*Original Article*

QJEP

Quarterly Journal of Experimental Psychology
2018, Vol. 71(3) 622–641

DOI: 10.1080/17470218.2016.1270976
qjep.sagepub.com
SAGE

# Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations

**Andrey Anikin[1] and César F. Lima[2,3,4]**

**Abstract**

Most research on nonverbal emotional vocalizations is based on actor portrayals, but how similar are they to the vocalizations produced spontaneously in everyday life? Perceptual and acoustic differences have been discovered between spontaneous and volitional laughs, but little is known about other emotions. We compared 362 acted vocalizations from seven corpora with 427 authentic vocalizations using acoustic analysis, and 278 vocalizations (139 authentic and 139 acted) were also tested in a forced-choice authenticity detection task ($N = 154$ listeners). Target emotions were: achievement, amusement, anger, disgust, fear, pain, pleasure, and sadness. Listeners distinguished between authentic and acted vocalizations with accuracy levels above chance across all emotions (overall accuracy 65%). Accuracy was highest for vocalizations of achievement, anger, fear, and pleasure, which also displayed the largest differences in acoustic characteristics. In contrast, both perceptual and acoustic differences between authentic and acted vocalizations of amusement, disgust, and sadness were relatively small. Acoustic predictors of authenticity included higher and more variable pitch, lower harmonicity, and less regular temporal structure. The existence of perceptual and acoustic differences between authentic and acted vocalizations for all analysed emotions suggests that it may be useful to include spontaneous expressions in datasets for psychological research and affective computing.

**Keywords**

Acoustic analysis; Actor portrayals; Authenticity; Emotion; Nonverbal vocalizations

Received: 20 July 2016; accepted: 26 November 2016

A researcher studying emotional expressions has three potential sources of stimuli (Scherer & Bänziger, 2010). The first and most common approach is to ask professional actors or amateurs to portray an emotion, often aided by a short vignette describing the context. The second option is to induce an emotional state in participants—for example, by showing them emotionally charged video clips or by asking them to relive a powerful personal memory. The third option is to record spontaneous expressions of emotion through field observation. Traditionally, the last approach has been under-utilized because it is time consuming and methodologically challenging (Douglas-Cowie, Campbell, Cowie, & Roach, 2003). However, the modern ubiquity of digital technologies and social media provides researchers with access to audio and video recordings of people engaged in dramatic and highly emotional activities, which would otherwise be difficult to obtain. Researchers are beginning to tap into this new source of data (Anikin & Persson, 2016; Dai, Han, Dai, & Xu, 2015; Parsons, Young, Stein, Craske, & Kringelbach, 2014), but little is known about perceptual and acoustic differences between observational material and the actor portrayals that dominate emotion research. In the current study, we compared a recently validated large corpus of authentic nonverbal vocalizations (Anikin & Persson, 2016) with acted vocalizations taken from seven published corpora.

What makes it desirable to extend emotion research beyond acted (posed) portrayals of emotion? Acted portrayals are intended to be easily recognized, and the most

[1]Division of Cognitive Science, Department of Philosophy, Lund University, Lund, Sweden
[2]Institute of Cognitive Neuroscience, University College London, London, UK
[3]Center for Psychology, University of Porto, Porto, Portugal
[4]Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

**Corresponding author:**
Andrey Anikin, Division of Cognitive Science, Department of Philosophy, Lund University, Box 192, Lund SE-221 00, Sweden.
Email: andrey.anikin@lucs.lu.se

accurately recognized tokens can be assumed to represent the conventional cultural code for a given expression (Krumhuber, Kappas, & Manstead, 2013; Scherer & Bänziger, 2010). However, more spontaneous displays of emotion also pervade everyday social interactions, and the ability to discriminate between "real" and "fake" emotions is an important social skill (Gervais & Wilson, 2005). From an evolutionary perspective, intraspecific communication presupposes the existence of honest, "hard-to-fake" signals that are reliably associated with particular emotional states (Searcy & Nowicki, 2005). For example, authentic laughter is an indicator of genuinely friendly intentions (Gervais & Wilson, 2005), but to be reliable, such honest signals must be distinct from potentially deceitful imitations. There is some recent evidence that listeners can indeed make such discriminations (Bryant & Aktipis, 2014; Lavan, Scott, & McGettigan, 2015), raising the question of what acoustic differences guide authenticity detection.

Systematic attempts to examine aspects of vocal emotional processing beyond acted vocal expressions have been relatively rare (Batliner, Fischer, Huber, Spilker, & Nöth, 2000; Douglas-Cowie et al., 2003; Gendron, Roberson, van der Vyver, & Barrett, 2014; Parsons et al., 2014). As for emotional speech processing (emotional prosody), while some studies found that listeners were unable to reliably judge authenticity (Jürgens, Drolet, Pirow, Scheiner, & Fischer, 2013; R. Jürgens, Grass, Drolet, & Fischer, 2015; Scherer, 2013), other studies have reported accurate authenticity detection (Drolet, Schubotz, & Fischer, 2012). In any case, authenticity of emotional speech should not be conflated with authenticity of nonverbal vocalizations, since verbal and nonverbal vocalizations involve partly distinct neural circuitry (Ackermann, Hage, & Ziegler, 2014; U. Jürgens, 2009; Scott, Sauter, & McGettigan, 2009).

The only nonverbal vocalization that has already become the object of authenticity research is laughter. For example, authentic and acted laughs could be correctly classified 67% of the time (against a chance level of 50%) in a study by Bryant and Aktipis (2014) and about 72% of the time in a study by Lavan et al. (2015). In the latter study, though, the stimuli were preselected for optimal authenticity detection from an initial corpus larger than the final one, possibly inflating the detection rate. Spontaneous laughs also activate different brain regions compared to volitional laughs (McGettigan et al., 2015; Scott, Lavan, Chen, & McGettigan, 2014; Wattendorf et al., 2013). Volitional laughs activate the anterior medial prefrontal cortex and the anterior cingulate gyrus more strongly than do spontaneous laughs, suggesting a greater engagement of mentalizing processes when laughter is less genuine. These sounds may therefore be perceived as more ambiguous and in need of active interpretation, whereas authentic laughs are processed more automatically. Consistent with this, spontaneous laughs activate auditory areas in the superior temporal gyrus more strongly than volitional laughs (McGettigan et al., 2015). Similarly, more activation in brain areas involved in mentalizing has been reported when processing acted as opposed to authentic emotional speech prosody (Drolet et al., 2012).

An important question is which acoustic features correlate with perceived authenticity. Bryant and Aktipis (2014) report that spontaneous laughs contain shorter syllables with relatively longer unvoiced breaks. They argue that the rate of five syllables per second, which is typical of spontaneous laughs, represents the highest possible oscillation rate of the intrinsic laryngeal muscles, making it a distinct and presumably hard-to-fake acoustic signature of genuine mirth. Interestingly, all laughs sounded more authentic when the recordings were sped up without modifying their pitch. The importance of the temporal characteristics of laughter is corroborated by Kipper and Todt (2001), who report a similar rate of five syllables per second in natural laughs. On the other hand, Bachorowski, Smoski, and Owren (2001) recorded students laughing at a comedy film and reported a slightly lower rate of 4.37 syllables per second. They further observed that natural laughs were extremely variable regarding both their temporal and spectral profiles. Kipper and Todt (2001) also concluded that laughs with more variable rhythm and pitch within one bout were judged as more natural than more stereotyped laughs. Comparing several studies, Vettin and Todt (2005) concluded that laughs produced in response to a funny episode, as opposed to social polite laughter, contained more syllables with a shorter inter-syllable interval and had a higher fundamental frequency. Finally, in a detailed acoustic comparison of spontaneous and volitional laughter, Lavan et al. (2015) found that spontaneous laughs had longer bouts, shorter syllables, higher and more variable fundamental frequency, more unvoiced frames, and lower mean intensity. Altogether, these studies suggest that multiple acoustic parameters might be involved in communicating laughter authenticity.

To our knowledge, no studies so far have addressed the authenticity of nonverbal vocalizations other than laughter. It remains unclear whether listeners can reliably judge whether other vocalizations are real or posed, and whether the acoustic markers of authenticity are similar or distinct across vocalizations. The current study compares for the first time a wide range of positive and negative authentic and acted vocalizations. Another novel aspect of our approach is the use of naturalistic authentic vocalizations taken from everyday emotional episodes. The studies of laughter reviewed above fall into Scherer and Bänziger's (2010) second methodological category—induced emotion. Spontaneous laughs, for example, have been evoked by showing participants amusing video clips (Bachorowski et al., 2001; Lavan et al., 2015; McGettigan et al., 2015; McKeown, Sneddon, & Curran, 2015). While this approach

has been shown to successfully produce authentic expressions, it is difficult to recreate in a laboratory setting the diversity of emotional elicitors typically encountered by people in everyday life. Crucially, due to ethical constraints, it is problematic to experimentally induce strong emotions like fear or anger to the point of making participants vocalize (Scherer & Bänziger, 2010).

Observational material offers a unique opportunity to transcend these limitations, and the Internet offers a promising alternative that is now being harnessed for research on vocal emotions. For instance, Parsons et al. (2014) introduced a corpus of authentic vocalizations intended for psychological testing, which includes laughter, crying, and neutral sounds obtained from amateur videos shared online (www.youtube.com). Using the same source, Anikin and Persson (2016) compiled and validated a broader corpus of authentic vocalizations that includes nine emotions. The emotion was inferred based on contextual cues, such as facial expression, verbal information, and the activity engaged in, such as: laughing because a friend took a tumble (amusement), roaring in frustration upon having lost a computer game (anger), cleaning a clogged toilet (disgust), being the victim of a scare prank (fear), suffering a sports injury (pain), having sex (pleasure), or crying with tears about someone's death (sadness). A similar approach of inferring the experienced emotion from contextual information is commonly adopted in emotional speech research. For example, Drolet et al. (2012) and Jürgens et al. (2013) obtained samples of authentic emotional speech from recordings of radio programmes and relied on both verbal and contextual cues to classify the underlying emotion of the speaker. To address concerns about the subjectivity of inferring the vocalizer's state of mind and the risk of them acting "for the camera" (Douglas-Cowie et al., 2003; Scherer, 2013), Anikin and Persson (2016) prioritized situations in which the vocalizer was unaware of being filmed, or the triggering event was sudden and intense, leaving little time for impression management. By including only the cleanest recordings in the corpus and performing manual filtering, the authors also ensured that sound quality was suitable for acoustic analysis. In fact, acoustic models in the validation study classified these vocalizations by emotion as accurately as did human listeners, with comparable confusion patterns (Anikin & Persson, 2016). Furthermore, since for each sound both the original context and the emotion perceived by listeners are reported, we were able to select the most unambiguous, best recognized tokens of each emotion from the corpus for the authenticity recognition task used here.

In sum, the current study takes advantage of comparing a wide range of authentic vocalizations taken from naturalistic settings with their acted counterparts to shed new light on the acoustic and perceptual basis of authenticity processing in vocal emotions. Participants performed a forced-choice authenticity detection task across 139 authentic and 139 acted vocalizations (96 positive and 182 negative). Acoustic analyses were conducted on all available vocalizations (427 authentic and 362 acted) to determine the correlates of actual (objective) as well as perceived (subjective) authenticity.

## Experimental study

### Method

*Stimuli.* Actor portrayals of emotional nonverbal vocalizations were taken from seven published corpora: Belin, Fillion-Bilodeau, and Gosselin (2008); Cordaro, Keltner, Tshering, Wangchuk, and Flynn (2016); Hawk, Van Kleef, Fischer, and Van der Schalk (2009); Lima, Castro, and Scott (2013); Maurage, Joassin, Philippot, and Campanella (2007); Sauter, Eisner, Calder, and Scott (2010); and Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, and Abramson (2009). The vocalizations from Sauter et al. (2010; $n = 70$) were only analysed acoustically and were not included in the behavioural experiment, as agreed with the author. We considered only the emotion categories for which authentic equivalents were available in the corpus by Anikin and Persson (2016): achievement, amusement, anger, disgust, fear, pain, pleasure, and sadness (Table 1).

All these sounds are nonverbal—that is, they contain no words and only a few semi-articulated interjections. In two corpora, by Belin et al. (2008) and Maurage et al. (2007), the speakers were instructed to hold a single vowel (the French *ah*). The original audio was used without modification, except that: (a) sounds were normalized for peak amplitude; (b) microphone hiss in the corpus by Simon-Thomas et al. (2009) was removed using the software Audacity (http://audacity.sourceforge.net); and (c) sounds exceeding 4 s in duration in the corpora by Hawk et al. (2009), Cordaro et al. (2016), and Anikin and Persson (2016) were shortened to approximately 4 s. These modifications were intended to make all corpora comparable in terms of the duration, loudness, and recording quality of sounds. Nevertheless, some differences among the selected 278 sounds remained. Notably, those from Hawk et al. (2009) had a significant amount of clipping, and authentic sounds had a longer average duration (2.1 s ± 1.5) than acted sounds (1.5 s ± 1.1), $t(351.1) = 3.7$, $p < .001$.

Since the corpora of acted vocalizations differed in the included emotion categories and contained more sounds than there were suitable authentic vocalizations for comparison, we selected a subset of stimuli from each corpus, ensuring that we: (a) had a comparable number of stimuli for each emotion, on average 17 authentic and 17 acted; (b) selected sounds with the highest recognition rate (these data were not available for the corpus by Belin et al., 2008; average emotion scores, rather than hit rates, were reported in Maurage et al., 2007); (c) avoided stimuli with high levels of background noise or clipping; (d) had a similar

**Table 1.** Sources of the vocalizations used in the behavioural experiment.

| Emotion category | Actor portrayals, M/F | | | | | | | Authentic (Anikin & Persson, 2016) | Actors + authentic |
|---|---|---|---|---|---|---|---|---|---|
| | Belin et al. (2008) | Cordaro et al. (2016) | Hawk et al. (2009) | Lima et al. (2013) | Maurage et al. (2007) | Simon-Thomas et al. (2009) | Total (M/F) | | |
| Achievement | — | 1/1 | — | 2/2 | — | 2/2 | 10 (5/5) | 10 (5/5) | 20 (10/10) |
| Amusement | 2/2 | 1/1 | 2/2 | 4/4 | — | 2/2 | 22 (11/11) | 22 (11/11) | 44 (22/22) |
| Anger | 1/1 | 1/1 | 1/1 | 2/2 | 1/3 | 2/2 | 18 (8/10) | 18 (13/5) | 36 (15/21) |
| Disgust | 1/1 | 1/1 | 2/2 | 2/2 | 0/4 | 2/2 | 20 (8/12) | 20 (10/10) | 40 (22/18) |
| Fear | 2/2 | 1/1 | — | 2/2 | 4/2 | 2/2 | 20 (11/9) | 20 (6/14) | 40 (23/17) |
| Pain | 5/5 | 1/1 | — | — | — | — | 12 (6/6) | 12 (9/3) | 24 (9/15) |
| Pleasure | 2/2 | — | — | 4/4 | — | 2/2 | 16 (8/8) | 16 (7/9) | 32 (17/15) |
| Sadness | 2/2 | 1/1 | 2/2 | 2/2 | 2/2 | 1/2 | 21 (10/11) | 21 (10/11) | 42 (22/20) |
| Total (M/F) | 30 (15/15) | 14 (7/7) | 14 (7/7) | 36 (18/18) | 18 (7/11) | 27 (13/14) | 139 (67/72) | 139 (71/68) | 278 (138/140) |
| Accuracy, %[a] | 68[c] | 71.1 | 94.8 | 92.5 | —[d] | 79.9 | | 64.3 | |
| Proportion index, %[b] | 94.4 | 94.3 | 99.3 | 98.7 | — | 96.8 | | 90.5 | |
| Language | French (Canada) | English (USA) | Dutch (Holland) | Portuguese (Portugal) | French (Belgium) | English (USA) | | English (mixed) | |
| Professional actors | No | No | Yes | No | No | No | | No | |

Note: $N = 278$. Values indicate the number of male/female vocalizations per corpus and per emotion. M = male; F = female. Belin et al. (2008): Sounds labelled *happiness* in the original corpus were used as *amusement* (available from: http://vnl.psy.gla.ac.uk/sounds/Montreal_Affective_Voices.zip). Cordaro et al. (2016): Sounds labelled *triumph* in the original corpus were used as *achievement* (available from: http://socrates.berkeley.edu/~keltner/resources.htm). Hawk et al. (2009): Sounds labelled *joy* in the original corpus were used as *amusement* (kindly provided by S. T. Hawk). Lima et al. (2013): (available from: https://protect-us.mimecast.com/s/3RmXBohVZ9x7Fa?domain=link.springer.com). Maurage et al. (2007): (available from: http://www.ledonline.it/NeuropsychologicalTrends/allegati/NeuropsychologicalTrends_2_Maurage.zip). Simon-Thomas et al. (2009): (kindly provided by E. R. Simon-Thomas). Anikin and Persson (2016): (available from: http://cogsci.se/publications.html).

[a]Accuracy of emotion recognition (unadjusted hit rates) averaged for all the sounds selected from this corpus. [b]Proportion index adjusts hit rates to compensate for varying numbers of categories, as described in Rosenthal and Rubin (1989). [c]Hit rates for individual sounds were not available; 68% is aggregated accuracy for the entire corpus by Belin et al. (2008), not our selection of 30 sounds. [d]Maurage et al. (2007) report the average scores on each emotion for each sound; we used this data to select the best recognized sounds, but it is not convertible into hit rates without access to original disaggregated responses.

number of well-recognized authentic vocalizations that could be matched with the acted ones; (e) had the best possible match in terms of the number of vocalizations produced by male and female speakers from each corpus and for each emotion; and (f) avoided including highly similar vocalizations produced by the same speaker (Table 1).

Authentic vocalizations were selected from a recently validated corpus (Anikin & Persson, 2016). Achievement was not included in the original corpus as a separate category, but we chose 10 naturalistic sounds from the following contexts to represent achievement: students passing an important exam ($n=2$ sounds), welcome news of an expected baby ($n=2$), and sport fans witnessing a victory of their team ($n=6$). Since the category of "achievement" was not used in the validation study, the comparison of naturalistic and acted sounds of achievement is best seen as tentative, and we did not include this emotion in the acoustic analysis.

Following the same procedure as that for actor portrayals, we selected authentic sounds with the highest recognition accuracy as reported by Anikin and Persson (2016). In addition, we excluded sounds containing noises that could give away the non-studio environment and thus enable participants to make authenticity judgments based on extraneous cues. Of the 139 authentic sounds used in this study, 127 were taken from the validated set of 260 sounds and 12 from previously untested material (six sounds of anger, five of disgust, and one of achievement). Adjusting for the number of categories in different studies (Rosenthal & Rubin, 1989), emotion recognition accuracy was consistently high for all corpora (Proportion Index > 90%; see Table 1).

*Procedure*. The behavioural study was conducted as an online experiment. Although online experiments allow for a limited control over the testing conditions (e.g., sound volume, background noise), they have been increasingly used in psychological research as they facilitate the access to large and diverse samples, potentially improving external validity and generalizability of findings (Birnbaum, 2004; Hewson, Vogel, & Laurent, 2016) and facilitating cross-cultural research (e.g., Cordaro et al., 2016).

Participants were informed that half of the sounds were "authentic (taken from YouTube videos of people engaged in emotionally charged activities)", and half were "fake (taken from several recent studies of emotion)". They were then presented with sounds from all corpora in random order and clicked one of two buttons to classify each sound as either "real (authentic)" or "fake (pretending)". To test whether knowledge of the experienced or portrayed emotion would affect the accuracy of authenticity detection, we compared two experimental conditions. In the cued condition, the name of the emotion being expressed was shown on the screen, while in the uncued condition the sound was presented without any emotional label. Each participant performed the entire test in either the cued or the uncued condition (between-subjects manipulation).

To facilitate recruitment and maintain the motivation of participants, the test was deliberately kept short and game-like. Participants were directed to one of two versions of the experiment and were asked to rate either 152 or 126 stimuli (152 + 126 = 278), which took on average 12 min. In both cases, 50% of sounds were authentic, and 50% were acted. Participants could replay the sounds, and they were given two types of feedback: The response button flashed green if the answer was correct and red if it was incorrect, and the current score was displayed at the bottom of the screen as the percentage of correct responses.

*Participants*. The experiment was available in English, Swedish, Russian, European Portuguese, and French. Participants were recruited through advertisements. They performed the experiment on their own computer and were not paid for their participation. Participants were informed about the aim of the experiment prior to taking part, and we did not record any personal information that could jeopardize the anonymity of respondents, apart from their first language. The total number of participants from each language group is shown in Table 2. The test could be interrupted at any time, and incomplete sessions were included in the analysis, provided that there were at least 100 responses per participant. Eighty-three participants had to be excluded because they had fewer than 100 trials, but most of them completed very few trials (median = 13), so that their data represented only ~8% of total responses.

*Controlling for potential extraneous acoustic cues of authenticity*. Acted vocalizations recorded in laboratory conditions are typically free from extraneous noises, but it is difficult to achieve the same level of acoustic purity with observational material. It is therefore conceivable that in some cases participants might have made authenticity judgments based on acoustic cues not related to the vocalization itself but instead indicative of the recording environment, such as traces of echo or noises in the background. To control for this possibility, we performed a second round of filtering to remove traces of extraneous noises. We used Audacity to

**Table 2.** Number of participants in the behavioural experiment.

| Participants' first language | Audio stimuli | |
|---|---|---|
| | Original | With masking noise |
| English | 58 | 25 |
| French | 12 | — |
| Portuguese | 19 | — |
| Dutch | — | 1 |
| Other (Swedish, German, etc.) | 17 | 20 |
| Total/number of times each sound was rated | 108/54.5 | 46/16.5 |

remove short clicks (by deleting a few milliseconds of audio), hiss, echo, and background noise when it was present and easily removable without degrading the audio quality (by using the "noise removal" feature, low-pass, high-pass, and notch filters). This did not change the fundamental frequency of original sounds. Only a small proportion of sounds needed to be filtered at all, since most of the material in these published corpora is already "clean", and we also pre-selected both authentic and acted sounds with the least acoustic impurities. After this additional filtering we added a controlled amount of noise to all vocalizations, both authentic and acted.

We used noise with amplitude equal to 50% of the maximum amplitude of each sound and with spectral shape described by power law with exponent coefficient alpha = 1.2. This is roughly midway between pink noise (alpha = 1) and Brownian noise (alpha = 1.5). The level and spectral slope of noise were chosen so as to make it effective at masking acoustic impurities but minimally intrusive. Subjectively, this noise was quite loud, and participants reported that certain sounds were practically inaudible. We hypothesized that, if the authenticity of sounds could still be detected in this masked condition, this would provide evidence that authenticity judgments were not made solely on the basis of extraneous noises.

The noise was generated and added to sound files in R (https://www.r-project.org). The same 278 sounds were then tested in this masked condition. Participants were asked to rate a random selection of 100 sounds, which took on average 8 minutes. We recruited 46 new participants in this control condition (who had not taken part in the test without masking noise).

*Acoustic analysis*. All available sounds from all corpora ($N = 903$, including 278 in the experiment described above) were acoustically analysed in R. Since some of the original sounds had been modified (shortened and/or filtered) in preparation for the experiment, the dataset of 278 experimental sounds was also re-analysed separately for the purpose of modelling acoustic predictors of authenticity judgments.

Syllable segmentation was performed using a custom algorithm described in Anikin and Persson (2016). Spectral features were extracted using fast Fourier transform with 50 ms Bartlett window and 50% overlap. To measure pitch and other related variables, we developed a custom pitch tracker implemented in R. This algorithm combines Praat's autocorrelation method described in Boersma (1993) with the BaNa algorithm, which is based on comparing ratios between harmonics in the spectrum (Ba, Yang, Demirkol, & Heinzelman, 2012). Since accurate pitch tracking is hard to achieve with such a wide variety of sounds, median pitch was also checked manually. Manual and automatic measurements of median pitch were highly correlated: $r = .95$ on a logarithmic scale. Nevertheless, measures of variability (*SD* of pitch, *SD* of energy in higher harmonics, etc.) may be inflated for relatively aperiodic sounds, such as roars of anger or croaky grunts of disgust. Where authentic and acted sounds differ in these acoustic variables, we can therefore conclude that there is some objective difference between the corresponding audio files, but the way this difference is perceived by the human ear may not be accurately captured by the acoustic features reported.

Our analysis was focused on the following acoustic variables, which were selected based on (a) previous research on acoustic correlates of emotion and authenticity (Anikin & Persson, 2016; Banse & Scherer, 1996; Lavan et al., 2015) and (b) screening possible acoustic predictors using Random Forests (see Statistics, supplementary data):

*Pitch (Hz)*: fundamental frequency for relatively tonal sounds or the lowest dominant frequency band for voiced sounds with blurred harmonics. We used pitch floor of 75 Hz and ceiling of 3500 Hz for all measurements. Pitch and other measures expressed in Hz were log2 transformed for statistical modelling.

*Harmonics-to-noise ratio (HNR; dB)*: the ratio of energy in harmonics to the total amount of spectral energy. HNR was calculated for all non-silent frames with at least one pitch candidate identified by the autocorrelation method, whether or not this candidate exceeded the voicing threshold.

*First quartile of spectral energy distribution (Hz)*: one quarter of the total acoustic energy in the spectral region from 75 to 6000 Hz is found under this frequency. The cut-off points were chosen to ensure that low-frequency noise and the differences in the sampling rate of original sounds would not influence spectral measures.

*Energy above F0 (dB)*: the ratio of energy above 1.25 times fundamental frequency (F0) in the spectral region from 75 to 6000 Hz to the entire amount of spectral energy in this frequency range.

*Spectral slope (% of amplitude range per kHz)*: the slope of regression line fitted to the spectrum of an analysis window in the region from 75 to 6000 Hz.

*Syllable length (ms)*: the length of segments continuously, or with interruptions no longer than 50 ms, exceeding an amplitude threshold chosen dynamically as a proportion of the global mean amplitude in the smoothed amplitude envelope of the entire vocalization.

*Interburst interval (ms)*: interval between adjacent vocal bursts, defined as local maxima in smoothed amplitude envelopes that exceed the global maximum and surrounding points by certain thresholds (initially set by iterative optimization against manual measurements).

*Root mean square (RMS) amplitude*: a measure of subjective acoustic intensity of voiced frames. Since all files were normalized for peak amplitude prior to

processing, the median value of RMS amplitude shows how sustained, rather than loud, a sound is.

*Voiced (%)*: proportion of voiced frames out of the total duration of the sound.

All these acoustic features, except for the proportion of voiced frames, were extracted for each 50-ms analysis window and were summarized as median and standard deviation over the entire sound. We used median rather than mean values, since medians are more robust to outliers, such as frames with incorrectly measured pitch or external noise. Since the duration of sounds varied considerably both within and across corpora, all temporal measures were made independent of sound duration (e.g., we used interburst interval rather than the absolute number of bursts).

### *Statistics*

*Analysis of experimental results*. The outcome variable for all models was the classification of a single sound by a single participant as either "real" or "fake", which could be correct or incorrect. We analysed the effects of trial number, condition (cued or uncued), corpus, emotion, and the (mis)match between the first language of raters and speakers on these unaggregated individual answers using logistic regression with two random effects: sound and participant. This and other linear models were fitted using Markov chain Monte Carlo in the Stan computational framework (http://mc-stan.org/) accessed from R.

*Analysis of acoustic features*. To minimize the risk of false positives associated with multiple comparisons, we obtained confidence intervals from multiple regression models (rather than pairwise comparisons, e.g., with *t*-tests), in which all beta coefficients were further assumed to come from a single distribution. This method causes regression coefficients to shrink towards zero and provides an in-built correction for multiple comparisons (Kruschke, 2014). The influence of predictors is considered simultaneously in multiple regression, so that the conclusions depend on which predictors are included in the model. For example, the first quartile of spectral energy distribution is noticeably higher in authentic sounds, but once we have accounted for other variables, such as pitch and energy in upper harmonics, the corresponding beta-coefficient in Figure 4a approaches zero.

We also explored a wider range of potentially relevant acoustic variables using a machine learning algorithm—Random Forest (Breiman, 2001). This method, which builds a large number of decision trees using different combinations of predictors, can be a more powerful alternative to multinomial regression or discriminant analysis. It is particularly suitable for selecting the most important potential predictors among a large number of variables with complex interactions. This method identified a few additional predictors, such as peak frequency and lowest dominant frequency band, but these variables were strongly correlated with each other, with pitch, and with our primary measures of spectral energy distribution (first quartile and spectral slope). They were therefore not included in the final analysis.

We also used Random Forests to see to what extent models trained on authentic vocalizations would be able to recognize the intended emotion of actor portrayals, and vice versa. One class of vocalizations (authentic or acted) served as a training set for building decision trees, and the other class as the testing set. The estimates of classification accuracy within the training set are based on an internal cross-validation procedure: Roughly two thirds of data are used to train the model, and one third is used for internal cross-validation. Since a Random Forest model consists of hundreds of independently built trees, which use different training sets, the algorithm generates an estimate of out-of-the-bag classification accuracy for each sound (Breiman, 2001). (Supplementary materials for this article, including R scripts used for acoustic and statistical analyses, raw data and the corpus of authentic vocalizations, are available.)

## *Results*

### *Authenticity classification experiment*

*Authenticity recognition per corpus*. The first question we investigated was whether listeners could distinguish between authentic and acted vocalizations. Without masking noise, average accuracy was 65.4% (chance level 50%), corresponding to an odds ratio (*OR*) of 2.1 (95% confidence interval, CI [1.9, 2.4]) in favour of being correct. With masking noise, accuracy was 64.6% ($OR = 2.0$, 95% CI [1.7, 2.3]). Trial number was not a significant predictor of success (logistic regression with participant and sound as random effect, likelihood ratio $L = 0.74$, $df = 1$, $p = .39$), and after performing 100 trials the odds of answering correctly were only 1.03 [0.94, 1.1] times higher than at the beginning of the test. There was also no interaction between trial effect and condition (cued or uncued: $L = 1.04$, $df = 1$, $p = .31$). Despite the feedback received by participants after each trial, there was thus no learning effect: Accuracy remained the same throughout the experiment and did not depend on the number of completed trials.

Perceived authenticity of different corpora of acted sounds varied considerably. For two corpora, by Hawk et al. (2009) and Cordaro et al. (2016), the proportion of "real" responses was close to the level expected by chance (Figure 1a). In contrast, this proportion was below the chance level of 50% for the corpora by Simon-Thomas et al. (2009), Lima et al. (2013), Maurage et al. (2007), and Belin et al. (2008), indicating that listeners consistently perceived these acted expressions as unauthentic. The addition of masking noise did not affect the overall

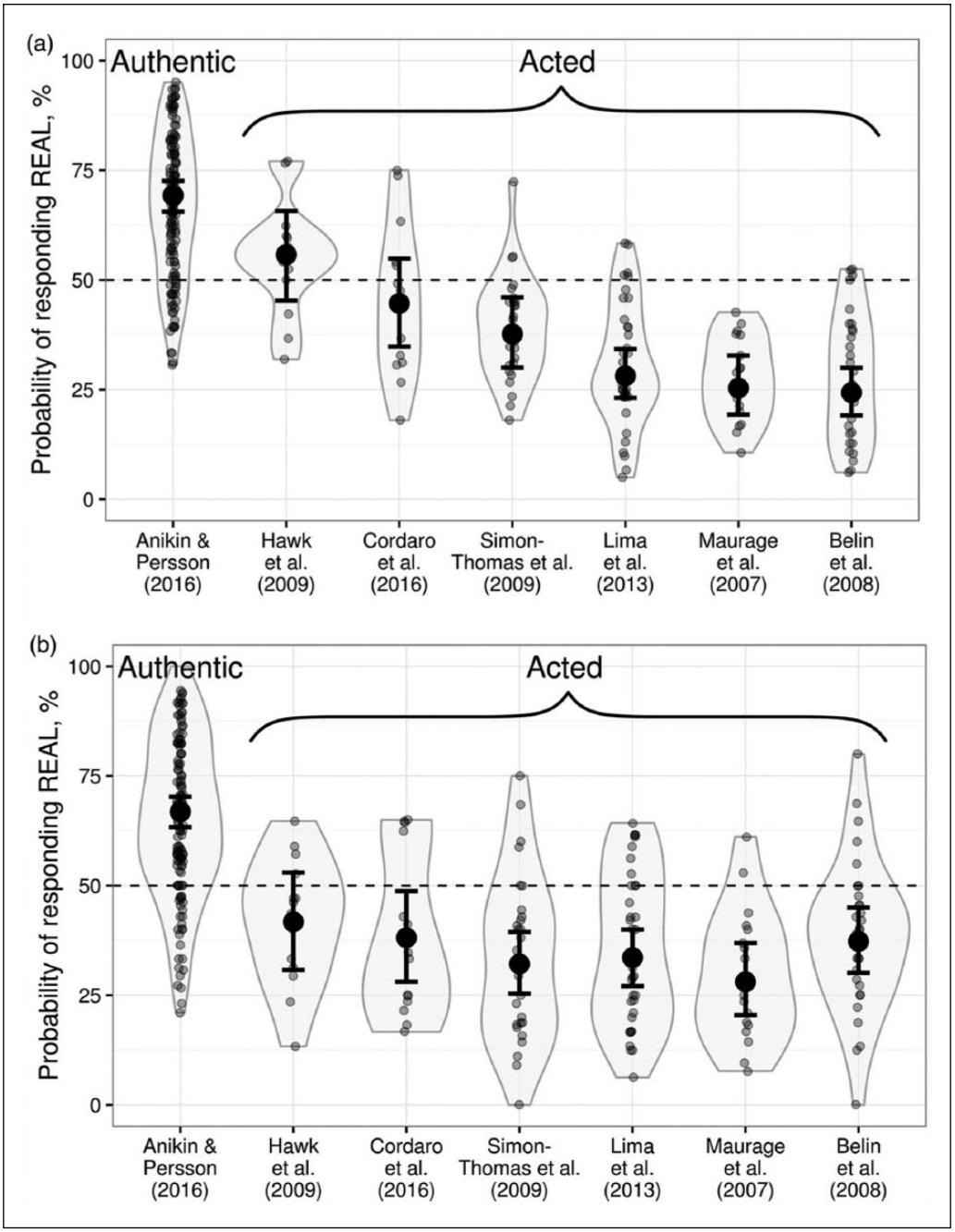


**Figure 1.** Perceived authenticity of sounds from each corpus collapsed across all emotions and calculated as the proportion of trials in which the sound was classified as "real" rather than "fake", (a) without and (b) with masking noise. Small dots show the authenticity of individual sounds, while the large circles with error bars show the median of posterior distribution for the entire corpus and 95% confidence interval (CI). The dotted line at 50% shows chance level.

proportion of "real" responses ($L = 0.002$, $df = 1$, $p = .96$), but there was an interaction between corpus and masking noise ($L = 49.9$, $df = 6$, $p < .001$), suggesting that sounds from different corpora were affected differently by the addition of the masking noise (Figure 1b). Specifically, there were two corpora for which the accuracy of authenticity detection changed after the addition of masking noise: by Belin et al. (2008) and Hawk et al. (2009).

The corpus by Hawk et al. (2009) is the only one that contains recordings by professional actors, and before the addition of masking noise it also had the highest perceived authenticity among corpora of acted sounds. It would be interesting to find out whether the high perceived authenticity of these sounds was related to employing professional actors or to some other characteristic of Hawk's corpus. One possible reason for higher perceived authenticity of

**Table 3.** Accuracy of classifying 278 sounds as "real" or "fake".

| Emotion | Audio stimuli | |
|---|---|---|
| | Original | With masking noise |
| Achievement | 74.8 [66.8, 80.3] | 69 [59.8, 76.7] |
| Amusement | 63.3 [57.1, 68.9] | 60.1 [53.2, 66.4] |
| Anger | 70.5 [64.5, 75.5] | 71.4 [65.1, 77.2] |
| Disgust | 62 [56.5, 67.7] | 63.9 [57.2, 70.6] |
| Fear | 76.3 [71.2, 80.6] | 74 [68.2, 79] |
| Pain | 65.4 [57.8, 72.2] | 58 [49.2, 66] |
| Pleasure | 69.1 [62.9, 74.8] | 68.8 [61.8, 75] |
| Sadness | 64.7 [58.8, 70.5] | 63.5 [57.1, 69.3] |
| Total | 67.9 [65.4, 70.3] | 66.2 [62.9, 69.7] |

Note: Accuracy in percentages. The values shown are medians of the posterior distribution in a logistic model with two random effects (sound and participant). They are slightly different from the actual observed proportions. Values in square brackets indicate 95% confidence intervals.

these sounds is their poorer acoustic quality—namely significant clipping—which contrasts with the typically "clean" studio recordings. Consistent with this explanation, the proportion of "real" answers for this corpus dropped by 13.8% (95% CI [5.2, 22.5]) after we added masking noise. In contrast, the perceived authenticity of sounds from the corpus by Belin et al. (2008) increased by 12.5% (95% CI [5.7, 19.3]) after the addition of noise, perhaps because it helped to mask some acoustic cues giving away the studio environment. Nevertheless, acoustic quality cannot be the only factor that influenced authenticity judgments in general, since the perceived authenticity of the remaining five corpora did not change after the addition of strong background noise. In particular, masking noise had no effect on the proportion of "real" responses for the authentic sounds (−2.1%, 95% CI [−6.3, 2.5]; see Figure 1b).

*Authenticity recognition per emotion*. The overall authenticity detection accuracy per emotion, with and without masking noise, is presented in Table 3. It was above the chance level of 50% for all emotions both with and without masking noise, with the possible exception of pain in the condition with masking noise, for which the 95% CI was [49.2%, 66%]. Accuracy may not be the most informative statistic for analysing responses per emotion, however, because of the possibility of an overall bias to classify all sounds in particular emotional categories as "real" or "fake". To account for this, we also analysed the frequency with which different sounds were judged to be "real"—their perceived authenticity. As shown in Figure 2a, authentic vocalizations were significantly more likely to be classified as "real" than were acted sounds for all eight emotions, but the magnitude of this effect varied considerably: It was largest for achievement, anger, fear, and pleasure, but relatively low for amusement, disgust, and sadness. There was an overall bias to perceive laughs (amusement) as authentic, causing the accuracy of authenticity detection to drop. In contrast, for disgust and sadness the relatively poor accuracy of authenticity detection was caused by the low perceived authenticity of authentic vocalizations.

These findings were largely replicated when masking noise was added, but with slightly lower accuracy for pain and achievement (Table 3; Figure 2b). There was no interaction between the presence of masking noise and emotion in models for predicting the probability of either classifying a sound correctly or perceiving it as "real" (likelihood ratio tests, $L = 10.1$ and 9.7, $p = .18$ and .20, respectively, $df = 7$ for both models). For each emotion, 95% CI for the change in perceived authenticity after the addition of masking noise included zero (details not shown).

*Effects of linguistic background and knowledge of target emotion*. The next question we asked was whether informing participants about the emotion being expressed would have any effect on authenticity detection. Without masking noise, accuracy was 1.4% higher in the cued condition (target emotion shown) than in the uncued condition (emotion not shown). After controlling for emotion and corpus, this corresponds to an *OR* of 1.07 (95% CI [0.89, 1.27]). There was no interaction between condition and authenticity ratings per corpus or per emotion (likelihood ratio tests: $L = 5.5$ and 6.5, $df = 6$ and 7, respectively; $p = .48$ in both cases). With masking noise added, accuracy was again only marginally higher in the cued than in the uncued condition (2.5%, $OR = 1.12$, 95% CI [0.89, 1.40]). Knowledge of the target emotion thus had little or no effect on the accuracy with which vocalizations were classified as authentic or acted. This variable was therefore not considered in other analyses, and the data from both cued and uncued conditions were pooled.

We also examined whether the listener's linguistic background affected authenticity detection. Pooling the data with and without masking noise, overall accuracy was 65.0% when the rater's first language was the same as the speaker's, and 65.4% when it was different, $OR = 0.97$

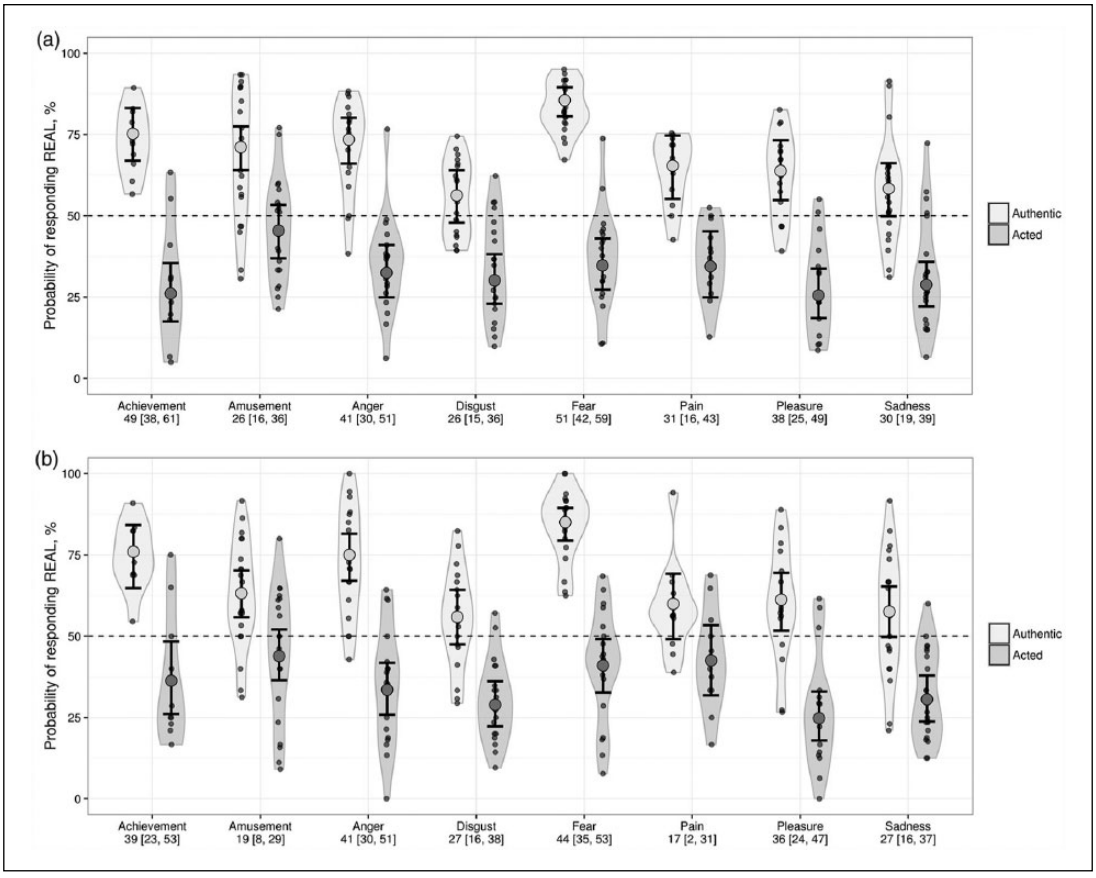


**Figure 2.** Perceived authenticity of authentic versus acted sounds in each emotional category calculated as the proportion of trials in which the sound was classified as "real" rather than "fake", (a) without and (b) with masking noise: median of posterior distribution and 95% confidence interval (CI). The most credible difference (%) between authentic and acted vocalizations is listed under each emotion, with 95% CI. The dotted line at 50% shows chance level.

(95% CI [0.89, 1.06]). When each corpus was considered separately, again no clear pattern emerged (Figure 3). Based on the available results, it appears that the (mis) match between the first language of the speaker and that of the rater had no effect on the detection of authenticity.

*Acoustic correlates of authenticity.* To understand which acoustic characteristics make it possible to distinguish between authentic vocalizations and actor portrayals, we performed three types of analysis. First, we compared all the available sounds to identify acoustic differences between authentic and acted emotional vocalizations. Second, we analysed acoustic predictors of subjective authenticity judgments for the smaller subset of 278 sounds that were included in the behavioural experiment. Finally, to explore the impact of the differences between authentic and acted vocalizations on the performance of automatic classification algorithms applicable to affective computing, we trained acoustic classification models using acted sounds and tested the ability of these models to recognize the emotion of authentic vocalizations (and vice versa).

*Acoustic differences between authentic and acted vocalizations.* For this analysis, we included all available authentic vocalizations, except for effort (because there were no acted counterparts to serve as comparison) and joy (because this category covered a number of positive states with unclear correspondence to the categories of achievement and triumph from other corpora). Acted vocalizations consisted of all sounds from the seven chosen emotional categories in the six corpora listed in Table 1, in addition to the sounds from the study by Sauter et al. (2010).

As shown in Table 4, authentic vocalizations differ from actor portrayals in a number of acoustic characteristics, including their considerably higher median pitch (Cohen's $d$ >1 for anger, fear and pleasure). To test the statistical significance of these differences, we used multiple logistic regression, in which acoustic variables were used to predict the status of each sound as authentic or acted. The resulting plot of beta-coefficients in Figure 4a shows the strength of independent contribution of each acoustic variable to separating authentic from acted vocalizations, controlling for other acoustic variables. According to this model, authentic sounds have a higher pitch, lower

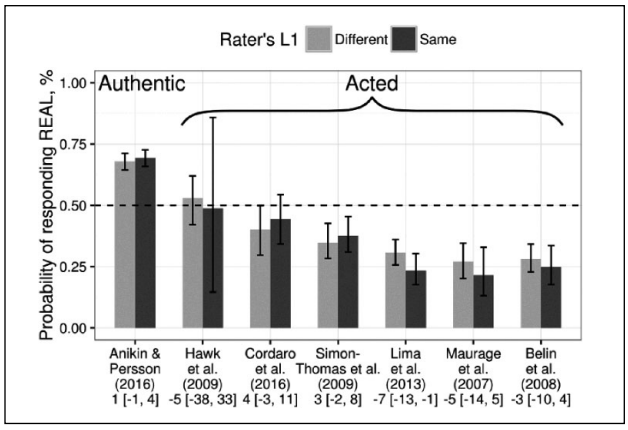


**Figure 3.** Perceived authenticity of sounds from each corpus: median of posterior distribution and 95% confidence interval (CI). The most credible difference (%) between raters with the same versus different first language as the caller's is shown under each emotion, with 95% CI. The dotted line at 50% shows chance level.

harmonicity, less variable spectral slope, and less variable amplitude. Interburst intervals in authentic vocalizations tend to be longer and more irregular. The proportion of voiced frames is higher for authentic sounds than for acted ones, but this measure is not robust with respect to the method of obtaining and cutting audio clips, which may have differed across studies. For the same reason, we did not consider duration and variables directly dependent on it, such as the number of syllables, as predictors in the logistic model (but duration was analysed as a potential acoustic cue used by participants to detect authenticity). The proportion of energy above F0 is more variable in authentic sounds (Table 4), but this difference is likely to be an artefact: Unlike the median value, the standard deviation of this variable is particularly sensitive to imperfect pitch tracking, which is more of a problem for the less tonal authentic sounds. This variable was therefore excluded from the regression model.

Since female voices are generally more high pitched than male voices, a possible concern is that some of the discovered acoustic markers of authenticity, especially differences in the fundamental frequency, may be related to the unequal numbers of male and female vocalizations in the authentic and acted samples (see Table 1). However, in a linear regression model, authenticity remained a significant predictor of higher pitch among the 789 analysed sounds, even after controlling for the sex of the speaker. Both authenticity and sex had a strong and independent effect on pitch, which was predicted to be 0.93 standard deviations higher if the speaker was female rather than male, $F(1, 785)=270$, $p<.001$, and 0.75 standard deviations higher if the sound was authentic rather than acted, $F(1, 785)=176$, $p<.001$. The higher pitch of authentic vocalizations thus cannot be explained only by a skewed sex ratio. Furthermore, there was no interaction between sex and authenticity, $F(1, 785)=1.8$, $p=.18$, across all emotional categories, indicating that authentic vocalizations tended to have a higher pitch for both male and female speakers.

The regression model in Figure 4a does not allow for complex interactions between predictors and ignores the considerable differences between emotions (cf. Table 4). Nevertheless, it can correctly classify ~76% of the 789 sounds as authentic or acted. Furthermore, to account for possible interactions, we employed a more powerful Random Forest algorithm. Using the same 16 predictors, this model achieves cross-validation classification accuracy of ~78%, which is similar to the accuracy of the regression model. Using emotion as a predictor improves the accuracy of the Random Forest model, but only slightly (to ~80%), suggesting that the model captures the acoustic differences between authentic and acted sounds in general, rather than emotion-specific acoustic markers of authenticity.

Finally, we narrowed down the range of predictors of authenticity to only five variables that are the best predictors of emotion, rather than authenticity: pitch, harmonicity, interburst interval, first quartile of spectral energy, and amplitude. A Random Forest model using only these five predictors classified 789 sounds as authentic or acted with cross-validation accuracy of 72%. Since this is close to the accuracy of the full model with 16 predictors (78%), the acoustic differences between authentic and acted sounds cannot be dismissed as artefacts related to less robust or imperfectly measured variables. On the contrary, the same acoustic variables that predict emotion can also serve as robust predictors of authenticity.

*Acoustic predictors of subjective authenticity judgments*. A logistic model with the same 16 predictors (without interaction) as those above was fitted to predict individual responses of participants, who classified a subset of 278 sounds as real or fake, with sound and participant as random effects. The strongest predictors of perceived authenticity (Figure 4b) were similar to predictors of objective authenticity (Figure 4a): higher and more variable pitch, lower harmonicity, and widely and irregularly spaced vocal bursts. The variability of amplitude and spectral slope were no longer strong predictors, suggesting that listeners did not rely on these acoustic characteristics for authenticity detection.

Since authentic sounds were on average ~0.5 s longer in duration, we tested separately whether duration was an important predictor of perceived authenticity. In a model with duration as the only predictor, its beta coefficient was 0.18 (95% CI [0.11, 0.27]). Controlling for other acoustic predictors listed in Figure 4, however, duration no longer predicted perceived authenticity (beta coefficient = 0.03, 95% CI [−0.07, 0.14]), and it did not improve the overall accuracy with which the model predicted authenticity

**Table 4.** Acoustic features of authentic and acted vocalizations.

| Acoustic variable | Source | Emotion | | | | | | | Mean Cohen's *d* per variable |
|---|---|---|---|---|---|---|---|---|---|
| | | Amusement (*Mean* ± *SD*) | Anger (*Mean* ± *SD*) | Disgust (*Mean* ± *SD*) | Fear (*Mean* ± *SD*) | Pain (*Mean* ± *SD*) | Pleasure (*Mean* ± *SD*) | Sadness (*Mean* ± *SD*) | |
| Duration (s) | Authentic | 2.46 ± 1.52 | 1.76 ± 1.65 | 0.92 ± 0.46 | 1.41 ± 0.8 | 1.77 ± 1.29 | 2.64 ± 2.09 | 2.97 ± 2.22 | |
| | Acted | 2.77 ± 2.9 | 1.43 ± 1.3 | 1.49 ± 1.77 | 1.78 ± 2.62 | 1.09 ± 0.97 | 1.33 ± 0.44 | 2.41 ± 2.63 | |
| | Cohen's *d* | −0.14 | 0.22 | −0.42 | −0.18 | 0.54 | 0.81 | 0.23 | 0.15 |
| Pitch, median (Hz) | Authentic | 472 ± 191 | 718 ± 512 | 265 ± 113 | 1104 ± 541 | 569 ± 357 | 331 ± 162 | 428 ± 176 | |
| | Acted | 317 ± 111 | 293 ± 111 | 236 ± 78 | 463 ± 256 | 378 ± 111 | 194 ± 62 | 314 ± 118 | |
| | Cohen's *d* | 0.95 | 1.11 | 0.31 | 1.58 | 0.57 | 1.06 | 0.77 | 0.91 |
| Pitch, *SD* (Hz) | Authentic | 129 ± 108 | 169 ± 154 | 77 ± 108 | 191 ± 136 | 141 ± 115 | 81 ± 79 | 136 ± 121 | |
| | Acted | 79 ± 66 | 63 ± 52 | 81 ± 90 | 75 ± 89 | 73 ± 52 | 46 ± 21 | 59 ± 49 | |
| | Cohen's *d* | 0.54 | 0.89 | −0.04 | 1.04 | 0.63 | 0.57 | 0.86 | 0.64 |
| Harmonicity, median (dB) | Authentic | 4.5 ± 3.1 | 2.9 ± 3.2 | 6.1 ± 4.2 | 10 ± 4.6 | 7.5 ± 5.6 | 6.8 ± 4.3 | 9.7 ± 5 | |
| | Acted | 4.9 ± 3.4 | 6.1 ± 4.6 | 7.3 ± 3.5 | 11.3 ± 4.6 | 8.9 ± 4.8 | 11.9 ± 2.9 | 10.3 ± 4.3 | |
| | Cohen's *d* | −0.12 | −0.82 | −0.31 | −0.28 | −0.25 | −1.35 | −0.13 | −0.47 |
| Harmonicity, *SD* (dB) | Authentic | 4.4 ± 1.2 | 3.6 ± 1.4 | 3.8 ± 1.3 | 4.8 ± 1.5 | 4.8 ± 1.5 | 4.5 ± 1.5 | 5 ± 1.4 | |
| | Acted | 4.3 ± 1.3 | 3.9 ± 1.2 | 4.3 ± 1.3 | 4.6 ± 1.8 | 5 ± 2.2 | 3 ± 1.5 | 4.3 ± 1.6 | |
| | Cohen's *d* | 0.08 | −0.23 | −0.38 | 0.12 | −0.12 | 1 | 0.46 | 0.13 |
| First quartile, median (Hz) | Authentic | 824 ± 259 | 1080 ± 423 | 695 ± 306 | 1228 ± 505 | 988 ± 309 | 580 ± 296 | 705 ± 353 | |
| | Acted | 610 ± 251 | 669 ± 278 | 620 ± 307 | 715 ± 327 | 919 ± 202 | 313 ± 215 | 471 ± 211 | |
| | Cohen's *d* | 0.84 | 1.13 | 0.24 | 1.24 | 0.23 | 1.01 | 0.82 | 0.79 |
| First quartile, *SD* (Hz) | Authentic | 213 ± 104 | 202 ± 130 | 145 ± 130 | 218 ± 135 | 174 ± 115 | 180 ± 123 | 313 ± 188 | |
| | Acted | 195 ± 77 | 155 ± 78 | 154 ± 111 | 155 ± 105 | 138 ± 59 | 89 ± 80 | 127 ± 91 | |
| | Cohen's *d* | 0.19 | 0.43 | −0.08 | 0.53 | 0.33 | 0.85 | 1.29 | 0.51 |
| Energy above F0, median (dB) | Authentic | 7.7 ± 4.5 | 9.7 ± 5.8 | 10.9 ± 4 | 1.8 ± 5.4 | 9.6 ± 6.3 | 7.1 ± 4.3 | 6 ± 3.9 | |
| | Acted | 7.3 ± 3.5 | 10.4 ± 4.2 | 9.8 ± 3.9 | 7.9 ± 4.6 | 14 ± 3.6 | 3.8 ± 4.8 | 5.4 ± 4 | |
| | Cohen's *d* | 0.1 | −0.14 | 0.28 | −1.23 | −0.73 | 0.73 | 0.15 | −0.12 |
| Energy above F0, *SD* (dB) | Authentic | 4.2 ± 1.5 | 4.6 ± 2.3 | 3.1 ± 1.8 | 3.9 ± 2.2 | 3.9 ± 1.7 | 3.5 ± 1.5 | 4.2 ± 1.4 | |
| | Acted | 3.2 ± 1.1 | 2.5 ± 1 | 3 ± 1.5 | 2.5 ± 1.3 | 2.7 ± 1.5 | 2.1 ± 0.7 | 3 ± 1.2 | |
| | Cohen's *d* | 0.74 | 1.15 | 0.06 | 0.8 | 0.72 | 1.14 | 0.93 | 0.79 |
| Spectral slope, median (% of range per kHz) | Authentic | −0.6 ± 0.3 | −1 ± 0.5 | −0.9 ± 0.5 | −0.6 ± 0.4 | −0.8 ± 0.5 | −0.7 ± 0.4 | −0.4 ± 0.2 | |
| | Acted | −0.7 ± 0.4 | −1.3 ± 0.7 | −0.8 ± 0.5 | −0.7 ± 0.4 | −0.9 ± 0.6 | −0.7 ± 0.4 | −0.7 ± 0.5 | |
| | Cohen's *d* | 0.29 | 0.5 | −0.2 | 0.25 | 0.19 | 0 | 0.77 | 0.26 |
| Spectral slope, *SD* (% of range per kHz) | Authentic | 0.5 ± 0.2 | 0.4 ± 0.2 | 0.4 ± 0.2 | 0.3 ± 0.1 | 0.3 ± 0.2 | 0.4 ± 0.2 | 0.3 ± 0.1 | |
| | Acted | 0.6 ± 0.2 | 0.6 ± 0.3 | 0.5 ± 0.2 | 0.4 ± 0.2 | 0.4 ± 0.3 | 0.3 ± 0.2 | 0.4 ± 0.2 | |
| | Cohen's *d* | −0.5 | −0.8 | −0.5 | −0.61 | −0.46 | 0.5 | −0.62 | −0.43 |
| Syllable length, median (ms) | Authentic | 129 ± 85 | 469 ± 242 | 326 ± 153 | 326 ± 146 | 375 ± 170 | 354 ± 165 | 234 ± 189 | |
| | Acted | 106 ± 44 | 495 ± 212 | 391 ± 158 | 373 ± 267 | 360 ± 107 | 561 ± 211 | 255 ± 170 | |
| | Cohen's *d* | 0.32 | −0.11 | −0.42 | −0.21 | 0.09 | −1.11 | −0.12 | −0.22 |

*(Conintued)*

**Table 4.** (Continued).

| Acoustic variable | Source | Emotion | | | | | | | Mean Cohen's *d* per variable |
|---|---|---|---|---|---|---|---|---|---|
| | | Amusement (*Mean* ± *SD*) | Anger (*Mean* ± *SD*) | Disgust (*Mean* ± *SD*) | Fear (*Mean* ± *SD*) | Pain (*Mean* ± *SD*) | Pleasure (*Mean* ± *SD*) | Sadness (*Mean* ± *SD*) | |
| Syllable length, *SD* (ms) | Authentic | 121 ± 128 | 289 ± 244 | 191 ± 126 | 315 ± 186 | 333 ± 220 | 208 ± 178 | 230 ± 162 | |
| | Acted | 87 ± 77 | 368 ± 222 | 157 ± 104 | 220 ± 206 | 232 ± 131 | 332 ± 191 | 177 ± 123 | |
| | Cohen's *d* | 0.31 | −0.34 | 0.3 | 0.48 | 0.48 | −0.68 | 0.37 | 0.13 |
| Interburst interval, median (ms) | Authentic | 320 ± 267 | 917 ± 467 | 699 ± 398 | 626 ± 294 | 942 ± 440 | 998 ± 757 | 655 ± 453 | |
| | Acted | 253 ± 83 | 2281 ± 909 | 1220 ± 746 | 863 ± 569 | 1467 | 927 ± 177 | 470 ± 386 | |
| | Cohen's *d* | 0.32 | −1.93 | −0.84 | −0.51 | — | 0.12 | 0.44 | −0.4 |
| Interburst interval, *SD* (ms) | Authentic | 143 ± 175 | 276 ± 255 | 86 ± 114 | 385 ± 222 | 319 ± 296 | 226 ± 166 | 285 ± 274 | |
| | Acted | 115 ± 148 | 354 | 569 ± 349 | 479 ± 537 | 25 | — | 254 ± 303 | |
| | Cohen's *d* | 0.17 | — | −1.77 | −0.22 | — | — | 0.11 | −0.43 |
| RMS amplitude, median (% of range) | Authentic | 16 ± 6 | 29 ± 12 | 20 ± 6 | 27 ± 10 | 25 ± 7 | 20 ± 7 | 14 ± 7 | |
| | Acted | 16 ± 9 | 26 ± 16 | 19 ± 9 | 29 ± 18 | 20 ± 5 | 26 ± 8 | 18 ± 9 | |
| | Cohen's *d* | 0 | 0.21 | 0.13 | −0.13 | 0.74 | −0.81 | −0.49 | −0.05 |
| RMS amplitude, *SD* (% of range) | Authentic | 10 ± 4 | 10 ± 5 | 9 ± 3 | 10 ± 4 | 10 ± 3 | 11 ± 4 | 11 ± 3 | |
| | Acted | 12 ± 7 | 14 ± 9 | 12 ± 7 | 15 ± 7 | 11 ± 3 | 13 ± 3 | 12 ± 7 | |
| | Cohen's *d* | −0.37 | −0.56 | −0.53 | −0.85 | −0.33 | −0.55 | −0.18 | −0.48 |
| Proportion of voiced frames (%) | Authentic | 62 ± 22 | 78 ± 17 | 76 ± 20 | 90 ± 11 | 81 ± 16 | 63 ± 25 | 69 ± 23 | |
| | Acted | 52 ± 21 | 73 ± 23 | 75 ± 19 | 69 ± 23 | 75 ± 19 | 81 ± 18 | 70 ± 26 | |
| | Cohen's *d* | 0.46 | 0.25 | 0.05 | 1.12 | 0.37 | −0.81 | −0.04 | 0.2 |
| Cohen's *d* per emotion | | **0.36** | **0.64** | **0.38** | **0.63** | **0.42** | **0.77** | **0.49** | |

Note: Authentic vocalizations: $n = 427$; acted vocalizations: $n = 362$. RMS = root mean square. Values in bold in the last row show Cohen's *d* per emotion.

judgments. It is therefore unlikely that duration was an important acoustic cue guiding the judgments of participants.

Overall, beta coefficients were considerably smaller in the model predicting subjective authenticity judgments than in the model predicting objective authenticity (Figure 4b vs. Figure 4a). The accuracy of predicting perceived authenticity was ~62% (~68% using Random Forest instead of logistic regression), which is considerably lower than that for the prediction of objective authenticity based on the same acoustic characteristics (78%). This may partly be due to the smaller number of sounds (278 vs. 789), but it also suggests that more sophisticated acoustic predictors might be needed. As an illustration, Figure 5 shows the spectrograms of two sounds from each emotional category: one with very high and one with very low perceived authenticity. Which acoustic features create this large difference in perceived authenticity is an issue that future models will have to address more comprehensively.

*Recognition of emotion by classifiers trained on acted and tested on authentic sounds*. Algorithms for automatic recognition of emotion in the voice are often trained on corpora of acted sounds but ultimately intended to be used in a real-life context of human–machine interaction. To test the transfer of emotion classification from acted to authentic material, we trained a Random Forest classifier on acted sounds ($n = 362$) using 16 acoustic predictors (see Figure 4). This model classified the sounds in the training set into seven emotions with cross-validation accuracy of ~54% (chance level: 14%). We then tested the accuracy with which this model classified the emotion of authentic vocalizations ($n = 427$). Most emotions were recognized equally well in both sets, demonstrating good transfer. The main exception was pleasure: Its recognition rate dropped from 70% to 16% (Table 5).

The opposite situation—namely, training a model on authentic material and then using it to classify acted vocalizations—is less likely to arise, because most available corpora consist of acted expressions, but it is also possible. To evaluate the transfer in this direction, we trained a Random Forest model with the same 16 predictors on authentic sounds and then tested it on acted sounds. Recognition accuracy in the training set (authentic) was 59%, and in the testing set (acted) it was 39%. Actor portrayals of amusement, disgust, pain, and pleasure were recognized as well, or even better, in the test set as in the training set. In contrast, a model trained on authentic vocalizations failed to recognize acted sounds of fear, anger, and sadness.

Transfer between training and testing sets can be successful for a particular emotion only if acoustic differences between authentic and acted vocalizations of this emotion are small relative to the acoustic differences between emotions in the training set. To take a simple example, the average pitch of authentic screams in the corpus by Anikin and Persson (2016) is so high (~1100 Hz) that the lower cut-off point for this variable in decision trees may exceed the average pitch of acted screams (~460 Hz). As a result, a Random Forest model trained on authentic screams of fear fails to recognize actor portrayals of fear, misclassifying them as sounds of disgust or pain. In contrast, a model trained on acted vocalizations has no problem with recognizing authentic sounds of fear, since the learned rule (e.g., pitch higher than 460 Hz) is easily satisfied by authentic screams. Extending this simplified reasoning to multivariate comparisons, the current study suggests that acoustic differences between authentic and acted vocalizations of fear, anger, sadness, and pleasure may be large enough to have practical implications for affective computing: An algorithm trained on one type of vocalization (authentic or acted) may fail to generalize to the other type.

This problem can be largely avoided by training the classifier on a mix of authentic and acted vocalizations: The cross-validation accuracy then becomes more consistent for both authentic and acted vocalizations, overall as well as for each emotion (Table 5, right two columns). Partly this improvement may be due simply to having a larger training set. However, recognition accuracy remained consistent across emotions for both authentic and acted vocalizations even when we used half the sounds to train the classifier (a random selection of 183 authentic and 180 acted sounds in the training set; Table 5).

## Discussion

The extensive reliance on posed facial expressions and vocalizations in emotion research raises the question of how similar they are to expressions produced more spontaneously. Recent research suggests that listeners can to some extent discriminate between authentic and acted laughter (Bryant & Aktipis, 2014; Lavan et al., 2015), but whether this generalizes to other nonverbal vocalizations remained unknown. In this study we tested for the first time whether vocalizations emitted spontaneously in a wide range of emotionally charged situations could be distinguished, by human raters and acoustic models, from acted vocalizations intended to portray the same emotions.

Nonverbal vocalizations from a previously validated observational corpus based on amateur videos (Anikin & Persson, 2016) were consistently judged as more authentic than acted vocalizations from a range of published corpora, across all eight analysed emotions: achievement, amusement, anger, disgust, fear, pain, pleasure, and sadness. Crucially, this was not due to the presence of extraneous noises giving away the non-studio environment, as the effect was replicated when background masking noise was added to all sounds. The fact that we could not guarantee ideal testing conditions (because the experiment was conducted online), if anything, makes our study a more stringent test of

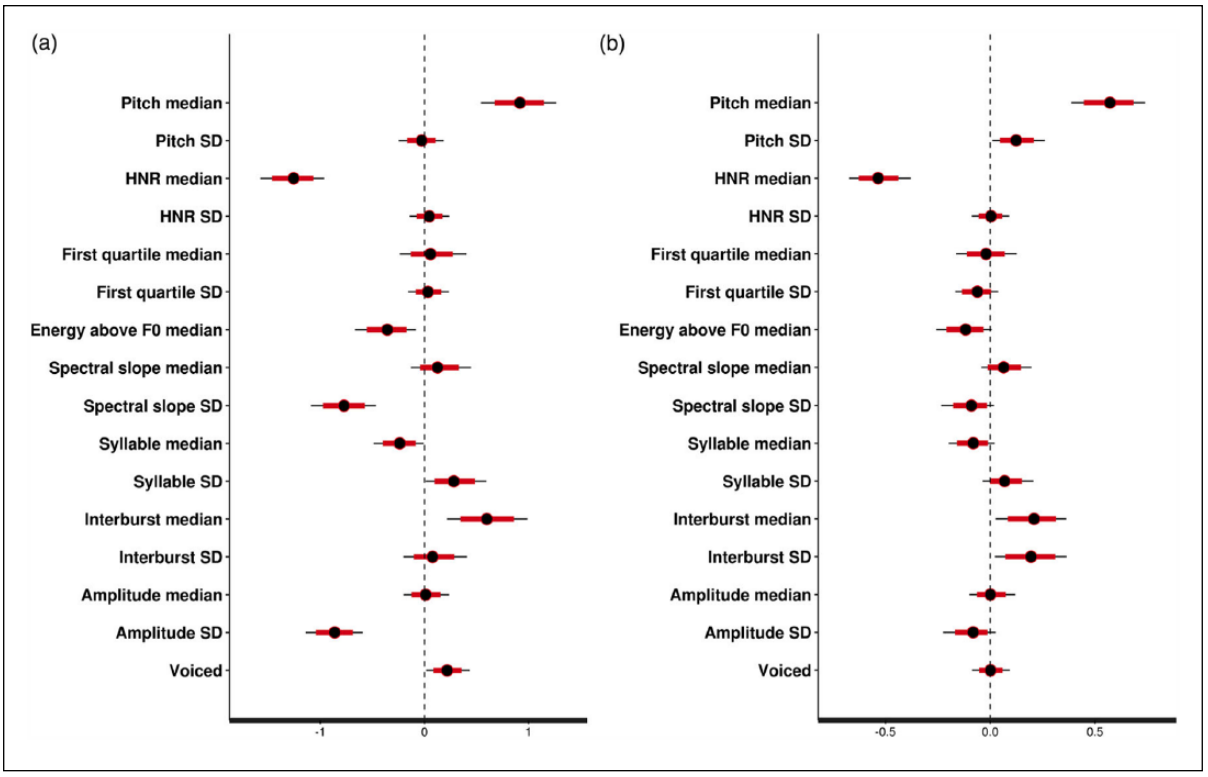


**Figure 4.** Standardized beta coefficients on a logit scale with 80% and 95% confidence interval (CI) in multiple logistic regression. The outcome variable is (a) the status of each sound as authentic or (b) the real/fake subjective judgment provided by participants in the behavioural experiment. Both models employ shrinkage to correct for multiple comparisons (Kruschke, 2014). HNR = harmonics-to-noise ratio.

whether listeners can indeed detect authenticity across a range of emotion categories. Furthermore, the fact that performance remained the same throughout the experiment indicates that (a) our findings cannot be accounted for by a learning effect and that (b) providing participants with feedback regarding their responses did not affect performance.

The accuracy of authenticity detection, however, varied considerably depending on the emotional category. Authentic sounds of fear, anger, and pleasure were much more likely (30–50%) to be perceived as authentic than posed vocalizations. This difference in authenticity appears to be equally large for achievement, although it was hard to obtain unambiguous authentic examples of this emotion, and therefore we present the current result as tentative. In contrast, listeners found it relatively harder to discriminate between authentic and acted sounds of amusement (laughing), sadness (crying), and disgust. This is not surprising in the case of disgust, since this emotion is elicited by relatively slow-acting external stimuli and presents ample opportunities for impression management, making it hard to ascertain whether the obtained "authentic" sounds of disgust really are spontaneous. Laughter is a more interesting case, being the only vocalization whose perceived authenticity has already been systematically tested in previous studies.

Laughs produced in response to seeing something amusing, such as a funny video clip, were previously reported to be judged as authentic 67–75% of the time (Bryant & Aktipis, 2014; Lavan et al., 2015). In the current study, authentic laughs were judged as authentic 67% of the time, although there was a slight bias to label all laughs "real", resulting in overall accuracy of 61% for this emotion. In agreement with previous studies (Bryant & Aktipis, 2014; Lavan et al., 2015), authentic laughs in this study were characterized by higher pitch and pitch variability. However, our findings did not confirm the previous observations that authentic laughs have a higher rate of syllables per second than acted laughs (Vettin & Todt, 2005), shorter syllables (Bryant & Aktipis, 2014), or a lower proportion of voiced frames (Bryant & Aktipis, 2014; Lavan et al., 2015). This discrepancy may be related to specific characteristics of different corpora. In particular, the method of obtaining authentic laughs in this study (recordings of amusing everyday situations) differed from the method employed in the previous studies (elicitation of laughs at the research centre). In future studies it will be of interest to directly compare authentic laughs based on observational material with authentic laughs elicited in a laboratory context. We did not record the age and sex of participants in order to facilitate recruitment, but in future

**Table 5.** Classification accuracy for a Random Forest model trained on acted and tested on authentic sounds, trained on authentic and tested on acted sounds, or trained and tested on both authentic and acted sounds.

| Emotion | Acted → authentic | | Authentic → acted | | Mixed → mixed, half[a] | | Mixed → mixed, all[b] | |
|---|---|---|---|---|---|---|---|---|
| | Training: acted ($n = 362$) | Testing: authentic ($n = 427$) | Training: authentic ($n = 427$) | Testing: acted ($n = 362$) | Authentic ($n = 244$) | Acted ($n = 182$) | Authentic ($n = 427$) | Acted ($n = 362$) |
| Amusement | 87 [84, 88] | 81 [78, 84] | 80 [77, 82] | 82 [82, 82] | 81 [69, 92] | 83 [68, 96] | 84 [81, 86] | 84 [82, 88] |
| Anger | 55 [52, 59] | 42 [36, 50] | 57 [54, 60] | 8 [5, 10] | 59 [45, 72] | 44 [30, 60] | 63 [60, 66] | 46 [40, 50] |
| Disgust | 49 [44, 53] | 43 [37, 48] | 58 [55, 62] | 55 [51, 59] | 56 [39, 71] | 60 [43, 74] | 59 [56, 61] | 62 [59, 67] |
| Fear | 48 [44, 52] | 71 [69, 75] | 69 [66, 71] | 3 [2, 3] | 64 [45, 79] | 38 [21, 52] | 69 [65, 73] | 42 [37, 48] |
| Pain | 41 [31, 51] | 30 [26, 33] | 46 [42, 51] | 50 [50, 50] | 36 [22, 51] | 43 [17, 83] | 42 [38, 46] | 47 [42, 58] |
| Pleasure | 70 [67, 72] | 16 [14, 18] | 50 [46, 55] | 75 [73, 78] | 34 [19, 50] | 73 [57, 90] | 39 [36, 43] | 74 [73, 76] |
| Sadness | 34 [29, 38] | 31 [25, 35] | 56 [52, 60] | 28 [25, 30] | 50 [34, 66] | 35 [19, 50] | 56 [51, 60] | 42 [38, 45] |
| Overall | 54 [53, 56] | 45 [44, 47] | 59 [58, 61] | 39 [38, 40] | 54 [49, 59] | 53 [47, 59] | 59 [58, 60] | 56 [54, 58] |

Note: Classification accuracy in percentages. Average accuracy over 1000 Random Forest models with 1000 decision trees in each and 16 acoustic predictors. Values in square brackets indicate 95% confidence intervals.
[a]Trained on a mixed set of 183 authentic and 180 acted sounds, and tested on the remaining 244 authentic and 182 acted sounds (1000 iterations, with stratified random sampling of sounds for the training set at each iteration). [b]Trained and tested on a the complete set of 427 authentic and 362 acted sounds (cross-validation accuracy).

it might also be relevant to look at potential age-related and sex-related effects, since previous studies found modulations related to age in vocal emotion recognition (Lima, Alves, Scott, & Castro, 2014), as well as sex effects in authenticity detection (McKeown et al., 2015).

For vocalizations other than laughter, higher and more variable pitch and lower harmonicity were important predictors of perceived authenticity. Notably, the higher pitch of authentic vocalizations could not be explained by differences in the number of male and female vocalizations across corpora. It is also unlikely that peculiarities of the voices of individual speakers might have affected the results, since the number of speakers was very high relative to the number of sounds (hundreds of speakers for authentic vocalizations and dozens for acted vocalizations). This is a further strength of the current study compared to typical research in vocal emotional processing, where the number of speakers is often small. By including such a large number of speakers, we largely enhance the generalizability of our findings and decrease the likelihood of speaker-specific effects.

The results of acoustic analysis suggest that high arousal (Banse & Scherer, 1996; Gustison & Townsend, 2015) might be implicated in the detection of authentic emotional displays. The intensity of underlying emotion has previously been reported as an important contributing factor to authenticity judgments. McKeown et al. (2015) argue that the hard-to-fake qualities are only exhibited by high-intensity laughter associated with a genuinely funny episode. Similarly, in the study by Lavan et al. (2015) authenticity and arousal judgments were correlated, so that laughs rated higher on arousal were also perceived as more authentic. On the other hand, there was no correlation between arousal and authenticity ratings in the study by Lima et al. (2013). Research on facial expressions also indicates that participants can discriminate between authentic and acted smiles even when the stimuli are obtained in such a way as to be matched for arousal (e.g., Murphy, Lehrfeld, & Isaacowitz, 2010). We did not ask listeners to rate sounds on arousal, so it is hard to ascertain the extent to which authenticity was perceived independently of arousal in our study. Moreover, a large proportion of variance in human authenticity judgments remained unexplained by the analysed acoustic features. The identified acoustic correlates thus capture only the most salient differences between authentic and acted stimuli, while more subtle distinctions probably exist, and they might affect the perceived authenticity of emotional vocalizations.

Another finding of the current study was that authenticity detection was not affected by being explicitly cued about the expressed emotion or by the native language of the listeners. The emotion of a vocalization is rarely detected with perfect accuracy; for example, authentic sounds of pain, anger, and fear are often confused by listeners (Anikin & Persson, 2016). It seems reasonable to speculate that knowing which emotion the caller actually experienced (or intended to portray) could make it easier to decide whether the vocalization is authentic or posed. However, the difference in perceived authenticity of authentic and acted vocalizations remained the same whether or not participants were told which emotion each sound represented. Furthermore, authenticity judgments were equally accurate whether or not the speaker and the listener spoke the same language. This was the case for both acted and authentic vocalizations, although for the latter the native language of the speaker was not always known, introducing some noise in the analysis. It is well established that vocal emotions are recognized more accurately when the

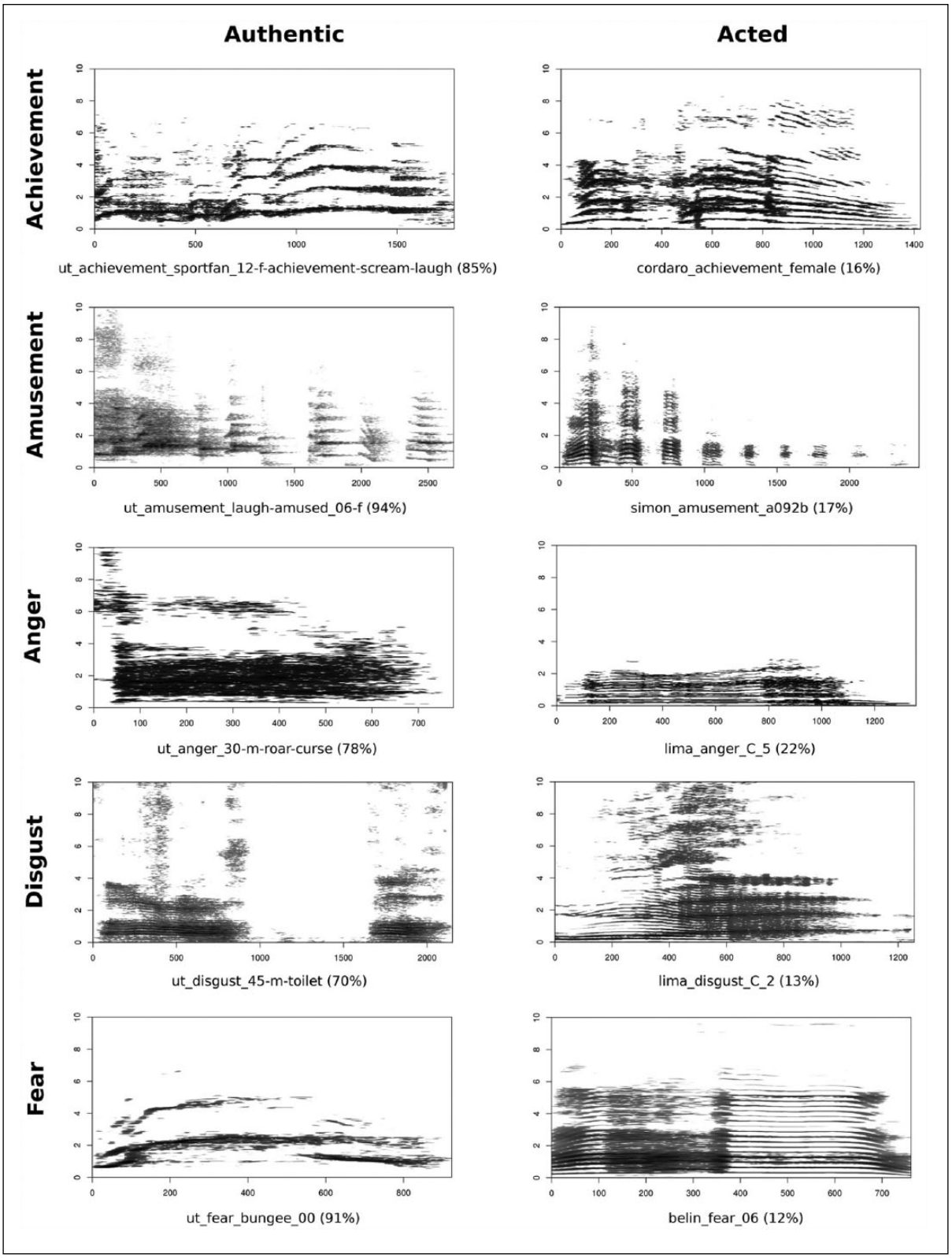


**Figure 5.** Continued.

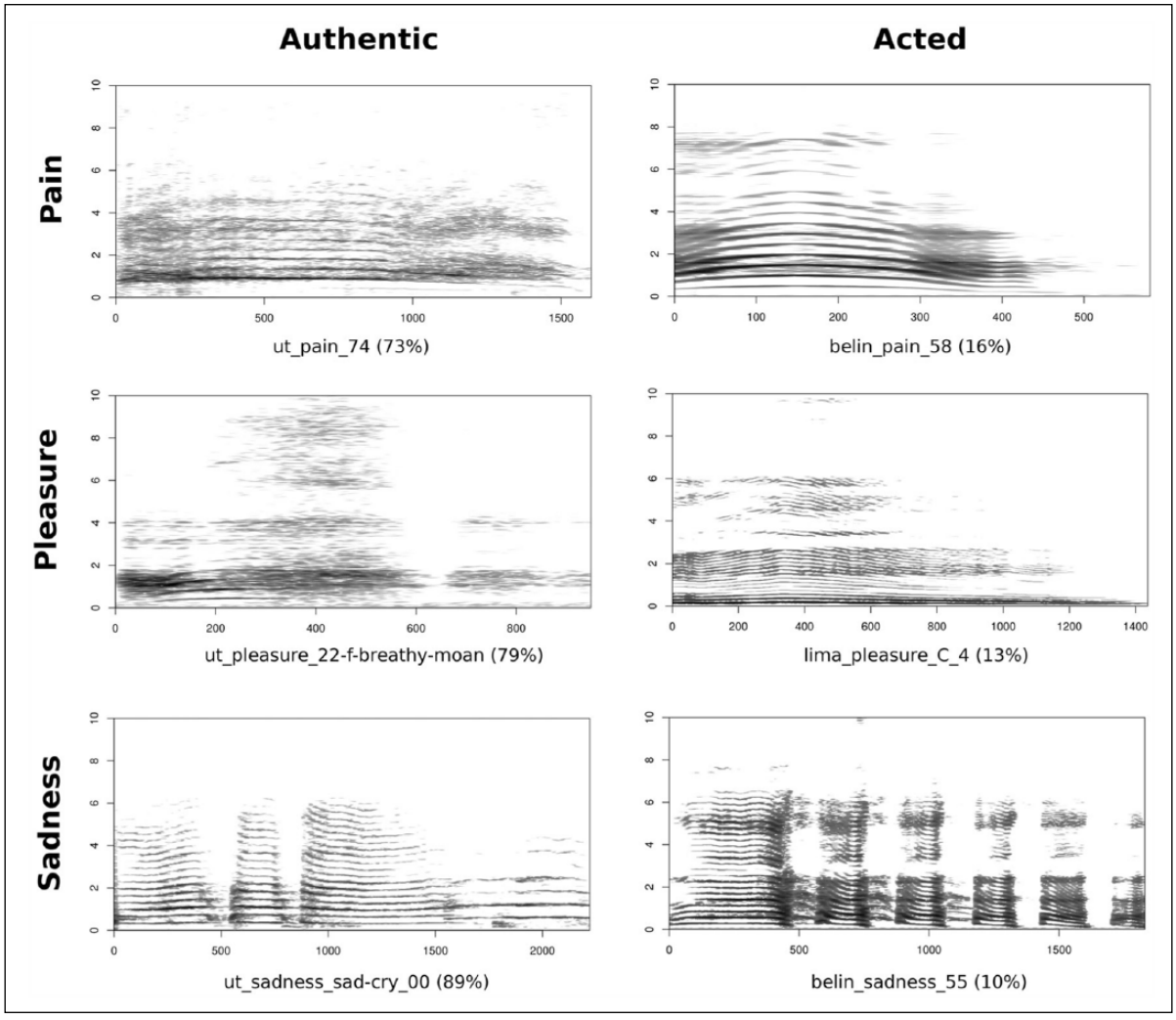


**Figure 5.** Contrast pairs of sounds with high and low perceived authenticity. Shown: emotion, spectrogram (time in ms, frequency in kHz), file name, and proportion of participants who responded "real".

producer and the receiver belong to the same sociocultural group—the so-called "in-group advantage" (Elfenbein & Ambady, 2002; Koeda et al., 2013; Laukka et al., 2013; but compare Anikin & Persson, 2016). Similarly, it might be expected that authenticity should be detected more accurately when the caller and the listener speak the same language. At least for the corpora tested, however, which are mostly free from quasi-verbal and clearly culture-specific interjections such as *Ouch!*, we failed to find a role for shared language in authenticity detection, further suggesting that the authenticity of nonverbal vocalizations is detected independently of their emotion.

In line with previous reports (Bryant & Aktipis, 2014; Lavan et al., 2015), the overall accuracy of authenticity detection in this study was not very high (65%). Besides, the perceived authenticity of both authentic and acted vocalizations varied widely, with a lot of overlap between these two types of expressions (Figure 2), so that some actor portrayals actually sounded more authentic than real-life vocalizations. This suggests that acoustic markers of authenticity may be only moderately salient in everyday interactions. From an evolutionary perspective, however, even minimal acoustic signatures of authenticity may be extremely important, since these hard-to-fake markers of the speaker's affective state distinguish between honest communication and bluff, with implications for the evolution of vocal signals (Searcy & Nowicki, 2005; Zahavi, 1982). The better than chance accuracy of authenticity detection ties in well with the evolutionary argument, particularly since accuracy was highest for "high-stakes" emotions associated with high deception costs, such as fear and anger. In addition, we tested authenticity detection for decontextualized stimuli, because our main goal was to determine whether isolated nonverbal vocalizations per se contain sufficient information to allow listeners to infer authenticity. In everyday interactions, in contrast, we often have multiple cues (e.g., auditory, visual) and rich contextual information, which may enhance the accuracy with which authenticity is detected in real-life interactions.

The discovered acoustic differences between authentic and posed emotional vocalizations also have important practical implications. In future studies it may be desirable to combine actor portrayals with authentic emotional displays in order to achieve a fuller understanding of the complexity and multifaceted nature of human vocal behaviour. In addition, future studies will need to delineate the neurocognitive mechanisms involved in authenticity detection across vocal emotions in order to test whether they are the same as those identified for laugher (McGettigan et al., 2015), or whether they are modulated by emotion category. Our findings also bear on affective computing. As more corpora of emotional vocalizations are becoming available, classifiers are sometimes trained on one corpus and tested on another to evaluate their generalizability (e.g., Petridis et al., 2015). However, since so few collections of authentic vocalizations are available for machine learning, the question remains whether even the best learning algorithms will prepare computers for recognizing the users' emotion in real-life human–machine interaction.

We have demonstrated, for the first time, that machine learning algorithms achieved robust transfer between authentic and acted vocalizations, in both directions, for amusement, disgust, and pain. In contrast, transfer was considerably more problematic for anger, fear, pleasure, and sadness, depending on which set was the training one. The previously untested assumption that a classifier trained on actor portrayals will be ready to deal with real-life emotional expressions is thus not always warranted. In contrast, an algorithm trained on a mix of authentic and acted vocalizations was successful at classifying both. Access to data banks of both naturalistic and posed vocalizations may thus be beneficial for optimizing real-life performance of automated systems for affect recognition. Finally, building on an emerging body of work (Dai et al., 2015; Parsons et al., 2014), this study has demonstrated the feasibility of using social media as a source of material for vocal emotion research, which opens up exciting new avenues for research.

## Acknowledgements

We would like to thank Emiliana R. Simon-Thomas, Skyler T. Hawk, and Disa Sauter, who kindly made their sounds available for analysis. Pierre Maurage and researchers from his team translated the experiment into French and suggested the control condition with masking noise. Tomas Persson, Christian Balkenius, and Carin Graminius provided many useful comments throughout the project. We are also grateful to our participants for volunteering their time.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Supplementary material

Supplementary material is available at journals.sagepub.com/doi/suppl/10.1080/17470218.2016.1270976.

## References

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, *37*(6), 529–546.

Anikin, A., & Persson, T. (2016). Non-linguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior Research Methods*, *49*: 758–771.

Ba, H., Yang, N., Demirkol, I., & Heinzelman, W. (2012, August). *BaNa: A hybrid approach for noise resilient pitch detection*. Statistical Signal Processing Workshop (SSP), 2012 IEEE (pp. 369–372).

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*(3), 1581–1597.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

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

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.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*(1), 803–832.

Boersma, P. (1993). *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. Proceedings of the institute of phonetic sciences (*Vol. 17*, No. 1193, pp. 97–110).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior*, *35*(4), 327–335.

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, *16*(1), 117–128.

Dai, W., Han, D., Dai, Y., & Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management*, *52*(7), 777–788.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*(1), 33–60.

Drolet, M., Schubotz, R. I., & Fischer, J. (2012). Authenticity affects the recognition of emotions in speech: Behavioral and fMRI evidence. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 140–150.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235.

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.

Gervais, M., & Wilson, D. S. (2005). The evolution and functions of laughter and humor: A synthetic approach. *The Quarterly Review of Biology*, *80*(4), 395–430.

Gustison, M. L., & Townsend, S. W. (2015). A survey of the context and structure of high- and low-amplitude calls in mammals. *Animal Behaviour*, *105*, 281–288.

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion*, *9*(3), 293–305.

Hewson, C., Vogel, C., & Laurent, D. (2016). *Internet research methods* (2nd ed.). London: Sage.

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

Jürgens, R., Grass, A., Drolet, M., & Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of Nonverbal Behavior*, *39*(3), 195–214.

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

Kipper, S., & Todt, D. (2001). Variation of sound parameters affects the evaluation of human laughter. *Behaviour*, *138*(9), 1161–1178.

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

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, *5*(1), 41–46.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London: Academic Press.

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

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

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), 145–160.

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.

Maurage, P., Joassin, F., Philippot, P., & Campanella, S. (2007). A validated battery of vocal emotional expressions. *Neuropsychological Trends*, *2*(1), 63–74.

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.

McKeown, G., Sneddon, I., & Curran, W. (2015). Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions. *Emotion Review*, *7*(1), 30–38.

Murphy, N. A., Lehrfeld, J. M., & Isaacowitz, D. M. (2010). Recognition of posed and spontaneous dynamic smiles in young and older adults. *Psychology and Aging*, *25*(4), 811–821.

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

Petridis, S., Pantic, M., Rudovic, O., Pantic, M., Patras, I., Liwicki, S., ... Bilakhia, S. (2015). Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. *IEEE Transactions on Affective Computing*, *23*, 1624–1636. doi:10.1109/TAFFC.2015.2446462

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, *106*(2), 332–337.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2251–2272.

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, *27*(1), 40–58.

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

Scott, S. K., Lavan, N., Chen, S., & McGettigan, C. (2014). The social life of laughter. *Trends in Cognitive Sciences*, *18*(12), 618–620.

Scott, S. K., Sauter, D., & McGettigan, C. (2009). Brain mechanisms for processing perceived emotional vocalizations in humans. In S. Brudzynski (Eds.), *Handbook of mammalian vocalization: An integrative neuroscience approach* (Vol. 19, pp. 187–197). London: Academic Press.

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

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9*(6), 838–846.

Vettin, J., & Todt, D. (2005). Human laughter, social play, and play vocalizations of non-human primates: An evolutionary approach. *Behaviour*, *142*(2), 217–240.

Wattendorf, E., Westermann, B., Fiedler, K., Kaza, E., Lotze, M., & Celio, M. R. (2013). Exploration of the neural correlates of ticklish laughter by functional magnetic resonance imaging. *Cerebral Cortex*, *23*(6), 1280–1289.

Zahavi, A. (1982). The pattern of vocal signals and the information they convey. *Behaviour*, *80*(1), 1–8.