptual mechanisms designed to process
211 communicative information in voices. These results provide an excellent example of how an
212 ecologically-based theoretical framework can be used to understand what might otherwise
213 appear to be novel features in contemporary cultural phenomena such as vocal and instrumental
214 music.

215

216 **Acknowledgements**

217 The authors thank Hugo Trad for help running the experiment. All data collected at the
218 Sorbonne-Université INSEAD Center for Behavioural Sciences.

219 **Author contributions**

220 ML and EP made equal contributions for conceiving the study, recording and designing stimuli,
221 running the experiment, analyzing data, and drafting the manuscript. GB contributed to the
222 interpretation of the data analysis and formalization of the results. JJA contributed to the
223 conception of the experimental design. All authors contributed to the writing of the manuscript.

224 **Data accessibility**

225 Matlab files and stimuli to run the experiment, R files and python notebook to analyze the
226 results, as well as a .fxb file for the guitar distortion plugin are available at the following URL:

227 <https://nubo.ircam.fr/index.php/s/QAMG78HPymso26o>

228 **Funding**

229 This study was supported by ERC Grant StG 335536 CREAM to JJA, and by a Fulbright
230 Visiting Scholar Fellowship to ML.

231

232 **References**

- 233 Abitbol, R., Lebreton, M., Hollard, G., Richmond, B. J., Bouret, S., and Pessiglione, M. (2015).
234 Neural mechanisms underlying contextual dependency of subjective values: Converging
235 evidence from monkeys and humans. *The Journal of Neuroscience*, 35(5), 2308–2320.
- 236 Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., and Poeppel, D. (2015). Human
237 screams occupy a privileged niche in the communication soundscape. *Current Biology*, 25(15),
238 2051-2056.
- 239 Begault, D. R. (2008). Forensic analysis of the audibility of female screams. In *Audio*
240 *Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and*
241 *Practice*. Audio Engineering Society.
- 242 Belin, P. and Zatorre, R. J. (2015). Neurobiology: Sounding the alarm. *Current Biology*, 25(18),
243 R805–R806.
- 244 Blumstein, D. T., Bryant, G. A., and Kaye, P. (2012). The sound of arousal in music is context-
245 dependent. *Biology Letters*, 8(5), 744–747.
- 246 Blumstein, D.T. & Recapet, C. (2009). The sound of arousal: The addition of novel non-
247 linearities increases responsiveness in marmot alarm calls. *Ethology*, 115(11), 1074-1081.
- 248 Blumstein, D. T., Whitaker, J., Kennen, J., & Bryant, G. A. (2017). Do birds differentiate
249 between white noise and deterministic chaos? *Ethology*, 123(12), 966-973.
- 250 Boersma, P. (2011). Praat: doing phonetics by computer [Computer program]. <http://www.praat.org/>.
251
- 252 Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the
253 semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1), 49-59.
- 254 Briefer, E.F., (2012). Vocal expression of emotions in mammals: mechanisms of production and
255 evidence. *Journal of Zoology*, 288, 1–20.
- 256 Bryant, G. A. (2013). Animal signals and emotion in music: Coordinating affect across groups.
257 *Frontiers in Psychology*, 4, 990.
- 258 Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on
259 ‘sensory-specific’ brain regions, neural responses, and judgments. *Neuron*, 57(1), 11-23.
- 260 Fitch, W. T., Neubauer, J., and Herzel, H. (2002). Calls out of chaos: the adaptive significance of
261 nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, 63(3):407–418.
- 262 Glasberg, B. R., & Moore, B. C. (2002). A model of loudness applicable to time-varying sounds.
263 *Journal of the Audio Engineering Society*, 50(5), 331-342.
- 264 Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music
265 performance: Different channels, same code?. *Psychological bulletin*, 129(5), 770.

- 266 Mothes-Lasch, M., Miltner, W. H., and Straube, T. (2012). Processing of angry voices is
267 modulated by visual load. *Neuroimage*, 63(1), 485–490.
- 268 Neubauer, J., Edgerton, M., Herzel, H., (2004). Nonlinear phenomena in contemporary vocal
269 music. *Journal of Voice*, 18, 1–12.
- 270 Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The
271 case of small sample sizes and nonnormal data. *Behavior Research Methods*, 45(3), 792-812).
- 272 Sakakibara, K. I., Fuks, L., Imagawa, H., & Tayama, N. (2004). Growl voice in ethnic and pop
273 styles. *Leonardo*, 1, 2.
- 274 Thompson, W. F., Geeves, A. M., & Olsen, K. N. (2018). Who enjoys listening to violent music
275 and why. *Psychology of Popular Media Culture*.
- 276 Vannier, M., Misdariis, N., Susini, P., & Grimault, N. (2018). How does the perceptual
277 organization of a multi-tone mixture interact with partial and global loudness judgments? *The*
278 *Journal of the Acoustical Society of America*, 143(1), 575-593.
- 279 Wilden, I., Herzel, H., Peters, G., Tembrock, G., (1998). Subharmonics, biphonation, and
280 deterministic chaos in mammal vocalization. *Bioacoustics*, 9, 171–196.