Human nonverbal vocalizations

ANDREY ANIKIN
COGNITIVE SCIENCE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

Faculties of Humanities and Theology
Department of Philosophy
Cognitive Science

Lund University Cognitive Studies 178
ISBN 978-91-88899-81-1
ISSN 1101-8453

Human nonverbal vocalizations

Andrey Anikin

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Humanities, Lund University, Sweden.
To be defended at SOL, room H104, Lund, on February 28, 2020 at 10:00.

Faculty opponent
Tecumseh Fitch

Organization LUND UNIVERSITY Cognitive Science Department of Philosophy Author(s) Andrey Anikin	Document name **Doctoral dissertation** **Date of issue: February 28, 2020** Sponsoring organization

Title and subtitle: Human nonverbal vocalizations

Abstract

Language is a very special ability, but human communication also includes a wealth of nonverbal signals: body language, facial expressions, and nonverbal vocalizations such as laughs, moans, and screams. Vocalizations are particularly interesting because they share the same modality as language but are more similar in function and structure to the calls of non-human animals. Accordingly, this thesis is an attempt to study human nonverbal vocalizations from a comparative and evolutionary perspective in order to explore the nonverbal repertoire and to understand how information is encoded in these signals.

While nonverbal vocalizations are typically obtained by asking participants to portray a particular emotion, a less structured observational approach is explored in Paper I. By collecting unscripted examples of nonverbal vocalizations from the social media, it may be possible to obtain a more representative sample of vocal behaviors, which are also judged to be more authentic compared to actor portrayals (Paper II). Moreover, when each sound is not intended to convey a single emotion, it becomes more obvious that the repertoire of nonverbal vocalizations consists of several perceptually distinct acoustic classes as well as intermediate variants (Paper III). This means that, like other mammals, humans have a limited number of species-typical call types. These fundamental acoustic categories are the building blocks of nonverbal communication, but their acoustic properties also inform the intonation and other prosodic features of spoken language.

Nonverbal vocalizations are interpreted flexibly in real-life interactions, taking into account the accompanying facial expression and other contextual information. To learn what information is available in the sound itself, it is desirable to be able to modify individual acoustic properties and to observe how the listeners' responses change as a result. A new method of voice synthesis is proposed in Paper IV and then used to test the perceptual effects of manipulating two aspects of voice quality: nonlinear vocal phenomena (Paper V) and breathiness (Paper VI). In addition to shedding new light on the acoustic code involved in nonverbal vocalizations, Papers V and VI confirm the importance of distinguishing between call types because the meaning of the same acoustic property – for example, voice roughness – can vary depending on the type of vocalization in which it occurs.

A red thread going through this dissertation is that humans are mammals and vocalize like mammals despite being linguistic creatures. The structure of the vocal repertoire and the general principles of voice modulation are broadly similar across many animal species, including humans. One reason for this convergence may be the existence of wide-spread crossmodal correspondences such as the tendency to associate low frequencies with a large body size. In Paper VII, I propose another possible cognitive mechanism for some non-arbitrary acoustic properties associated with intense emotion in humans and other species. In the case of human nonverbal vocalizations, high-intensity calls possess all the acoustic properties associated with bottom-up auditory salience – that is, these sounds appear to be "designed" to attract the listeners' attention. This may be the result of vocal production and perception coevolving, or it may mean that the acoustic structure of high-intensity vocalizations exploits preexisting perceptual biases.

To summarize, knowing the evolutionary history and cognitive mechanisms behind vocal behaviors, such as human nonverbal vocalizations studied in this dissertation, provides a deeper understanding of their role in communication.

Key words: nonverbal, communication, acoustic, emotion

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language: English**
ISSN 1101-8453 Lund University Cognitive Sciences 178		**ISBN** 978-91-88899-81-1
Recipient's notes	**Number of pages**	Price
	Security classification	

Signature  Date 2020-01-14

Human nonverbal vocalizations

Andrey Anikin

LUND
UNIVERSITY

Faculties of Humanities and Theology
Cognitive Science
Department of Philosophy

ISBN 978-91-88899-81-1
ISSN 1101-8453 Lund University Cognitive Sciences 178

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

To hope for a better future that makes a better future possible

Acknowledgments

Many people have contributed to this work and to my PhD experience. I can't mention everything and everyone here, but I'm deeply grateful for all the help, encouragement, and support!

I would like to begin by acknowledging the cataclysmic impact of Rasmus Bååth, who introduced me to R and Bayesian statistics while I was still doing my Master's in cognitive science. Dewey-eyed and impressionable, I became an instant convert and have since spent at least half of my working hours in RStudio. Niklas Johansson has been a constant intellectual companion and a terrific collaborator on three papers. Tomas Persson, Manuel Oliva, Kerstin Gidlöf, Annika Wallin, Peter Gärdenfors, Nikolai Aseyev, and many others were always a pleasure to write papers with and taught me a lot about research. Cesar Lima and Ana Pinheiro have been great collaborators through the years, although we have never met in person – I hope we will soon!

I would also like to thank my supervisors Christian Balkenius and Tomas Persson. Tomas supervised me already during my time as a Master's student, and both he and Christian gave me a lot of encouragement and completely free reins to explore all kinds of research ideas, no matter how far-out. The department of Cognitive Science in general has provided a lot of support over the years and funded my experimental work – thanks to Anna Cagnan Enhörning for her infinite patience with my clumsy paperwork! I am particularly grateful for the generous scholarship from Vitterhetsakademien that financed my stay at the University of Sussex.

Speaking of Sussex, it was a very special opportunity to be a guest researcher there and to work with Karen McComb, who has my warmest gratitude for inviting me (which turned out to involve an awful lot of red tape), providing unstinting support and guidance, and introducing me to her gorgeous cat Kimi. Chris and Kate Darwin were kind enough to take me in together with my family and were perfect hosts. My warm regards also to the rest of the Sussex team: Tazmin, Anna, Val, Holly, and everyone else! Special thanks to Karen Hiestand for introducing me to life in the British countryside and for all the enlightening discussions, from charities to cattle welfare.

Ever since my stay in Sussex, it has been my privilege to work with David Reby and Kasia Pisanski, and the week spent with them in St. Etienne was among the most intense and exciting periods of my PhD. I was also fortunate to meet and work with other outstanding researchers, particularly during the unforgettable week-long Dagstuhl retreat with the VIHAR group: Dan Stowell, Elodie Briefer, Nick Campbell, Roger Moore, and many others.

On a more personal note, the open, friendly, and pressure-free environment at Lund Cognitive Science has always been a source of great comfort. The relaxed atmosphere with a strong sense of camaraderie has given me and other students the confidence to explore and grow without any of the squabbles and rivalries that often mar academic departments. Manuel Oliva and Zahra Gharaee were always quick to lighten the mood, not to mention our great paddling adventure and exchange of spoofs. So many laughs every day! Trond Tjöstheim, Andreas Stephens, Can Kabadayi, Megan Lambert, Betty Tärning, Kristin Ingvarsdottir, Stephan Reber, Tobias Mahlmann, Emmanuel Genot and many others were always there to discuss the latest experiment with or just to chat over lunch. Trond, thank you for all the profound discussions and outings, starting already in the good old days of the Master's program, and for keeping my neuroscience sharp together with Kalle Palm and Sophie Nielsen! Stephan, I owe to you my first-hand experience of working with alligators – and of building an alligator facility! Emmanuel, thank you for teaching me that exercising is an exact science!

My family and friends all over the world have provided constant emotional support and piloted all my experiments. I am particularly grateful to them and to many other volunteers for taking part in my earlier, crowd-sourced studies. Special thanks to my wife and son for inspiring my research with their rich repertoire of nonverbal vocalizations!

Finally, it is easy to give encouragement and support when everything is working smoothly, and much harder to do so when there are disagreements and trouble. I would therefore like to end by thanking Rita Astuti, my supervisor during my year as a PhD student in anthropology at the LSE, who helped me find my way with great integrity and tact. Our Friday seminars at the LSE made that year worthwhile and led me directly to cognitive science, and your support made it possible to start over.

So thank you, one and all!

Table of Contents

List of original papers

Paper I

Anikin, A. & Persson, T. (2017). Non-linguistic vocalizations from online amateur videos for emotion research: a validated corpus. *Behavior Research Methods, 49*(2), 758-771. © Psychonomic Society, Inc. 2016.

Paper II

Anikin, A. & Lima, C. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology, 71*(3), 622-641. © SAGE Publications 2018.

Paper III

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: call types and their meaning. *Journal of Nonverbal Behavior, 42*(1), 53-80. © The Authors 2017.

Paper IV

Anikin, A. (2019). Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behavoir Research Methods, 51*(2), 778-792. © The Author 2018.

Paper V

Anikin, A. (2019). The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics*.
doi: 10.1080/09524622.2019.1581839. © The Author 2019.

Paper VI

Anikin, A. (in press). A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica.* © 2019 S. Karger AG, Basel.

Paper VII

Anikin, A. (in review). The link between auditory salience and emotion intensity in human nonverbal vocalizations. © The Author 2019.

Other papers by the author not included in the thesis

Gidlöf, K., Anikin, A., Lingonblad, M. & Wallin, A. (2017). Looking is buying. How visual attention and choice are affected by consumer preferences and properties of the supermarket shelf. *Appetite, 116*, 29-38.

Oliva, M. & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports, 8*(1), 4871.

Lima, C., Anikin, A., Monteiro, A., Scott, S., & Castro, S. (2019). Automaticity in the recognition of nonverbal emotional vocalizations. *Emotion, 19*(2), 219-233.

Anikin, A. & Johansson, N. (2019). Implicit associations between individual properties of color and sound. *Attention, Perception, & Psychophysics, 81*(3), 764–777.

Pinheiro, A., Lima, D., Albuquerque, P., Anikin, A., & Lima, C. (2019). Spatial location and emotion modulate voice perception. *Cognition & Emotion, 33*(8), 1577-1586.

Johansson, N., Anikin, A., & Aseyev, N. (2019). Color sound symbolism in natural languages. *Language & Cognition*, 1-28.

Amorim, M., Anikin, A., Mendes, A., Kotz, S., Lima, C., & Pinheiro, A. (2019). Changes in vocal emotion recognition across the life span. *Emotion*, 1-11.

Johansson, N., Anikin, A., Carling, G., & Holmer, A. (in press). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*.

1. Nonverbal vocalizations as a form of communication

This thesis is about a particular form of communication, namely human nonverbal vocalizations – that is, any voiced sounds that we communicate with and that are not speech: laughs, screams, moans, cries, etc. This definition excludes primarily physiological sounds, such as burps and sneezes, as well as emblems with some language-specific phonemic structure, such as *Ouch* and *Wow*. Nonverbal vocalizations in infants are abundant and extensively studied (Green, Whitney, & Potegal, 2011; Koutseff et al., 2018; Lingle, Wyman, Kotrba, Teichroeb, & Romanow, 2012; Scheiner, Hammerschmidt, U. Jürgens, & Zwirner, 2002; Zeifman, 2001), but their relationship to the adult repertoire is complex, and in this thesis I mainly focus on adults. Although the target species is *Homo sapiens*, the signals being studied have more in common with vocalizations of non-human animals (hereafter, simply "animals") than with language. I therefore approach these vocalizations from a comparative perspective, aiming to understand how they contribute to communication and, more broadly, how human vocal behavior is informed by our phylogenetic history. The purpose of the first chapter is to make explicit the general theoretical framework for this investigation, situating human nonverbal vocalizations in relation to language and animal calls. I then formulate research questions (section 1.3) and discuss the papers in relation to these questions (sections 2 and 3).

Definitions of communication vary depending on the field of inquiry. If the main focus is on language, it seems intuitive to use the "conduit metaphor" (Lakoff & Johnson, 2008[1980]) and to conceptualize communication as transfer of information from the sender to the receiver. Successful communication, according to this classical view, enables the receiver to reconstruct the mental representations that the sender intended to convey. Biology, on the other hand, supplies many examples of communicative interactions in which both production and perception of signals are too direct to plausibly involve mental representations, intentionality, or advanced cognitive processing. Accordingly, biological theories of communication may prefer to eschew the concept of information and to define communication as the process of altering the receiver's behavior via evolved mechanisms (Rendall, Owren, & Ryan, 2009; Stegmann, 2013).

In this thesis I combine elements of both approaches, building the argument on an evolutionary foundation while preserving the concept of information and meaning as central to describing communication. As argued by Fischer (2011), the effect – or meaning – of a signal depends on the receiver's set of sensory organs, cognitive architecture, and unique life history (such as human cultural environment). Accordingly, the informational content of a signal is best treated not as an intrinsic property of the signal itself, but as a product of its interaction with a particular receiver in a particular context. With this proviso, communication can be defined as exchange of information via an evolved (for biological systems) or designed (for artificial systems) mechanism, where information corresponds to potential reduction in uncertainty about the state of the world (Fischer, 2011; Wheeler & Fischer, 2012).

At the same time, the language-inspired notion of communication as the process of intentionally transferring a mental representation from the sender to the receiver via a symbolic code represents only the tip of the iceberg – a highly specialized form of communication that is rather unusual in the natural world and that does not cover all kinds of human nonverbal signals. Instead, a useful starting point for studying human nonverbal communication may be to specify the various cognitive mechanisms involved in the production and perception of all communicative signals, from fairly direct to the most cognitively sophisticated. These mechanisms are listed in sections 1.1 and 1.2, treating production and perception separately and using examples of both human and animal signals throughout, so as to emphasize that these levels of cognitive sophistication are not about "us versus them" but are found in many animals, including humans. The proposed classification of production and perception mechanisms is functional: the main focus is on how communication works on an algorithmic level (Marr, 1982) rather than on the exact computational mechanisms or their localization in the brain.

1.1 Signal production

To begin with the sender, a communicative signal can be produced in many different ways. In this review I distinguish between the following types of signals based on their production mechanisms: long-term somatic features such as sexual ornaments; transient signals whose form and eliciting context are innate; innate signals that are produced more flexibly or intentionally; and finally, socially learned signals. Throughout the text, a signal is considered to be innate if it is predictably displayed by all members of the species without the need for learning.

1.1.1 Somatic signals

The least cognitively demanding communicative signals require neither learning nor a conscious intention to be produced – in fact, they may not even require a brain. There are numerous somatic signals – long-term modifications of the signaler's body that evolved in order to inform other organisms about the fitness, age, sex, and social status of the signaler. For example, males of many animal species possess ornaments such as antlers in deer, large tail feathers in peacocks, brightly colored spots in fishes, and so on. These decorations evolve via sexual selection driven by male competition and female preferences. In many cases the ornaments are not only perceptually salient, but also metabolically expensive or endangering; by growing and maintaining them, males can simultaneously advertise and prove their own fitness. The high cost ensures that the resulting communication is hard-to-fake and thus "honest", which is often referred to as the handicap principle (Zahavi, 1975). More generally, honest signaling can be maintained despite some conflict of interest between the signaler and the receiver if the cost of producing a signal depends on fitness, so that production is more expensive for individuals of poor quality, or if the signal conveys the level of need rather than fitness (Searcy & Nowicki, 2005).

There is often some wiggle room that makes it possible to exploit perceptual biases by exaggerating a trait without fully undermining its status as an honest fitness indicator. For instance, vocal tract length is readily perceived from the spacing of resonance frequencies (formants), and together with the rate at which vocal folds vibrate (fundamental frequency, which is perceived as pitch) formant spacing can serve as an indicator of the overall body size. Because males in non-monogamous species are under pressure to appear as large as possible in order to intimidate rival males and impress females, in some species adaptations have evolved to exploit the low-is-large perceptual bias. One mechanism of acoustic size exaggeration is to produce loud low-pitched calls using anatomical adaptations such as fleshy pads on the vocal folds of roaring cats, hypertrophied larynges in howler and colobus monkeys, or an additional set of non-laryngeal vocal folds in koalas. Another method is to extend the vocal tract by growing mobile larynges or additional resonators such as nasal proboscises or air sacs (Charlton & Reby, 2016). Because there are usually anatomical limits on how far acoustic size exaggeration can be pushed, the resulting signals still preserve a correlation with the actual body size and remain useful as fitness indicators. For example, the mobile larynx in deer stags cannot descend below the sternum, so the vocal tract length at full extension provides honest information about the animal's age and size as well as his stamina (Reby & McComb, 2003).

Sexual selection in humans is an object of lively and occasionally sensationalist debates, but it does furnish excellent examples of somatic signals in humans. For example, it is possible that the descended larynx and beard in males were driven

by female preferences and male competition in the context of attempting to exaggerate the apparent body size (Fitch, 2018; Puts, 2010). More generally, sexual dimorphism in the structure of the human vocal tract is larger than expected from the overall difference in body size and more extreme than in any other living ape (Aung & Puts, 2019; Puts et al., 2016), suggesting strong sexual selection in the hominin line. While men are on average just 10% taller and 20% heavier than women (Miller, 2011), their vocal folds become enlarged at puberty, lowering the average pitch in relaxed male speech a full octave below female voices (Puts et al., 2016). Furthermore, high levels of testosterone at puberty cause a gradual descent of the larynx in boys, which makes the vocal tract about 20-25% longer in men (Simpson, 2009), dramatically lowers formant frequencies, and further enhances the impression of large size (Puts et al., 2016). In turn, lower pitch and formant frequencies in men have been shown to affect both female preferences and the perceived dominance in the context of male competition (Feinberg, Jones, Little, Burt, & Perrett 2005; Fraccaro et al., 2013; Puts, Gaulin, & Verdolini, 2006).

Because of these profound differences between male and female voices and the sex-specific selective pressures that must have produced them, research on human vocal behavior often includes comparisons between male and female vocal behavior and perception (e.g., Charlton, Taylor, & Reby, 2013), and some studies are designed to test sex-specific acoustic hypotheses, particularly in the context of mate choice and dominance (e.g., Evans, Neave, & Wakelin, 2006). To reiterate, this sexual dimorphism is caused by a permanent, hormonally controlled modification of the vocal tract, which can be viewed as a somatic communicative signal and understood in evolutionary terms. Of course, the operation of sexual selection is not limited to static signals such as the permanently descended larynx in men: dynamic voice modulation in both sexes can often be analyzed from the perspective of body size exaggeration. Furthermore, complex behavioral traits, such as songs of oscine birds or roaring contests of deer stags (Reby et al., 2005), also evolve to regulate mating. Likewise, it has been suggested that such uniquely human abilities as music and language (Fitch, 2010; Miller, 2011) were affected by sexual selection. Evolutionary forces thus affect all kinds of communicative signals, regardless of their production mechanism.

1.1.2 Innate form, innate context

Moving on from hormonally triggered, long-term somatic features to transient signals whose production is rapid and controlled by the brain, many of these signals are innate in terms of both the form of the signal and the context of its production. For example, worker ants returning from a food site lay down a pheromone trail, which helps to recruit and guide other workers, who in turn strengthen the trail with fresh pheromone markers until the food supply is

exhausted. By using several types of attractant and repellent pheromones with varying half-life, ants can coordinate the behavior of the entire colony in an adaptive and highly flexible manner (Jackson & Ratnieks, 2006). Despite the complexity of the resulting behavior, however, the physical form of the signal (the choice of a particular pheromone) and the timing of its expression appear to be determined by if-then rules that leave little room for learning, context, or conscious intentions.

Innate and relatively inflexible signals are by no means unique to invertebrates – on the contrary, a large proportion of animal signals fall into this category. For example, the basic structure of most primate vocalizations and many gestures is genetically determined or "production-first" (Owren, Amoss, & Rendall, 2011; Seyfarth & Cheney, 2018; Snowdon, 2009), and each expression is associated with a range of typical eliciting contexts. In humans, congenitally hearing-impaired infants laugh and cry in a manner similar to hearing infants (Scheiner, Hammerschmidt, U. Jürgens, & Zwirner, 2006). Furthermore, even anencephalic human infants and decerebrated animals are capable of crying (Newman, 2007). This indicates that the appropriate motor programs (a coordinated activity of the diaphragm and muscles of the larynx) are species-typical behaviors that are encoded in the brain stem, mature without auditory feedback, and are executed when triggered by a predetermined eliciting context: social play and tickling for laughs (van Hooff & Preuschoft, 2003), separation for infant cries (Newman, 2007), and so on. Nor do we grow out of such innate signaling as adults: if suddenly frightened, most people will scream and display the classical primate fear face before being able to monitor or suppress this involuntary reaction (Paper I). In neurological terms, phylogenetically conservative circuitry for the production of species-typical signals in relatively narrow, predetermined contexts remains operative even in organisms endowed with a strong capacity for social learning and intentional control, including humans. Human vocal production is thus under dual neural control: the limbic pathway is responsible for triggering species-typical vocalizations, while the motor-cortical pathway enables direct voluntary control over vocalizing (Ackermann, Hage, & Ziegler, 2014; U. Jürgens, 2009). In fact, speech prosody has many similarities with nonverbal vocalizations (section 1.1.3), suggesting that, as language was evolving in our ancestors, it built upon the phylogenetically older mammalian vocalization system and had to remain compatible with it (Fitch, 2010, Ch. 4).

In sum, nonverbal vocalizations such as laughs and screams are prime examples of innate, species-typical vocal behaviors in humans (Sauter et al., 2019), and they have clear parallels in our primate relatives (Lingle et al., 2012; McCune, Vihman, Roug-Hellichius, Delery, & Gogate, 1996; Newman, 2007; Ross, Owren, & Zimmermann, 2009). An important corollary is that these vocalizations are very similar in different human cultures (Cordaro, Keltner, Tshering, Wangchuk, &

Flynn, 2016; Sauter, Eisner, Ekman, & Scott, 2010), providing a kind of nonverbal Esperanto that has no doubt facilitated cross-cultural contacts throughout history. Limited cross-cultural variation does not in itself prove innateness, but developmental studies in hearing-impaired individuals coupled with neurological research provide much stronger evidence, demonstrating that at least some of nonverbal vocalizations are part of our species-typical repertoire.

1.1.3 Innate form, flexible context

Whereas ants laying pheromone tracks or infants laughing when tickled appear to follow simple *if-then* rules, other species-typical signals can be deployed with varying degrees of flexibility. For example, learning has some role in determining the context in which vervet monkeys produce the aerial alarm call. While young monkeys initially produce the eagle alarm call to any disturbance in the air, such as falling branches and harmless birds, they gradually learn which species of raptors are particularly dangerous and call only when they spot those (Seyfarth, Cheney, & Marler, 1980). The acoustic structure of the call itself is innate; what's more, there is a strong predisposition to produce this alarm call to threats from above rather than to terrestrial predators like leopards or snakes, for which vervet monkeys use different alarm calls. Learning serves to fine-tune the eliciting context, but the occurrence of alarm calls and their structure remain predictable.

In comparison, calls of chimpanzees are less context-specific, even if their acoustic structure is innate, and some calls may even be produced with intention to inform. For example, chimpanzees appear to produce more alarm calls when conspecifics are not aware of the threat (Crockford, Wittig, Mundry, & Zuberbühler, 2012), and they may be able to inhibit the production of food grunts when it would be disadvantageous to disclose this information to others (Zuberbühler, 2015), although this inhibition appears to be effortful and is not always successful (Goodall, 1986). Complex audience effects of this type, as well as maintaining eye contact and using a variety of vocalizations and gestures until the desired response is obtained, further suggest that apes can communicate intentionally (Leavens & Hopkins, 1998).

It is important to point out that the same signal can be produced with varying degrees of flexibility or intentional control. The question of intentionality in animal communication is fraught with difficulty (Manser, 2013; Wheeler & Fischer, 2012). For humans, however, it is well established that nonverbal vocalizations and facial expressions can be produced spontaneously, as when laughing at something amusing or showing a genuine, Duchenne smile (Ekman, Davidson, & Friesen, 1990), but they can also be used in a more controlled fashion, as when smiling or chuckling politely on social occasions. Interestingly, different neural circuits appear to be involved depending on whether an emotional

expression like a laugh is produced spontaneously or volitionally (Scott, Lavan, Chen, & McGettigan, 2014). Because of these neurological differences in production mechanisms, there are relatively subtle, but detectable differences between spontaneous and volitional facial expressions (Ekman et al., 1990) and vocalizations (Paper II), indicating that at least some markers of genuine affect may be hard to fake and thus relatively "honest". The crucial point is that this honesty stems precisely from imperfect intentional control. The less the context of production is open to manipulation, the more reliably the signal expresses the true mental state of the sender. As the amount of flexibility increases, the signal can potentially express a wider range of meanings (Wheeler & Fischer, 2012), but it also places a greater burden on the receiver, who now has to take into account the broader context, and possibly also the reputation of the sender, since the "honesty" of the message is no longer guaranteed.

Some aspects of language also belong in the category of innate signals with relatively flexible usage. Emotional prosody in spoken language shows strong similarities around the world (Banse & Scherer, 1996; Bryant & Barrett, 2008; Paulmann & Uskul, 2014; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009), making it straightforward to tell whether a speaker of an unfamiliar language is angry, happy, or sad. The accompanying changes in voice quality, rate of speaking, intonation and other acoustic features are partly derived from the even more universal nonverbal vocalizations (Paper I; Cordaro et al., 2016). For instance, although it is relatively uncommon for humans to produce purely nonverbal, animal-like roars, expletives are often yelled out with an intensity and voice quality characteristic of true roars (Paper I). In addition to emotional prosody, spoken language utilizes a number of largely universal grammatical markers, such as rising intonation in questions (Ohala, 1984), as well as interjections like *Huh?*, which are also similar in many languages (Dingemanse, Torreira, & Enfield, 2013). While their usage is flexible and subject to intentional control, the form of these signals appears to be constrained by the need to conform to the repertoire of communicative signals that humans are genetically endowed with.

1.1.4 Learned form

Signals with a completely arbitrary, purely learned form are uncommon in the natural world. The most obvious example is language, although even language is now regarded as less arbitrary than originally thought due to the widespread presence of onomatopoeia (direct sound imitation such as *meow*) and other forms of sound symbolism in basic vocabulary (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Johansson, Anikin, Carling, & Holmer, in press). In the animal world, the gestural repertoire of great apes is often considered to be more flexible

than their vocalizations (Arbib et al., 2008; Genty, Clay, Hobaiter, & Zuberbühler, 2014). Although the role of social learning in the acquisition of gestures by free-living apes appears to be limited (Genty, Breuer, Hobaiter, & Byrne, 2009; Hobaiter & Burne, 2011; but see Fröhlich, Müller, Zeiträg, Wittig, & Pika, 2017), all species of great apes can be taught to understand and produce hundreds of signs from the American sign language. The grammatical structure of their sentences remains relatively impoverished (Terrace, Petitto, Sanders, & Bever, 1979), but rigorous testing has confirmed that they do understand the meaning of the signs and can produce them appropriately, not only to obtain reward but also to request information or inform others of their intended course of action (Rumbaugh & Savage-Rumbaugh, 1994).

The work with language-trained apes and parrots (Pepperberg, 2006) provides the most convincing examples of intentional use of symbolic signals by non-human animals, but non-symbolic socially learned signals are not uncommon in the natural world. Vocal dialects have now been reported not only among songbirds, but also in some marine mammals (Deecke, Ford, & Spong, 1999; Rendell & Whitehead, 2003) and bats (Prat, Azoulay, Dor, & Yovel, 2017). Less pronounced dialectal variation may also be present in great apes such as chimpanzees (Crockford, Herbinger, Vigilant, & Boesch, 2004). Once learned, however, dialectal vocalizations may well be produced without intention to inform and with only limited sensitivity to context, placing them closer to the relatively inflexible signals discussed above.

As for human nonverbal vocalizations, their basic acoustic structure appears to be species-typical, with relatively minor differences between different cultures (Cordaro et al., 2016; Sauter, Eisner, Ekman, et al., 2010). However, a number of studies have demonstrated a small in-group advantage – that is, an improvement in the accuracy of recognizing the emotion conveyed by both speech prosody (Bryant & Barrett, 2008; Elfenbein & Ambady, 2002; Neiberg, Laukka, & Elfenbein, 2011; Scherer, Banse, & Wallbott, 2001) and nonverbal vocalizations (Elfenbein & Ambady, 2002; Gendron, Roberson, van der Vyver, & Barrett, 2014; Koeda et al., 2013; Laukka et al., 2013; Sauter, Eisner, Ekman, et al., 2010; Sauter & Scott, 2007) when the speaker and listener belong to the same linguistic and cultural group. Furthermore, although hearing-impaired infants and adults produce many recognizable nonverbal vocalizations (U. Jürgens, 2009; Scheiner et al., 2006), there are clear acoustic differences between nonverbal vocalizations of hearing and deaf individuals (Makagon, Funayama, & Owren, 2008; Pisanski, personal communication; Sauter et al., 2019). These observations imply that social learning affects the context in which nonverbal vocalizations are typically produced, and to some extent their acoustic structure as well. This is not surprising given that humans are neurologically equipped to control vocal production at will, so that every vocal behavior falls on a continuum from spontaneous to volitional (Scherer

& Bänziger, 2010). This volitional control introduces a major confound into research on core human nonverbal repertoire, raising issues of cultural specificity, the authenticity of conveyed emotion, and other methodological concerns discussed in Section 2.

1.1.5 Summary of signal production

The signals that animals and humans communicate with can be produced via different mechanisms, from hormonally triggered morphological changes to socially learned signals emitted under direct conscious control. Human communication, including both language and nonverbal communication, employs the full range of these possibilities. Accordingly, the production of nonverbal vocalizations can be profitably analyzed from several perspectives. Sexual dimorphism in voice characteristics can be regarded as a somatic signal shaped by sexual selection, and the same logic of body size exaggeration that determines the morphology of the vocal tract also applies to vocal behavior – for example, the tendency to lower the pitch and to extend the vocal tract in order to appear more dominant. The core repertoire of human nonverbal vocalizations appears to be species-typical (Paper III), but the same vocalizations can be produced in a manner ranging from largely spontaneous or "honest" (Paper I) to purely volitional or "deceptive", with some revealing acoustic differences between the two (Paper II). There is also a socially learned and thus culture-specific component to the production of nonverbal vocalizations, particularly in the gray zone between purely non-linguistic exclamations and semi-verbal onomatopoeic interjections (emblems) such as *Urgh!* and *Ouch!* Because of this variety of mechanisms involved in vocal production, it turns out to be a non-trivial task to describe the species-typical component of human vocal behavior. This task is a major part of my dissertation and the subject of Chapter 2.

1.2 Signal perception

Moving on from the producer to the receiver, the perceived signal can be processed in various ways, which mirror the hierarchy of production mechanisms discussed in section 1.1. From least to most cognitively sophisticated, communicative signals can have direct perceptual effects, trigger innately specified responses, or be associated with one or more response strategies through learning.

1.2.1 Direct effects

The most direct effect a signal can have on a receiver – in the sense of involving the smallest amount of neural processing – is largely determined by the properties of peripheral receptors and low-level sensory circuits. For example, harsh and loud shrieks effectively attract the listeners' attention and have a generally aversive effect because of their acoustic properties, leading some authors to propose a distinction between direct and indirect affect induction in the audience (Owren & Bachorowski, 2003; Owren & Rendall, 1997). It is also possible that the cries of infants in humans and other mammalian species are under selective pressure to (1) maximize their subjectively experienced loudness by carrying a significant amount of energy in the range of frequencies to which adults are particularly sensitive (Lingle et al., 2012), potentially causing pain and even hearing loss in the listener (Calderon, Carney, & Kavanagh, 2016), and to (2) prevent habituation by means of introducing frequency modulation, nonlinear vocal phenomena, and other acoustic irregularities (Koutseff et al., 2018; Lingle et al., 2012).

I argue in Paper VII that the acoustic properties of all high-arousal calls are such as to maximize their bottom-up salience, attracting and holding the listeners' attention with minimum engagement of task-directed, top-down attention. If this is correct, the attention-grabbing and often aversive effect of such sounds is not mediated by learned associations, but primarily stems from excessive stimulation of the listener's auditory system. Some degree of neural processing is always necessary, however, so in my opinion it is not meaningful to separate "direct" perceptual effect from other innate responses discussed in section 1.2.2. Even so, the contribution of low-level perceptual processing is interesting theoretically because it underscores the danger of approaching all communication with a toolkit borrowed from semantics. The informational content of a startling shriek or gun shot, if any, is clearly very different from that of a propositional utterance.

1.2.2 Innate responses

Even when the receiver's response is not directly predicated on the physical properties of the signal, it can nevertheless be unconditional and innate – that is, it can develop in all members of a species without being learned. The simplest examples are close to Owren and Rendall's (1997) definition of direct effects discussed above and may appear to require little cognitive processing, as in the case of the acoustic startle reflex – a rapid, unconditional defensive reaction to a threatening stimulus such as a sudden loud noise. However, even such simple responses do not have to be impervious to contextual effects. For instance, in humans the eyeblink to a sudden noise is attenuated by positive and enhanced by negative affective states (Lang, Bradley, & Cuthbert, 1990). Non-associative learning can also play some role in modulating the response. For example, the

startle response is attenuated if the eliciting stimulus is presented repeatedly (habituation) or preceded by a weaker prestimulus – a phenomenon known as prepulse inhibition (Braff, Geyer, & Swerdlow, 2001). The defining feature of this category of innate responses, however, is that the basic pattern of the eliciting stimulus and response are "hard-wired" rather than learned.

In the animal world, innate responses are extremely common and crucial for survival. To refer back to the example of somatic signals that regulate mating, female preferences for features like bright plumage or long tail feathers are not the product of associative learning, but rather innately specified responses to the appropriate triggering stimuli. In other words, a female peacock does not learn from personal experience that males with large tails produce healthy offspring; instead, their brain is predisposed to respond favorably to a particular combination of visual features on a large tail (Miller, 2011). Innately specified responses can persist not only without a chance to learn the meaning of the signal through previous exposure, but without even a theoretical possibility of such exposure. For instance, moths that migrated to Pacific islands relatively recently continue to drop to the ground upon hearing an ultrasound, although this defensive measure against bats is rather pointless in their bat-free environment. In contrast, this motor response has been decoupled from the detection of bat cries in species endemic to the islands, who no longer drop down, although their ears are still somewhat sensitive to ultrasound (Fullard, Ratcliffe, & Soutar, 2004).

A well-documented example of an innately prepared response in humans is rapid detection of threatening stimuli by subcortical circuits centered on the amygdala, which orchestrates a reflexive fearful response to pictures of snakes and spiders (LeDoux, 2012; Öhman, 1986). Interestingly, the amygdala also appears to respond similarly to facial expressions of fear in other humans – specifically, to the increased visibility of the sclera as the sender's eyes open wide in fear (Whalen et al., 2004). In this case both the production of the facial expression of fear and its detection appear to be innate and relatively inflexible – that is, hard to control or inhibit intentionally. Revealingly, the responsible neural mechanisms are largely subcortical, which makes both production and response very fast, but also hinders intentional control.

Returning to human vocal behavior, there is evidence that the processing of emotional vocalizations has a very rapid subcortical component (Sauter & Eimer, 2010), although the underlying neurological mechanisms are not yet sufficiently well understood (Bestelmeyer, Maurage, Rouger, Latinus, & Belin, 2014; Frühholz, Trost, & Kotz, 2016; Oliva & Anikin, 2018). Interestingly, the tendency to associate low auditory frequency with large and heavy objects is found in congenitally blind individuals (Hamilton-Fletcher et al., 2018), suggesting that these crossmodal correspondences are not learned from experience. Accordingly,

the tendency to associate low-pitched voices with masculinity and dominance can be seen as an example of an innate response to a vocal signal, particularly in the light of the well-documented sex differences in the sensitivity to these vocal cues, which is a signature of sexual selection (Charlton et al., 2013; Evans et al., 2006). Crossmodal associations and other innate response mechanisms thus appear to play an important role in the processing of nonverbal vocalizations, as discussed in section 3.2 and Paper VII.

An interesting special case of learning is imprinting, which plays an important role in creating a powerful bond between the mother and her offspring. In highly vocal and colonial animals such as seals and walruses, the ability of the mother to learn the voice of her pup is crucial for them to reunite after the mother's hunting expeditions (Charrier, Aubin, & Mathevon, 2010). Likewise, human parents – particularly mothers – are good at recognizing the voice of their infants, and hearing their child's cries triggers an unconditional nurturing response, which includes both a powerful emotional component and the milk letdown reflex (Zeifman, 2001). The cries of baby seals and human infants – more specifically, the unique acoustic signatures that enable individual recognition – are thus learned signals that trigger innate nurturing behavior in the mother.

1.2.3 Learned responses

When there is no innate predisposition to respond to a signal in a particular way, the receiver has to learn the signal's meaning and the most appropriate response from experience. In behavioral terms, it means learning that the signal predicts future changes in environmental conditions or in the sender's behavior, which requires some form of associative learning. Depending on exactly what is learned and how this information is processed, learned responses can be more or less flexible. The simplest strategy would be to learn a single deterministic *if-then* rule – that is, to associate a signal with a standard response that does not depend on the broader context. Because such "mindless" conditioning is seldom advantageous in nature, however, reinforcement learning is usually flexible enough to make the stimulus-response association context-dependent (Pearce, 2008). In a communicative context, the animal may take into account additional factors such as the sender's identity, the history of previous interactions with the sender, the presence of other group members, etc.

While the resulting behavior can still be described using a large number of increasingly complicated, probabilistic *if-then* rules, the relationship between the signal and the response becomes less predictable. As a result, at some point it becomes more parsimonious to describe signal perception in terms of the sender learning to extract the relevant information from the signal and to respond appropriately. For example, vervet monkeys respond to alarm calls depending on

their current position. An animal who hears an eagle alarm call while on the ground will rush up into the branches, whereas an animal who is already high up will descend from the exposed treetops (Seyfarth et al., 1980). Furthermore, if an alarm call is later followed by the sound made by the actual predator, this otherwise frightening sound no longer provokes a strong response, presumably indicating that the presence of a predator has already been inferred from the alarm call and remembered. Thus, it appears that an eagle alarm call evokes a mental representation of an eagle in the audience, a snake alarm call brings to mind a representation of a snake, and so on (Wheeler & Fischer, 2012).

The idea of signals evoking mental representations in animals remains a somewhat controversial, but parsimonious explanation for flexible responses to context-specific, or functionally referential, signals such as alarm calls (Manser, 2013; Wheeler & Fischer, 2012). Whether or not mental representations are involved, highly flexible cognitive processing is required when the same signal can be produced in a broad range of contexts. For instance, chimpanzees who hear a sequence of screams from two familiar individuals seem to be able not only to tell who is the aggressor and who is the victim, but also to judge whether these roles conform to their expectations based on the existing social hierarchy (Slocombe, Kaller, Call, & Zuberbühler, 2010), suggesting that they build mental models of the situation based on what they hear. Likewise, people are highly attuned to such nuances as laughing *with* someone versus *at* someone (Szameitat et al., 2009; Wood, Martin, & Niedenthal, 2017). They also find it a natural task to guess whether the people laughing together are friends or strangers, even when listening to recordings from a different culture (Bryant et al., 2016). Characteristically, comprehension develops earlier and far outstrips production both in human infants and in language-trained animals (Rumbaugh & Savage-Rumbaugh, 1994), again demonstrating that the capacity for highly flexible, context-dependent interpretation of learned signals is more widespread in the animal world and less cognitively costly than the corresponding production skills.

For many animals, and certainly for humans, signal perception can thus be described in terms of the inferences that receivers make on the basis of the information that they extract from a signal. This view is closely aligned with the pragmatic approach to human communication, which emphasizes social aspects of communication (Scott-Phillips, 2015; Sperber & Wilson, 1986). From this perspective, the distinction between language and mammalian vocalizations, including human nonverbal vocalizations, is rather blurry on the receiver's side, even though their production mechanisms are distinct (Ackermann et al., 2014; U. Jürgens, 2009). The pragmatic meaning of an utterance – whether a sentence or a bout of vocalizing – still needs to be inferred and integrated into a situation model (Zwaan & Radvansky, 1998) in a manner that goes beyond pure semantics. Where humans arguably push the boundaries the most compared to animal

communication is in establishing a true dialogue, in which the speaker ostensively communicates the intention to communicate, and both the speaker and the listener cooperatively obey Gricean maxims (Fitch, 2010, Ch. 3; Scott-Phillips, 2015), which requires a developed ability to understand the others' mental states (theory of mind) and high-order intentionality.

A dialogical perspective that explicitly acknowledges an active, bidirectional interaction between the speaker and the listener is an influential approach to conversation analysis (Garrod & Pickering, 2004). There is also abundant evidence that nonverbal vocalizations, such as laughter, tend to obey the rules of turn taking when they punctuate speech, suggesting that they can be fully integrated in ordinary conversation (Provine, 2001). Furthermore, purely nonverbal vocalizations grade smoothly into emblems (*Huh? Wow!)*, so they can presumably function as semantically impoverished but highly expressive words. At the same time, as argued above, purely nonverbal vocalizations – especially those of a more spontaneous nature (Paper I) – are closer to mammalian calls than to language in terms of their production mechanism. For example, a scream of sudden fright appears to be broadcast without taking into account the audience, social appropriateness, etc. As a result, a dialogical perspective is mostly appropriate for vocalizations that are intentionally integrated in conversation, and arguably less so when the main focus is on the species-typical vocal repertoire, as in this dissertation.

1.2.4 Summary of signal perception

As with signal production (section 1.1), a variety of cognitive mechanisms are involved in the processing of communicative signals, including nonverbal vocalizations. Their effect on the audience can be strongly affected by low-level perceptual features, and the response can be largely stereotypical, as in the case of a generalized startle reflex to any unexpected noise. At the other end of the spectrum, subtle acoustic variation in a particular vocalization, such as a laugh, can be integrated with contextual information into a detailed mental representation of the situation, enabling complex inferences about who is laughing, what is happening, who else is present, etc. All these mechanisms are part of nonverbal communication, however, and it would be a mistake to focus only on the most cognitively sophisticated aspects of signal perception to the exclusion of less flexible, involuntary or innate responses. In this thesis, I emphasize the role of relatively low-level perceptual mechanisms in determining the meaning of nonverbal vocalizations, testing the contribution of their bottom-up auditory salience (Paper VII) and specific aspects of voice quality (Papers IV-VI).

1.3 Research questions

As stated at the beginning of Section 1, the goal of this dissertation is to explore the role of nonverbal vocalizations in human communication from a comparative and evolutionary perspective in order to elucidate how human vocal behavior is informed by our phylogenetic history. Simply put, this means describing what nonverbal vocalizations humans produce and comparing them with the vocal communication of other animals. Having presented the theoretical framework within which this investigation is conducted, I can now break down its overarching goal into specific research questions and show how my work has contributed to answering them.

In order to compare the vocalizations of humans and other animals, we first have to describe the human nonverbal repertoire – that is, the acoustically distinct classes of nonverbal vocalizations (call types) that all humans produce without having to learn them. This may sound like a trivial task; however, like marine mammals and bats and unlike other apes, humans are accomplished vocal learners. It is therefore necessary to separate the species-typical component from socially learned or idiosyncratic vocal behavior. As argued above, human nonverbal vocalizations are controlled by a phylogenetically old, prelinguistic mammalian vocalization system, and these sounds form the species-typical core that also informs speech prosody. Nonverbal vocalizations have also been shown to be relatively similar cross-culturally, but a systematic investigation of this core nonverbal repertoire has not yet been performed. The ambition of Paper III is to advance this task by examining, in a cross-cultural setting, the categorization of nonverbal vocalizations into classes defined by their acoustics and meaning. Papers I and II prepare the ground for this investigation: I present a case for using spontaneous vocalizations as less culture-specific and more suitable for phylogenetic comparisons (Paper I) and show that they are indeed different from vocalizations intentionally produced on cue (Paper II).

Once we know what nonverbal vocalizations humans communicate with, the next step is to compare them with vocal communication in non-human animals. One way to do so is to look for similar call types – sounds that occur in different species with a recognizable acoustic structure and functionally similar eliciting contexts. There is some promising work in this direction, notably the demonstration that all great apes laugh (Ross et al., 2010), but I did not perform comparative analyses of this type. The direction I followed was to investigate the "acoustic code" of human nonverbal vocalizations – the principles of voice modulation that underlie nonverbal communication. One aspect of this work was methodological: I developed and tested an open-source toolbox for parametric voice synthesis that made it possible to synthesize nonverbal vocalizations (Paper IV) and to test hypotheses about the role of specific vocal characteristics such as

nonlinear phenomena (Paper V) and breathy voice quality (Paper VI). Although these manipulations were performed on human vocalizations, the results are in line with theoretical expectations based on previous work on bioacoustics and can later be replicated in animal playback studies.

Papers IV-VI thus represent an attempt to better understand the acoustic code involved in human nonverbal vocalizations and to compare this code with what is known of mammalian vocal communication in general. More fundamentally, however, it is important to understand *why* the acoustic code is the way it is. In Paper VII, I investigate the link between the acoustic properties of high-intensity calls and the allocation of bottom-up auditory attention in the brain. The close match between the acoustic characteristics of salient acoustic events and high-intensity vocalizations suggests that some aspects of vocal production may have evolved to exploit sensory biases.

In sum, this dissertation engages with two main questions. They are too broad to be answered conclusively within the scope of this work, but the objective is to contribute to their better understanding. These research questions are:

(1) *What nonverbal vocalizations do humans possess as a species?*
This is the subject of Section 2 and Papers I-III.

(2) *How is information encoded acoustically in these sounds?*
This question is addressed in Section 3 and Papers IV-VII.

2. Species-typical component

As discussed in Chapter 1, the repertoire of human nonverbal vocalizations appears to have a strong innate or species-typical component – that is, all humans develop these vocalizations, largely regardless of their first language and other environmental input. While the existence of a core, species-typical human vocal repertoire is now widely acknowledged on a theoretical level, its precise descriptions remain scarce. The first reason for this difficulty is historical: the great theoretical and practical significance of language has understandably made its study a priority at the expense of nonverbal vocalizations. Many prosodic features of speech are probably derived from nonverbal vocalizations, and there is increasing convergence between research on emotion in speech and nonverbal vocalizations (Elfenbein & Ambady, 2002; Kamiloglu, Fischer, & Sauter, 2019), but even so, looking at speech prosody alone would be a roundabout way to learn about phylogenetically older vocal behaviors. Yet, it is only in the last decade that research on nonverbal vocalizations has really taken off (section 2.1). As a result, even broad questions, such as what is universal and what is culture-specific in human nonverbal communication, remain open.

The second problem with extracting the species-typical vocal component is that humans have dual vocal control and can intentionally produce, suppress, or manipulate all kinds of vocalizations, including putatively innate sounds such as laughs and screams (section 1.1). For the purposes of understanding the vocal repertoire that humans possess as a species, it would be preferable to minimize the intentional control of vocal behavior and to look at more spontaneous forms. Vocalizations triggered by an unexpected event and associated with a genuine, strong emotion may be particularly valuable for the purpose of identifying the species-typical component in vocal behavior because their sudden occurrence may minimize impression management. On the contrary, in most previous studies human nonverbal vocalizations were elicited under controlled conditions by asking participants to vocalize on cue, deliberately aiming to portray a particular emotion or context (e.g., Belin, Fillion-Bilodeau, & Gosselin, 2008; Cordaro et al., 2016; Lima, Castro, & Scott, 2013; Maurage, Joassin, Philippot, & Campanella, 2007; Sauter, Eisner, Ekman, et al., 2010).

Aiming to contribute to the nascent research on nonverbal vocalizations and to transcend potential limitations of studies based on actor portrayals, I investigated

the possibility of using observational material – vocalizations that are produced more spontaneously, in real-life situations. This chapter is about finding suitable sources of such vocalizations, comparing them with actor portrayals, and characterizing the core repertoire of nonverbal vocalizations based on these observations. Papers I-III are briefly summarized here, situated in the larger context of describing the species-typical component of human vocal behavior, and in some cases updated to include more recent research that was not yet available at the time when Papers I-III were written. Some limitations of the available data and future challenges are also highlighted.

2.1 Sources of spontaneous vocalizations (Paper I)

From the point of view of data availability, the perfect scenario would be to place countless cameras and microphones all over the world and to record people from culturally isolated groups vocalizing as they seek and obtain food, encounter predators, compete for resources, bond, attract mates, suffer disappointments and accidents, etc. Over time, this ideal database would accumulate thousands of instances of nonverbal vocalizations from functionally diverse, survival-relevant contexts of varying intensity. Culturally invariant acoustic properties could then be abstracted, providing a complete and ecologically valid catalog of human vocal behavior that does not depend on the first language or cultural tradition. If it was also possible to simultaneously record neural activity in each vocalizer, we would be in possession of a truly comprehensive account of vocal production.

This master plan is of course impossible for logistical and ethical reasons, but it can be insightful to view other research projects as approximations to this idealized scenario under a number of simplifying assumptions. The most common approach has been to focus on only one or two cultural settings and to elicit the vocalizations by asking participants to pretend that they are experiencing a particular emotion (Belin et al., 2007), often accompanied by a short vignette describing a particular context, such as being tickled for amusement, a sudden fright for fear, and so on (Cordaro et al., 2016; Lima et al., 2013; Sauter, Eisner, Ekman, et al., 2010). Some effort has been invested into testing the assumption of cultural universality: several research groups have obtained recordings and performed playback studies in remote locations, minimizing the risk of cultural contamination by other populations and the globalized entertainment media (Bryant et al., 2016; Cordaro et al., 2016; Sauter, Eisner, Ekman, et al., 2010).

The second major assumption is that people are good actors – that is, that they can produce realistic vocalizations on cue. The resulting vocalizations are certainly recognizable and presumably representative of everyday vocal interactions, in

which impression management is ubiquitous (Scherer, 2003). At the same time, the notion that vocalizations elicited in the lab are fundamentally similar to spontaneous vocal behavior remains an assumption. The extent to which this assumption is justified is discussed in the following section, but in order to test it, we first have to obtain spontaneous vocalizations for a comparison.

Emotions can sometimes be induced in the lab using a combination of Stanislawski's system employed by professional actors and experimental procedures such as watching an amusing video clip (Scherer & Bänziger, 2010). There is also the relatively underexplored option of serendipitously recording vocalizations as events unfold in real life. This approach was pioneered in the speech community by using recordings of radio programs, interactions with customers at information helpdesks, communications with airplane pilots under severe stress, and other real-life interactions for which it was possible to determine the most likely emotional state of the speaker (e.g., Erickson, 2005; R. Jürgens, Drolet, Pirow, Scheiner, & Fischer, 2013). More recently, the rise of social media platforms has offered a new and potentially limitless source of publicly available data, some of which could never have been collected otherwise. To take the most extreme example, no experimenter would make a participant undergo a physical injury to study pain vocalizations, but people sometimes share videos of themselves in acutely painful situations (e.g., sports accidents and giving birth) via the social media.

The possibility of using online sources for research on nonverbal vocalizations is discussed in Paper I. A search on www.youtube.com uncovered many types of recognizable contexts accompanied by vocalizing, which were classified into several emotions (amusement, anger, disgust, fear, joy, pleasure, and sadness) as well as pain and physical effort. In a validation study, listeners from several countries could usually recognize the context in which the vocalization was produced, confirming that spontaneous vocalizations effectively communicate affective states (Parsons, Young, Craske, Stein, & Kringelbach, 2014; Sauter & Fischer, 2018). Curiously, recognition accuracy did not depend on the first language of listeners, suggesting that spontaneous vocalizations may be less culture-specific than actor portrayals.

An important tradeoff of working with amateur recordings from online sources is their low acoustic quality and the prevalence of background noise. Nevertheless, acoustic analysis of the recordings followed by machine learning was sufficiently powerful to reach the same recognition accuracy as human raters. Moreover, a large number of vocalizations were obtained within a reasonable amount of research time: 260 vocalizations were selected for the validation study, but in total about 600 recordings were collected from hundreds of unique speakers. Paper I thus achieved two objectives: it proved the feasibility of obtaining large numbers

of nonverbal vocalizations from social media and produced a collection of spontaneous vocalizations spanning a wide range of contexts and emotion intensity levels, which proved useful for further testing.

In retrospect, another important lesson to draw from Paper I concerns the importance of doing fieldwork, whether in the physical world or online. This anthropological or ethological approach to studying nonverbal communication in a natural environment has long been championed by Robert Provine (reviewed in Provine, 2016), but most publications focus on hypothesis testing under controlled experimental settings. Admittedly, researchers of human communication already have a great deal of insight when the object of study is our own species rather than, say, pheromone-laying ants. At the same time, confining the explorations of vocal behavior to a laboratory setting constrains the range of phenomena that can be observed, particularly when participants are explicitly told what emotion to portray or what sound to produce (e.g., an isolated vowel in Maurage et al., 2007 and Belin et al., 2008). Completely new patterns can emerge when vocal behavior is studied in a more natural and less structured environment. For example, an investigation of confusion patterns in Paper I unexpectedly revealed a strong tendency to classify sounds based on their acoustic class rather than the speaker's affective state, suggesting a shift of perspective from emotion to call types (Paper III; see also Engelberg & Gouzoules, 2019; Schwartz, Engelberg, & Gouzoules, 2019).

At the time when Paper I was published, the only other available collection of non-acted nonverbal vocalizations by human adults was apparently the OxVoc database, which included 19 cries and 30 laughs collected from www.youtube.com (Parsons et al., 2014). Several other corpora of non-acted vocalizations have been published since then, contributing to the total pool of available vocalizations. For instance, Raine, Pisanski, and Reby (2017) compiled and tested a corpus of tennis grunts recorded during real matches. The team led by Harold Gouzoules have focused on screams, obtaining some non-acted examples from YouTube clips, newscasts, and unscripted television programs (Engelberg & Gouzoules, 2019; Engelberg, Schwartz, & Gouzoules, 2019; Schwartz et al., 2019). In addition to using online videos, Atias et al. (2019) recorded the first reactions of 153 lottery winners, thus obtaining ecologically valid vocalizations of extreme joy. There is also ongoing work on recording vocalizations emitted by women during childbirth directly in hospitals and simultaneously obtaining physiological measurements, so as to correlate acoustic characteristics with the actual level of pain and physical condition (Reby & Pisanski, personal communication). Last but not least, there is a long tradition of studying infant cries. A full discussion of this vast field is beyond the scope of this thesis, but by its very nature research on infants has to rely on spontaneous behaviors (infants cannot be asked to portray an emotion), and infant vocalizations have been successfully recorded and analyzed in multiple

real-life contexts such as vaccinations (Koutseff et al., 2017) and temper tantrums (Green et al., 2011).

Taken together, these projects have convincingly demonstrated the feasibility of collecting and analyzing spontaneous examples of human vocal behavior. The main challenge for the future is to scale up data collection. At the time of writing, the largest corpus of emotional speech that I am aware of contains about 2000 stimuli from 54 actors (Lassalle et al., 2019), and for adult nonverbal vocalizations the largest corpora reach about 600 (Paper I) to 1000 (Bachorowski, Smoski, & Owren, 2001) sounds. A noteworthy recent effort is the large-scale study by Cowen and colleagues, who collected and tested about 2000 nonverbal vocalizations by 56 speakers from several countries and included a separate comparison corpus of spontaneous vocalizations from social media (Cowen, Elfenbein, Laukka, & Keltner, 2019). In another impressive project, over 3000 bouts of spontaneous laughter were extracted from conversations and analyzed acoustically (Wood, 2019). However, even these numbers pale in comparison with datasets compiled in infants (e.g., 15,000-30,000 infant vocalizations in Scheiner et al., 2002, 2006) and non-human mammals (e.g., 15,000 bat calls analyzed by Prat, Taub, & Yovel, 2016; 10,000 marmoset calls in DiMattina & Wang, 2006). In the field of acoustic communication, chasing large numbers can sometimes be essential for making progress, particularly if the goal is to map the entire vocal repertoire of a species, to uncover relatively subtle differences in the acoustic structure of calls between different populations (Hammerschmidt & Fischer, 2019), or to achieve accurate recognition of emotion in the voice by machine learning algorithms (e.g., Hershey et al., 2017).

Human data ought to be quite straightforward and cheap both to collect and to test compared to animal vocalizations, so it is really an extraordinary situation that the tested corpora are typically so small – a few dozen calls from as few as four speakers is quite standard (e.g., Lima et al., 2013; Sauter, Eisner, Calder, & Scott, 2010). Moreover, these corpora are often not available for pooling because of administrative and ethical restrictions. To draw a parallel with the more technically and ethically challenging genetic research, major breakthroughs have been associated with compiling truly comprehensive datasets containing the genomes of tens of thousands of people from all over the world and making them available for research (Lek et al., 2016). In the case of vocalizations, a success story is the Xeno Canto online repository of bird songs (https://www.xeno-canto.org), which has provided easy access to enormous datasets and has been used, for example, to benchmark machine-learning algorithms for species detection (Stowell & Plumbley, 2014). A similarly large-scale, open databank of human vocalizations could catalyze the field, and it would certainly make the reported effects considerably more robust.

2.2 Are spontaneous vocalizations different? (Paper II)

In the previous section I discussed some reasons to supplement volitional vocalizations (also known as actor portrayals, simulated or play-acted calls, etc.; on terminology, see Engelberg & Gouzoules, 2019) with more spontaneous examples recorded in real-life interactions (Paper I). Having done that, an important question is whether these spontaneous vocalizations are indeed different from more conventional examples, namely nonverbal vocalizations produced on cue to portray a particular emotion. This was tested in a simple experiment, in which participants heard a mixture of spontaneous and volitional vocalizations and classified them as either "real" or "fake" (Paper II). Controlling for recording quality and other extraneous factors that might have influenced the perceived authenticity, spontaneous vocalizations were perceived as more authentic for all eight analyzed emotions and all six published corpora of volitional vocalizations, although the difference in authenticity varied considerably across emotions and corpora.

The conclusions of Paper II are straightforward with regard to answering the question of whether or not spontaneous vocalizations sound more authentic than actor portrayals. They do. Machine learning further demonstrated that using a mixture of spontaneous and volitional vocalizations to train a classifier made it considerably more robust compared to training it on one type only. Taken together, these results speak strongly in favor of including both elicited and observational material in research on nonverbal communication. On the other hand, it turned out to be less straightforward to pinpoint the acoustic differences responsible for making certain vocalizations sound highly authentic and others "fake". An even more fundamental problem is that the ground truth of vocal production can be hard to ascertain – some of the putatively spontaneous vocalizations in Paper I may well have involved a good deal of volitional control. In terms of perception, the clearest pattern was that listeners treated highly intense (high-pitched, noisy, unpredictable) vocalizations as authentic, possibly because they expected intense emotion to cause dramatic, hard-to-fake vocal behaviors and changes in voice quality that are not easy to produce at will. A similar pattern of interpreting intense emotion as more likely to be authentic has since been demonstrated in a large-scale comparison of genuine and acted emotional speech (Juslin, Laukka, & Bänziger, 2018).

Interestingly, the notion that acoustically extreme nonverbal vocalizations, such as screams, are particularly difficult to fake was questioned by another team soon after Paper II was published. Engelberg & Gouzoules (2019) found that listeners could not tell the difference between volitional and relatively spontaneous human

screams. This study is unusual in that the sounds were selected based on their acoustic characteristics rather than the emotional state of the caller. The volitional screams were mostly produced by professional actors, and the relatively limited number of screams in both Engelberg and Gouzoules (2019) and Paper II makes it even more difficult to draw direct comparisons between these two studies. However, the work by Engelberg and Gouzoules (2019) does prove that actors can produce very convincing screams and presumably other "costly" vocal behaviors, just as they can achieve mastery over their facial expressions and body language.

Other teams have since continued research on authenticity perception in nonverbal vocalizations (Bryant et al., 2018; Engelberg & Gouzoules, 2019; Lavan et al., 2019; Sauter & Fischer, 2018), and a more fine-grained picture may eventually emerge. A particularly pressing task is to better understand the limits of volitional control over vocal production, identifying the acoustic signatures of genuine emotion. As argued in Paper II, both volitional and spontaneous expressions are common in everyday life, and both are legitimate objects for research. The existence of perceptually salient differences between them, however, means that emotion authenticity is a relevant characteristic that should be taken into account when studying vocal behavior.

2.3 Human nonverbal repertoire (Paper III)

The research on nonverbal vocalizations in adult humans began as a branch of the psychology of emotion and a direct extension of research on affective speech. In fact, nonverbal vocalizations are often referred to as "affect bursts" (e.g., Belin et al., 2008; Schröder, 2003), and practically all studies focus on the emotions that can be expressed with these sounds. By its very nature, the observational method of data collection championed in Papers I and II leads away from this emotion-centric view: if a vocalization is taken from social media, it is impossible to know exactly why the person is vocalizing, what they are feeling, or what message, if any, they intend to convey. The corpus validation study (Paper I) proved that listeners could often tell whether the vocalizer was amused or sad, afraid or in pain, etc. However, the confusion patterns suggested that listeners may have perceived these sounds in terms of a few acoustic classes or call types, which were then interpreted in terms of emotion to fit the offered classification categories (which were, in turn, inspired by previous psychological research). For example, scream-like sounds were usually interpreted as an expression of fear, which is in line with the way fear is typically portrayed, but not necessarily with the reality – based on what I observed when collecting the material, screams often expressed aggression, pain, and in fact even positive states like jubilation or a pleasant surprise.

These observations, as well as other evidence presented in Paper III, suggested that human nonverbal vocalizations were perceived as consisting of a number of fairly distinct call types such as laughs, cries, screams, and moans. Laughs (Bryant et al., 2016; Lavan, Scott, & McGettigan, 2016; Wood et al., 2017), screams (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015; Engelberg & Gouzoules, 2019; Engelberg et al., 2019), and certain infant vocalizations (Green et al., 2011; Lingle et al., 2012; McCune et al., 1996; Newman, 2007; Scheiner et al., 2002) have sometimes been treated as acoustically distinct vocalizations (call types), but no attempts have been made to explore the full range of human nonverbal vocalizations from this perspective. I therefore asked participants from three countries to classify nonverbal vocalizations verbally by call type and emotion as well as nonverbally using the odd-one-out method in a triad classification task (Paper III). As predicted, call types emerged as an intuitive, perceptually salient, and partly language-independent classification of nonverbal vocalizations that appeared to precede the attribution of a particular meaning.

The notion that vocalizations are acoustic classes rather than direct expressions of emotion is not very surprising for a biologist, considering the long tradition of classifying animal calls, including the relatively graded primate vocalizations (Fischer, Wadewitz, & Hammerschmidt, 2017; Scheiner et al., 2002), into acoustically defined categories. At the same time, in the field of human nonverbal communication Paper III was the first attempt to systematically map the underlying acoustic categories and explore their relationship with the interpretation of each vocalization. The actual classification of call types proposed in Paper III is best seen as preliminary – the number of stimuli and tested languages would have to be considerably greater before claims can be made that the entire species-typical vocal repertoire has been mapped exhaustively. However, the shift of perspective from emotion to call types can inform further research on acoustic communication and integrate it with the theoretical framework of bioacoustics, as discussed further in section 3.1.

3. Cracking the code

In section 2 I described my work on exploring the diversity of human nonverbal vocalizations – their spontaneous and volitional forms, perception and categorization in a cross-cultural context, and the acoustic types of which they consist. This exploratory research can ultimately be regarded as an attempt to describe the species-typical component of human vocal behavior – the kind of task an ethologist performs when documenting the vocal repertoire of a species. The studies presented in section 3 are different: here the focus is on determining the effect of particular acoustic characteristics of a stimulus on its meaning. I begin by presenting a new method for synthesizing vocalizations and evaluating the effect of specific acoustic manipulations (section 3.1) and then discuss some possible causes for the strong similarities found in the acoustic code across species (section 3.2). In other words, if section 2 was about *what* human nonverbal vocalizations are, section 3 is about *how* they function in communication – about the acoustic code that makes them meaningful.

3.1 Testing acoustic manipulations (Papers IV-VI)

Natural, unmodified recordings are more ecologically valid than synthetic or manipulated vocalizations, but the downside is that the effect of particular acoustic characteristics on listeners can only be investigated by testing a large number of stimuli and performing a correlational analysis. For example, I observed that vocalizations with a higher pitch sounded more authentic (Paper II), were likely to be perceived as indicating fear if they were also tonal (Paper I), etc. Likewise, acoustic correlates of particular emotional states or dimensions, such as valence and arousal, were reported in numerous studies on affective speech (Kamiloglu et al., 2019), human nonverbal vocalizations (Lima et al., 2013; Sauter, Eisner, Calder, et al., 2010), and animal calls (Briefer, 2012).

The main problem with this correlational approach is that many acoustic properties co-vary, and very large sample sizes are necessary to tease apart the contribution of specific aspects of prosodic characteristics or voice quality – much larger than what is typically available in voice research (see section 2.3). The alternative is to manipulate recorded vocalizations in systematic ways, so as to test

how their meaning changes if, for example, we raise the pitch without changing any other aspect of the sound. This approach has been used successfully to test hypotheses about the role of fundamental frequency and formant spacing on perceived speaker's size and dominance (Feinberg et al., 2005; Fraccaro et al., 2013; Puts et al., 2006, 2016; see section 1.1.1). As described in Paper IV, however, not every acoustic manipulation is technically feasible, so the most powerful option is to create fully synthetic stimuli, over which we have complete control.

The possibility of using parametric voice synthesis to study the acoustic code in nonverbal vocalizations is explored in Papers IV-VI. I wrote a computer program – *soundgen* – that creates human or animal nonverbal vocalizations based on manually specified source and filter characteristics (Paper IV). Synthetic stimuli are the exact opposite of the spontaneous vocalizations used in Papers I-III: they offer perfect experimental control but have the lowest ecological validity (Kamiloglu et al., 2019). To mitigate potential problems caused by the synthetic stimuli sounding artificial, I closely modeled them on actual vocalizations from Paper I and validated the synthesis by means of comparing the original recordings with their synthetic reproductions in terms of their authenticity as well as the emotion that they were perceived to express (Paper IV). The main conclusion was that the quality of synthesis was high enough to make relatively short synthetic vocalizations practically indistinguishable from the original recordings, although the authenticity began to suffer as the length and acoustic complexity increased. Fortunately, nonverbal vocalizations are perfect targets for parametric voice synthesis: they are relatively simple compared to speech, typically short or repetitive, and at the same time incredibly rich in nonlinear phenomena (Paper V) and other acoustic features that would be impossible to manipulate without synthesizing the sound de novo.

Paper IV was published online in summer 2018, so it may be premature to speculate about the long-term usefulness of the sort of parametric voice synthesis implemented in *soundgen* for other researchers. Apart from the studies reported in Papers V and VI, I have used it for experiments on crossmodal associations (Anikin & Johansson, 2019; Anikin, Rudling, Persson, & Gärdenfors, 2018) and in two ongoing projects on body size exaggeration and context-dependent meaning of particular acoustic characteristics (in preparation). *Soundgen* has also been used to create synthetic morphs of human nonverbal vocalizations for a test of categorical perception (Adrienne Wood, personal communication). Another promising field for its application would be in bioacoustic research, where the majority of vocalizations are short enough to be amenable to manual parametric synthesis, and where precise acoustic manipulations open the door to testing many novel hypotheses. In fact, several other tools have recently been proposed for synthesizing biological sounds (Moore, 2016; Tanner, Justison, & Bee, 2019;

Zúñiga & Reiss, 2019), so there is clearly a demand for this technology in the research community. For the purposes of my own research, I was particularly interested in using *soundgen* to manipulate relatively subtle aspects of voice quality in nonverbal vocalizations – aspects that were previously impossible to manipulate experimentally. The results of these manipulations are reported in Papers V and VI.

In Paper V, *soundgen* was used to add a controlled amount of different nonlinear phenomena, namely pitch jumps (sudden changes in voice pitch), subharmonics (an additional low-frequency component making the voice rough, as in some rock singing), chaos (broadband spectral noise with preserved traces of tonality), or their combination, to synthetic human nonverbal vocalizations. As described in the paper, these nonlinearities are difficult not only to synthesize, but even to measure – even today, the only reliable method of their detection is to manually inspect each spectrogram while listening to the sound. As a result, most evidence of their perceptual effects is indirect, based on nonspecific measures of vocal roughness or spectral noise. Although relatively small-scale, the two experiments reported in Paper V proved the feasibility of adding controlled amounts of specific nonlinearities to synthetic sounds and demonstrated that these acoustic phenomena were interpreted flexibly, depending on their type and the kind of sound in which they occur. Of all the studies included in this dissertation, Paper V is probably the most obvious candidate for follow-up research: after this proof-of-concept demonstration, the same technique can be applied to many other types of sounds (infant cries, animal screams, etc.), and vocal nonlinear phenomena are so complex and varied that many studies would be needed to investigate their communicative role in a comprehensive manner.

Two experiments reported in Paper VI had the same design; in fact, they were conducted simultaneously with the ones in Paper V, but in this case the manipulation was to adjust laryngeal voice quality along the tense-breathy dimension. As in the case of nonlinear phenomena, this manipulation would be difficult or impossible to achieve without completely resynthesizing the sound, and the effect of laryngeal voice quality in nonverbal vocalizations had not been examined experimentally prior to this study. The results revealed that breathiness had a strong effect on the perceived valence of relatively ambiguous vocalizations, such as moans and gasps, as well as on the perceived level of general alertness or arousal of the speaker. As with Paper V, this opens the door to further investigations of the role of voice quality in nonverbal communication using precise experimental manipulations instead of correlational analyses.

In addition to showcasing the potential usefulness of the proposed method of parametric synthesis for voice research, Papers V and VI added weight to the notion that nonverbal vocalizations are best analyzed in terms of graded, but partly

distinct call types (Paper III). Both nonlinear phenomena (Paper V) and breathiness (Paper VI) affected the perceived meaning of a vocalization primarily when it was inherently ambiguous, mirroring an earlier observation that spectral noise and high-frequency energy were associated with aversiveness only in the more ambiguous call types among the vocal repertoire of the squirrel monkey (Fichtel, Hammerschmidt, & Jürgens, 2001). The implication of these findings is that the same acoustic change (e.g., a shift from tonal to rough voice quality) may signal different changes in the caller's affective state depending on the call type in which it occurs: in a moan, this may make a major difference between pleasure and pain; in a scream, a subtle shift from a purely fearful to a slightly aggressive attitude; etc. Generally, acoustic correlates of valence may remain elusive (Briefer, 2012) because the hedonistic or aversive nature of the eliciting stimulus mostly affects the choice of call type, whereas within-call variation may be determined primarily by the level of arousal or emotion intensity (Bastian & Schmidt, 2006; Fischer et al., 2017). As demonstrated by Papers V and VI and other recent work (Baciadonna, Briefer, Favaro, & McElligott, 2019; Briefer et al., 2017), however, within-call variation can also reflect valence, partly in a call-specific manner. Some markers of arousal may also be call-specific. For example, Linhart, Ratcliffe, Reby, & Špinka (2015) report that the acoustic changes associated with increasing distress in piglets were not the same in screams and grunts: amplitude marked higher arousal mostly in screams, while median frequency (a summary measure of spectral shape) increased only in grunts.

As discussed in section 2.3, this means that, instead of looking directly for acoustic correlates of discrete emotions or dimensions such as valence and arousal, voice research should distinguish explicitly between acoustic variation between and within call types (Briefer, 2012; Fischer et al., 2017). For nonverbal vocalizations, it may thus be more profitable to investigate the relationship between acoustic characteristics and meaning in specific types of vocalization, such as laughs (Wood et al., 2017) or screams (Arnal et al., 2015), instead of looking for acoustic correlates of discrete emotions or affective dimensions in all nonverbal vocalizations at once (as in numerous publications such as Paper I; Kamiloglu et al., 2019; Lima et al., 2013; Sauter, Eisner, Calder, et al., 2010; etc.).

Another corollary of the shift of perspective from emotion to call type introduced in Paper III and followed up in Papers IV-VI is that it brings the research on human nonverbal vocalizations more in line with the theoretical perspectives and analytical approaches employed in animal research. The distinction between within-call and between-call acoustic variation is a case in point, but it is also increasingly clear that the acoustic changes associated with high arousal (Briefer, 2012; Filippi et al., 2017) or aggressive vs. fearful attitude (Morton, 1977) display strong similarities across species, including humans. Hypotheses about human nonverbal communication can thus be guided by vocal research in other species,

and the manipulations tested in humans (as in Papers V and VI) can in turn shed new light on the role of these acoustic features in animal communication. Simply put, thinking of human nonverbal vocalizations in terms of call types makes research on human and animal vocal behavior more directly compatible.

3.2 The logic of the acoustic code (Paper VII)

The more we understand about how voice modulation can be used to communicate without language, and the more regularities we discover in the way this acoustic code functions across species, the more imperative it becomes to understand *why* it works this way and not another. For example, why are high-arousal vocalizations typically long, loud, high-pitched, and noisy (Briefer, 2012)? When this question is raised – which is actually not so often – explanations fall into two main categories: production mechanisms and perceptual biases. These acoustic characteristics might be consequences of physiological changes that affect vocal production in the sender, or they might be optimized to exploit perceptual biases in the receiver. As discussed below, these two explanations can be complementary rather than mutually exclusive.

To continue with the example of high-arousal vocalizations, general activation triggers a cascade of physiological effects via the autonomous nervous system (LeDoux, 2012; Scherer, 1986) and causes predictable changes in vocal production. For example, the voice becomes louder, brighter, and more high-pitched as the subglottal pressure and the tension of laryngeal muscles increase (Gobl & Ní Chasaide, 2010). These acoustic changes may therefore be observed in different species without necessarily being a design feature intended to optimize communication – they may be simply side effects of the way organisms physiologically respond to stress. On the other hand, some voice changes associated with high arousal may have been shaped by natural selection specifically for communicative purposes. For instance, nonlinear vocal phenomena are effective for attracting and holding the attention of listeners (Blumstein & Recapet 2009; Karp, Manser, Wiley, & Townsend, 2014; Townsend & Manser, 2011), and although they are more likely to appear at a high subglottal pressure (Cazau, Adam, Aubin, Laitman, & Reidenberg, 2016; Fitch, Neubauer, & Herzel, 2002; Herzel, Berry, Titze, & Steinecke, 1995), with good vocal control it is possible to suppress nonlinearities even in very loud and high-pitched calls such as opera singing or pant-hoots of chimpanzees (Riede, Arcadi, & Owren, 2007). In most cases, however, it is in the caller's interest to allow or even encourage nonlinearities in high-intensity calls to ensure that they are heard and noted by conspecifics. Accordingly, the prevalence of nonlinear phenomena in high-intensity calls may be regarded as an attempt to exploit perceptual biases in the

audience – an adaptation rather than merely a by-product of vocalizing in a stressed state.

The best-known hypothesis in vocal communication that appeals to perceptual biases is Ohala's frequency code (Ohala, 1984) and the closely related Morton's motivation-structural rules (Morton, 1977). The basic insight is that high auditory frequency is crossmodally associated with a small size, while low frequency is associated with a large size (Hamilton-Fletcher et al., 2018; Spence, 2011). As a result, in situations when it is to the caller's advantage to sound large (e.g., in dominance displays), it is adaptive to lower the pitch and formant frequencies, and vice versa: in situations when size should be downplayed (e.g., appeasement), pitch and formant frequencies should be raised. This simple, but powerful principle explains many of the acoustic properties of animal calls (August & Anderson, 1987; Briefer, 2012), human vocalizations and speech (Aung & Puts, 2019; Ohala, 1984; Pisanski et al., 2016), and even some aspects of sound symbolism in the vocabulary (Johansson et al., in print; Pitcher, Mesoudi, & McElligott, 2013).

In Paper VII, I tried to formulate a similarly general principle that would explain the acoustic characteristics of high-arousal vocalizations. An examination of literature on bottom-up attention (salience) in auditory processing revealed that the characteristics of salient acoustic events – events that involuntarily attract attention in a task-independent manner – closely mirrored the acoustic properties of emotionally intense vocalizations. Empirical tests reported in Paper VII confirmed that the self-reported salience of nonverbal vocalizations was closely related to the intensity of emotion that they were perceived to convey, that vocalizations rated as more salient were indeed distracting, causing a greater drop in task performance, and that the acoustic predictors of salience in nonverbal vocalizations were similar to those previously described in psychoacoustic studies with mixed environmental sounds. According to these findings, the acoustic characteristics of high-intensity vocalizations are tuned to match the optimal sensitivity of the auditory system. Assuming that this is not a coincidence, the "salience code" could be an adaptation on the part of vocal production to match the perceptual biases; alternatively, both production and perception may continuously coevolve so as to maintain this close match. In Paper VII I advocate the view that the sense of hearing is phylogenetically more conservative than vocal production, with the implication that the high salience of high-intensity calls can be understood in the light of the sensory bias hypothesis (Ryan & Cummings, 2013).

4. Summary

4.1 Conclusions

The work presented in this thesis makes the following contributions to the research questions laid down in section 1.3.

Question 1. What nonverbal vocalizations do humans possess as a species?

- I make a case for supplementing actor portrayals with examples of spontaneously produced nonverbal vocalizations.
- I show where to find spontaneous vocalizations (Paper I) and prove that they can be different from actor portrayals in terms of their acoustic structure and perceived authenticity (Paper II).
- I present a preliminary classification of the human nonverbal repertoire, describing the most distinct call types and their meanings (Paper III).

Question 2. How is information encoded acoustically in these sounds?

- I describe a novel method for synthesizing and manipulating human and animal vocalizations (Paper IV).
- Using this method, I demonstrate how nonlinear vocal phenomena (Paper V) and tense or breathy voice quality (Paper VI) affect the meaning of different types of nonverbal vocalizations.
- I suggest that processing biases in the auditory system contribute toward shaping the acoustic properties of high-intensity vocalizations (Paper VII), explaining certain similarities of high-arousal calls across species.

In terms of broader theoretical implications, I argue for a closer integration between research on human and animal vocal communication, including:

- a shift of focus from the recognition of emotion to meaningful acoustic variation within and across call types (Paper III),
- engagement with bioacoustics as a source of hypotheses to test in humans, and vice versa (Papers V and VI), and
- the adoption of an evolutionary perspective on human vocal behavior.

4.2 Broader significance

Broadening the scope beyond the main topics discussed in depth in this dissertation, potential practical applications of this research and more global directions for further exploration include:

- Human-machine interaction: better understanding of the acoustic principles of animal and human vocal communication can guide the development of interactive software capable of understanding affective prosody and producing simple nonverbal vocalizations or emotionally inflected speech, with numerous applications in social robotics, educational technology, entertainment industry, and other fields (for some pioneering attempts, see Breazeal & Aryananda, 2002; Read & Belpaeme, 2015).
- Animal welfare: there is a lot of interest in automatic monitoring of animal vocalizations to promote animal welfare, particularly for farm animals (Manteuffel, Puppe, & Schön, 2004; Mcloughlin, Stewart, & McElligott, 2019) and zoo animals (Whitham & Wielebnowski, 2013). This requires a good working model of vocal communication, including the identification of robust and readily detectable markers of emotion intensity and valence. The availability of parametric, automatically controlled sound synthesis offers the additional opportunity of providing auditory feedback to the animals – for example, in order to provide comfort in stressful situations or to create an enriched milieu.
- Evolution of language: unraveling the story of the origins of language requires a broad and profound knowledge of both human and animal communication in all its richness (Fitch, 2010). Human nonverbal repertoire is one piece of this enormous puzzle. In addition, some findings and theoretical perspectives discussed here (e.g., links between the acoustic code, crossmodal correspondences, and sensory biases) may help to shed light on early evolution of language as well as on language in its present form. My work on crossmodality and sound symbolism with Niklas Johansson, although not included in this dissertation, is a step in this direction.

5. 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. & Johansson, N. (2019). Implicit associations between individual properties of color and sound. *Attention, Perception, & Psychophysics, 81*(3), 764-777.

Anikin, A., Rudling, M., Persson, T., & Gärdenfors, P. (2018). Synesthetic associations between voice and gestures in preverbal infants: Weak effects and methodological concerns. *PsyArXiv*. https://doi.org/10.31234/osf.io/n2gvz

Arbib, M. A., Liebal, K., Pika, S., Corballis, M. C., Knight, C., Leavens, D. A., ... & Pika, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Current Anthropology, 49*(6), 1053-1076.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology, 25*(15), 2051-2056.

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2019). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General, 148*(10), 1842-1848.

August, P. V., & Anderson, J. G. (1987). Mammal sounds and motivation-structural rules: A test of the hypothesis. *Journal of Mammalogy, 68*(1), 1-9.

Aung, T., & Puts, D. (2019). Voice pitch: A window into the communication of social power. *Current Opinion in Psychology, 33*, 154-161.

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.

Baciadonna, L., Briefer, E. F., Favaro, L., & McElligott, A. G. (2019). Goats distinguish between positive and negative emotion-linked vocalisations. *Frontiers in Zoology, 16*(1), 25.

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

Bastian, A., & Schmidt, S. (2008). Affect cues in vocalizations of the bat, *Megaderma lyra*, during agonistic interactions. *The Journal of the Acoustical Society of America, 124*(1), 598-608.

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.

Bestelmeyer, P. E., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *Journal of Neuroscience, 34*(24), 8098-8105.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences, 113*(39), 10818-10823.

Blumstein, D. T., & Recapet, C. (2009). The sound of arousal: The addition of novel non-linearities increases responsiveness in marmot alarm calls. *Ethology, 115*(11), 1074-1081.

Braff, D. L., Geyer, M. A., & Swerdlow, N. R. (2001). Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies. *Psychopharmacology, 156*(2-3), 234-258.

Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots, 12*(1), 83-104.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology, 288*(1), 1-20.

Briefer, E. F., Mandel, R., Maigrot, A. L., Freymond, S. B., Bachmann, I., & Hillmann, E. (2017). Perception of emotional valence in horse whinnies. *Frontiers in Zoology, 14*(1), 8.

Bryant, G. A., Fessler, D. M., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., ... & De Smet, D. (2016). Detecting affiliation in colaughter across 24 societies. *Proceedings of the National Academy of Sciences, 113*(17), 4682-4687.

Bryant, G. A., Fessler, D. M., Fusaroli, R., Clint, E., Amir, D., Chávez, B., ... & Fux, M. (2018). The perception of spontaneous and volitional laughter across 21 societies. *Psychological Science, 29*(9), 1515-1525.

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

Calderon, L. E., Carney, L. D., & Kavanagh, K. T. (2016). The cry of the child and its relationship to hearing loss in parental guardians and health care providers. *Journal of Evidence-Informed Social Work, 13*(2), 198-205.

Cazau, D., Adam, O., Aubin, T., Laitman, J. T., & Reidenberg, J. S. (2016). A study of vocal nonlinearities in humpback whale songs: From production mechanisms to acoustic analysis. *Scientific Reports, 6*, 31660.

Charlton, B. D., & Reby, D. (2016). The evolution of acoustic size exaggeration in terrestrial mammals. *Nature Communications, 7*, 12739.

Charlton, B. D., Taylor, A. M., & Reby, D. (2013). Are men better than women at acoustic size judgements? *Biology Letters, 9*(4), 20130270.

Charrier, I., Aubin, T., & Mathevon, N. (2010). Mother-calf vocal communication in Atlantic walrus: A first field experimental study. *Animal Cognition, 13*(3), 471-482.

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.

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist, 74*(6), 698-712.

Crockford, C., Herbinger, I., Vigilant, L., & Boesch, C. (2004). Wild chimpanzees produce group-specific calls: A case for vocal learning? *Ethology, 110*(3), 221-243.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology, 22*(2), 142-146.

Deecke, V. B., Ford, J. K., & Spong, P. (1999). Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (*Orcinus orca*) dialects. *The Journal of the Acoustical Society of America, 105*(4), 2499-2507.

DiMattina, C., & Wang, X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *Journal of Neurophysiology, 95*(2), 1244-1262.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603-615.

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE, 8*(11), e78273.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology, 58*(2), 342-353.

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

Engelberg, J. W., & Gouzoules, H. (2019). The credibility of acted screams: Implications for emotional communication research. *Quarterly Journal of Experimental Psychology, 72*(8), 1889-1902.

Engelberg, J. W., Schwartz, J. W., & Gouzoules, H. (2019). Do human screams permit individual recognition? *PeerJ, 7*, e7087.

Erickson, D. (2005). Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology, 26*(4), 317-325.

Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology, 72*(2), 160-163.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour, 69*(3), 561-568.

Fichtel, C., Hammerschmidt, K., & Jürgens, U. (2001). On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. *Behaviour, 138*(1), 97-116.

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., ... & Newen, A. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences, 284*(1859), 20170990.

Fischer, J. (2011). Where is the information in animal communication. In R. Menzel & J. Fischer (eds.), *Animal thinking: Contemporary issues in comparative cognition* (pp. 151-161). Cambridge, MA: MIT Press.

Fischer, J., Wadewitz, P., & Hammerschmidt, K. (2017). Structural variability and communicative complexity in acoustic communication. *Animal Behaviour, 134*, 229-237.

Fitch, W. T. (2010). *The evolution of language*. New York: Cambridge University Press.

Fitch, W. T. (2018). The biology and evolution of speech: A comparative analysis. *Annual Review of Linguistics, 4*, 255-279.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour, 63*(3), 407-418.

Fraccaro, P. J., O'Connor, J. J., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour, 85*(1), 127-136.

Fröhlich, M., Müller, G., Zeiträg, C., Wittig, R. M., & Pika, S. (2017). Gestural development of chimpanzees in the wild: The impact of interactional experience. *Animal Behaviour, 134*, 271-282.

Frühholz, S., Trost, W., & Kotz, S. A. (2016). The sound of emotions—Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews, 68*, 96-110.

Fullard, J. H., Ratcliffe, J. M., & Soutar, A. R. (2004). Extinction of the acoustic startle response in moths endemic to a bat-free habitat. *Journal of Evolutionary Biology, 17*(4), 856-861.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8-11.

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.

Genty, E., Breuer, T., Hobaiter, C., & Byrne, R. W. (2009). Gestural communication of the gorilla (*Gorilla gorilla*): repertoire, intentionality and possible origins. *Animal Cognition, 12*(3), 527-546.

Genty, E., Clay, Z., Hobaiter, C., & Zuberbühler, K. (2014). Multi-modal use of a socially directed call in bonobos. *PLOS ONE, 9*(1), e84738.

Gobl, C., & Ní Chasaide, A. (2010). "Voice source variation and its communicative functions". In W. J. Hardcastle, J. Laver, & F. E. Gibbon (eds.). *The handbook of phonetic sciences (2nd ed.)* (pp. 378-423). Singapore: Wiley-Blackwell.

Goodall, J. (1986). *The chimpanzees of Gombe: Patterns of behavior*. Cambridge, MA: Harvard University Press.

Green, J. A., Whitney, P. G., & Potegal, M. (2011). Screaming, yelling, whining, and crying: Categorical and intensity differences in vocal expressions of anger and sadness in children's tantrums. *Emotion, 11*(5), 1124-1133.

Hamilton-Fletcher, G., Pisanski, K., Reby, D., Stefańczyk, M., Ward, J., & Sorokowska, A. (2018). The role of visual experience in the emergence of cross-modal correspondences. *Cognition, 175*, 114-121.

Hammerschmidt, K., & Fischer, J. (2019). Baboon vocal repertoires and the evolution of primate vocal diversity. *Journal of Human Evolution, 126*, 1-13.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M. (2017, March). CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 131-135).

Herzel, H., Berry, D., Titze, I., & Steinecke, I. (1995). Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 5*(1), 30-34.

Hobaiter, C., & Byrne, R. W. (2011). The gestural repertoire of the wild chimpanzee. *Animal Cognition, 14*(5), 745-767.

Jackson, D. E., & Ratnieks, F. L. (2006). Communication in ants. *Current Biology, 16*(15), R570-R574.

Johansson, N., Anikin, A., Carling, G., & Holmer, A. (in press). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*.

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.

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

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion. *Journal of Nonverbal Behavior, 42*(1), 1-40.

Kamiloglu, R., Fischer, A., & Sauter, D. A. (2019). Good vibrations: A review of vocal expressions of positive emotions. doi:10.31234/osf.io/86rmu. Preprint accessed from https://psyarxiv.com/86rmu/.

Karp, D., Manser, M. B., Wiley, E. M., & Townsend, S. W. (2014). Nonlinearities in meerkat alarm calls prevent receivers from habituating. *Ethology, 120*(2), 189-196.

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.

Koutseff, A., Reby, D., Martin, O., Levrero, F., Patural, H., & Mathevon, N. (2018). The acoustic space of pain: Cries as indicators of distress recovering dynamics in pre-verbal infants. *Bioacoustics, 27*(4), 313-325.

Lakoff, G., & Johnson, M. (2008[1980]). *Metaphors we live by*. Chicago: University of Chicago press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97*(3), 377-395.

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., ... & Baron-Cohen, S. (2019). The EU-emotion voice database. *Behavior Research Methods, 51*(2), 493-506.

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Iraki, F. K. E., ... & Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, 353.

Lavan, N., Domone, A., Fisher, B., Kenigzstein, N., Scott, S. K., & McGettigan, C. (2019). Speaker sex perception from spontaneous and volitional nonverbal vocalizations. *Journal of Nonverbal Behavior, 43*(1), 1-22.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior, 40*(2), 133-149.

Leavens, D. A., & Hopkins, W. D. (1998). Intentional communication by chimpanzees: a cross-sectional study of the use of referential gestures. *Developmental Psychology, 34*(5), 813-822.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron, 73*(4), 653-676.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-297.

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.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology, 58*(5), 698-726.

Linhart, P., Ratcliffe, V. F., Reby, D., & Špinka, M. (2015). Expression of emotional arousal in two different piglet call types. *PLOS ONE, 10*(8), e0135414.

Makagon, M. M., Funayama, E. S., & Owren, M. J. (2008). An acoustic analysis of laughter produced by congenitally deaf and normally hearing college students. *The Journal of the Acoustical Society of America, 124*(1), 472-483.

Manser, M. B. (2013). Semantic communication in vervet monkeys and other animals. *Animal Behaviour, 86*(3), 491-496.

Manteuffel, G., Puppe, B., & Schön, P. C. (2004). Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science, 88*(1-2), 163-182.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company.

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

McCune, L., Vihman, M. M., Roug-Hellichius, L., Delery, D. B., & Gogate, L. (1996). Grunt communication in human infants (*Homo sapiens*). *Journal of Comparative Psychology, 110*(1), 27-37.

Mcloughlin, M. P., Stewart, R., & McElligott, A. G. (2019). Automated bioacoustics: Methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface, 16*(155), 20190225.

Miller, G. (2011). *The mating mind: How sexual choice shaped the evolution of human nature*. New York: Anchor Books.

Moore, R. K. (2016). A real-time parametric general-purpose mammalian vocal synthesiser. In *INTERSPEECH* (pp. 2636–2640). Grenoble, France: International Speech Communication Association.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist, 111*(981), 855-869.

Neiberg, D., Laukka, P., & Elfenbein, H. A. (2011). Intra-, inter-, and cross-cultural classification of vocal affect. In *Twelfth Annual Conference of the International Speech Communication Association*. Florence, Italy, August 27-31, 2011.

Newman, J. D. (2007). Neural circuits underlying crying and cry responding in mammals. *Behavioural Brain Research, 182*(2), 155-165.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F$_0$ of voice. *Phonetica, 41*(1), 1-16.

Öhman, A. (1986). Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology, 23*(2), 123-145.

Oliva, M. & Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports, 8*(1), 4871.

Owren, M. J., & Bachorowski, J. A. (2003). Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior, 27*(3), 183-200.

Owren, M. J., & Rendall, D. (1997). An affect-conditioning model of nonhuman primate vocal signaling. In D. H. Owings, M. D. Beecher, & N. S. Thompson (eds.), *Communication. Perspectives in Ethology, vol 12*. Boston, MA: Springer.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology, 73*(6), 530-544.

Parsons, C. E., Young, K. S., Craske, M. G., Stein, A. L., & 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.

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition & Emotion, 28*(2), 230-244.

Pearce, J. M. (2008). *Animal learning and cognition: An introduction (3rd ed.)*. Hove, NY: Psychology Press.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37*(4), 417-435.

Pepperberg, I. M. (2006). Cognitive and communicative abilities of Grey parrots. *Applied Animal Behaviour Science, 100*(1-2), 77-86.

Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., & Feinberg, D. R. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Scientific Reports, 6*, 34389.

Pitcher, B. J., Mesoudi, A., & McElligott, A. G. (2013). Sex-biased sound symbolism in English-language first names. *PLOS ONE, 8*(6), e64825.

Prat, Y., Azoulay, L., Dor, R., & Yovel, Y. (2017). Crowd vocal learning induces vocal dialects in bats: Playback of conspecifics shapes fundamental frequency usage by pups. *PLOS Biology, 15*(10), e2002556.

Prat, Y., Taub, M., & Yovel, Y. (2016). Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. *Scientific Reports, 6*, 39419.

Provine, R. R. (2001). *Laughter: A scientific investigation*. New York: Penguin.

Provine, R. R. (2016). Laughter as a scientific problem: An adventure in sidewalk neuroscience. *Journal of Comparative Neurology, 524*(8), 1532-1539.

Puts, D. A. (2010). Beauty and the beast: Mechanisms of sexual selection in humans. *Evolution and Human Behavior, 31*(3), 157-175.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283-296.

Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D., Wheatley, J. R., ... & Jablonski, N. G. (2016). Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proceedings of the Royal Society B: Biological Sciences, 283*(1829), 20152830.

Raine, J., Pisanski, K., & Reby, D. (2017). Tennis grunts communicate acoustic cues to sex and contest outcome. *Animal Behaviour, 130*, 47-55.

Read, R., & Belpaeme, T. (2016). People interpret robotic non-linguistic utterances categorically. *International Journal of Social Robotics, 8*(1), 31-50.

Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour, 65*(3), 519-530.

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W. T., & Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society of London B: Biological Sciences, 272*(1566), 941-947.

Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour, 78*(2), 233-240.

Rendell, L. E., & Whitehead, H. (2003). Vocal clans in sperm whales (*Physeter macrocephalus*). *Proceedings of the Royal Society of London B: Biological Sciences, 270*(1512), 225-231.

Riede, T., Arcadi, A. C., & Owren, M. J. (2007). Nonlinear acoustics in the pant hoots of common chimpanzees (*Pan troglodytes*): Vocalizing at the edge. *The Journal of the Acoustical Society of America, 121*(3), 1758-1767.

Ross, M. D., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology, 19*(13), 1106-1111.

Rumbaugh, D. M., & Savage-Rumbaugh, E. S. (1994). Language in comparative perspective. In N. J. Mackintosh (ed.), *Animal learning and cognition. Handbook of perception and cognition series, 2nd ed.* (pp. 307-333). San Diego: Academic Press.

Ryan, M. J., & Cummings, M. E. (2013). Perceptual biases and mate choice. *Annual Review of Ecology, Evolution, and Systematics, 44*, 437-459.

Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience, 22*(3), 474-481.

Sauter, D. A., & Fischer, A. H. (2018). Can perceivers recognise emotions from spontaneous expressions? *Cognition and Emotion, 32*(3), 504-515.

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*(3), 192-199.

Sauter, D. A., Crasborn, O., Engels, T., Kamiloğlu, R. G., Sun, R., Eisner, F., & Haun, D. B. M. (2019). Human emotional vocalizations can develop in the absence of auditory learning. *Emotion*. doi: 10.1037/emo0000654

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.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107*(6), 2408-2412.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2002). Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice, 16*(4), 509-529.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2006). Vocal expression of emotions in normally hearing and hearing-impaired infants. *Journal of Voice, 20*(4), 585-604.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

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

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (eds.), *Blueprint for affective computing: A sourcebook* (pp. 166-178). Oxford: Oxford University Press.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology, 32*(1), 76-92.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40*(1-2), 99-116.

Schwartz, J. W., Engelberg, J. W., & Gouzoules, H. (2019). What is a scream? Listener agreement and major distinguishing acoustic features. *Journal of Nonverbal Behavior*, 1-20. doi:10.1007/s10919-019-00325-y

Scott-Phillips, T. (2015). Nonhuman primate communication, pragmatics, and the origins of language. *Current Anthropology, 56*(1), 56-80.

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

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

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science, 210*(4471), 801-803.

Seyfarth, R., & Cheney, D. (2018). Pragmatic flexibility in primate vocal production. *Current Opinion in Behavioral Sciences, 21*, 56-61.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621-640.

Slocombe, K. E., Kaller, T., Call, J., & Zuberbühler, K. (2010). Chimpanzees extract social information from agonistic screams. *PLOS ONE, 5*(7), e11473.

Snowdon, C. T. (2009). Plasticity of communication in nonhuman primates. *Advances in the Study of Behavior, 40*, 239-276.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971-995.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition (Vol. 142)*. Cambridge, MA: Harvard University Press.

Stegmann, U. E. (2013). Introduction: A primer on information and influence in animal communication. In U. E. Stegmann (ed.), *Animal communication theory: Information and influence* (pp. 1-39). Cambridge: Cambridge University Press.

Stowell, D., & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ, 2*, e488.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., & Sterr, A. (2009). Differentiation of emotions in laughter at the behavioral level. *Emotion, 9*(3), 397-405.

Tanner, J. C., Justison, J., & Bee, M. A. (2019). SynSing: Open-source MATLAB code for generating synthetic signals in studies of animal acoustic communication. *Bioacoustics*, 1-22. doi:10.1080/09524622.2019.1674694.

Terrace, H. S., Petitto, L. A., Sanders, R. J., & Bever, T. G. (1979). Can an ape create a sentence? *Science, 206*(4421), 891-902.

Townsend, S. W., & Manser, M. B. (2010). The function of nonlinear phenomena in meerkat alarm calls. *Biology Letters, 7*(1), 47-49.

Van Hooff, J. A., & Preuschoft, S. (2003). Laughter and smiling: The intertwining of nature and culture. In F. deWaal & P. Tyack (eds.), *Animal social complexity: intelligence, culture, and individualized societies* (pp. 261-287). Cambridge: Harvard University Press.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., ... & Johnstone, T. (2004). Human amygdala responsivity to masked fearful eye whites. *Science, 306*(5704), 2061-2061.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews, 21*(5), 195-205.

Whitham, J. C., & Wielebnowski, N. (2013). New directions for zoo animal welfare science. *Applied Animal Behaviour Science, 147*(3-4), 247-260.

Wood, A., Martin, J., & Niedenthal, P. (2017). Towards a social functional account of laughter: Acoustic features convey reward, affiliation, and dominance. *PLOS ONE, 12*(8), e0183811.

Wood, A. (2019, December 7). Social context influences the acoustic properties of laughter. https://doi.org/10.31234/osf.io/npk8u

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology, 53*(1), 205-214.

Zeifman, D. M. (2001). An ethological analysis of human infant crying: Answering Tinbergen's four questions. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 39*(4), 265-285.

Zuberbühler, K. (2015). Linguistic capacity of non-human animals. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(3), 313-321.

Zúñiga, J., & Reiss, J. D. (2019). Realistic procedural sound synthesis of bird song using particle swarm optimization. In *Audio Engineering Society Convention e-Brief 555*, Oct 16-19 2019, New York.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185.

Paper I

Behav Res (2017) 49:758–771
DOI 10.3758/s13428-016-0736-y

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

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

Published online: 29 April 2016

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

Keywords Emotion · Nonlinguistic vocalizations · Naturalistic vocalizations · Acoustic analysis

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Method

Compilation of the corpus

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

Priority was given to eliciting contexts that were:

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

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

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

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

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

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

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

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

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

Preparation of audio clips

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

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

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

Methods of acoustic analysis

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

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

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

Statistical modeling based on acoustic measurements

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

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

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

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

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

Rating experiment

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

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

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

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

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

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

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

Results

Rating task

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

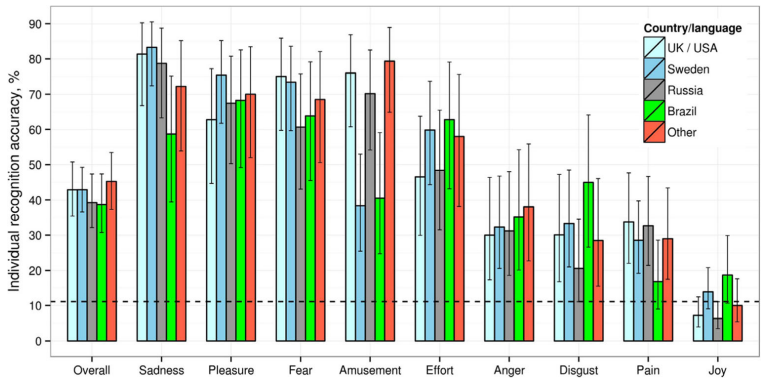

Fig. 1 Individual recognition accuracies by emotion and the rater's linguistic–cultural group: Posterior medians and 95% confidence intervals. The dotted line shows the probability of guessing correctly by chance (11%)

Acoustic analysis

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

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

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

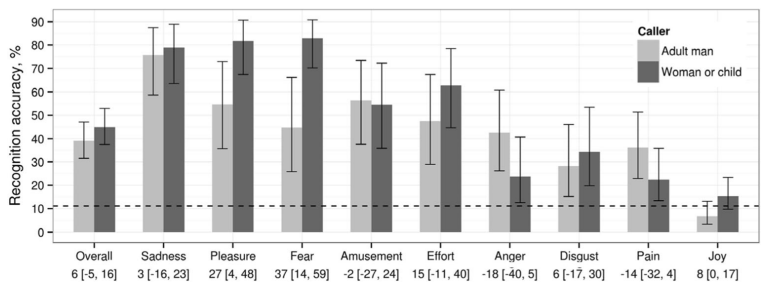

Fig. 2 Individual recognition accuracies by emotion and the caller's gender–age group: Posterior medians and 95% CI's. The most credible difference (%) between gender–age groups is shown underneath each bar, with 95% CI. The dotted line shows chance level (11%)

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

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

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

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

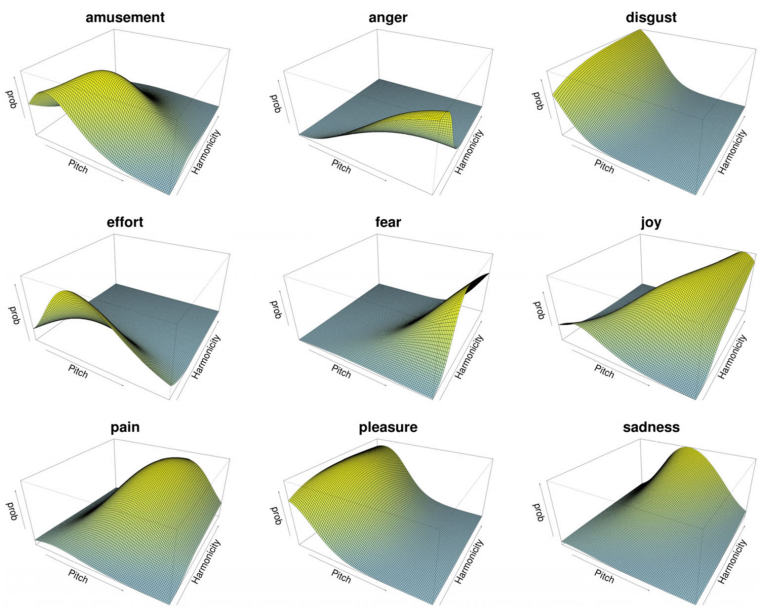

Fig. 3 Perspective plots showing the predicted probability that a sound of particular pitch and harmonicity belongs to each of the nine emotional categories. The probabilities were calculated with multinomial logistic regression. For each point on the pitch–harmonicity plane, the nine probabilities sum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

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

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

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

Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

Conclusions

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Paper II

Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–20

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 ($2.1\,s \pm 1.5$) than acted sounds ($1.5\,s \pm 1.1$), $t(351.1) = 3.7$, $p < .001$.

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

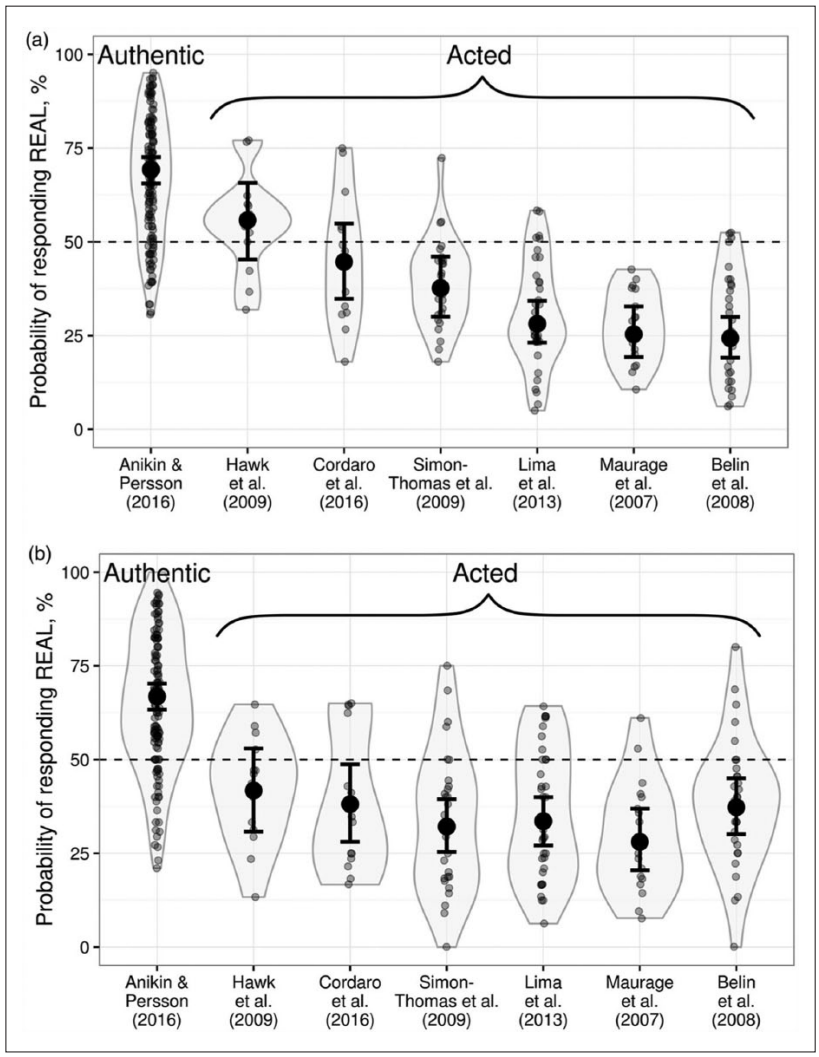

Figure 1. Perceived authenticity of sounds from each corpus collapsed across all emotions and calculated as the proportion of trials in which the sound was classified as "real" rather than "fake", (a) without and (b) with masking noise. Small dots show the authenticity of individual sounds, while the 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, $OR = 0.97$

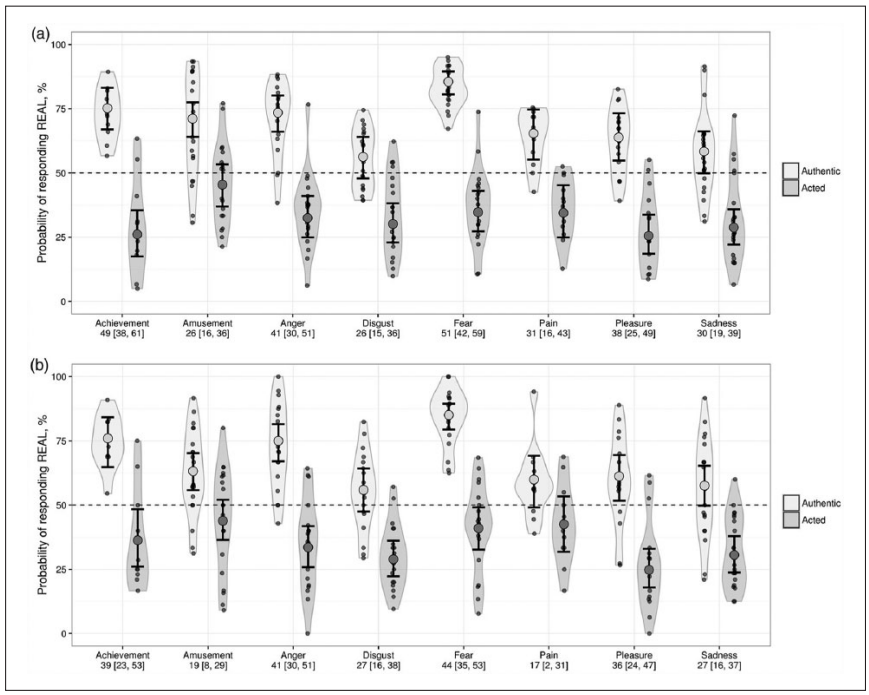

Figure 2. Perceived authenticity of authentic versus acted sounds in each emotional category calculated as the proportion of trials in which the sound was classified as "real" rather than "fake", (a) without and (b) with masking noise: median of posterior distribution and 95% confidence interval (CI). 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

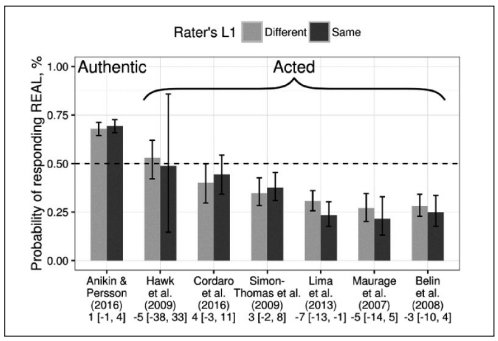

Figure 3. Perceived authenticity of sounds from each corpus: median of posterior distribution and 95% confidence interval (CI). The most credible difference (%) between raters with the same versus different first language as the caller's is shown under each emotion, with 95% CI. The dotted line at 50%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

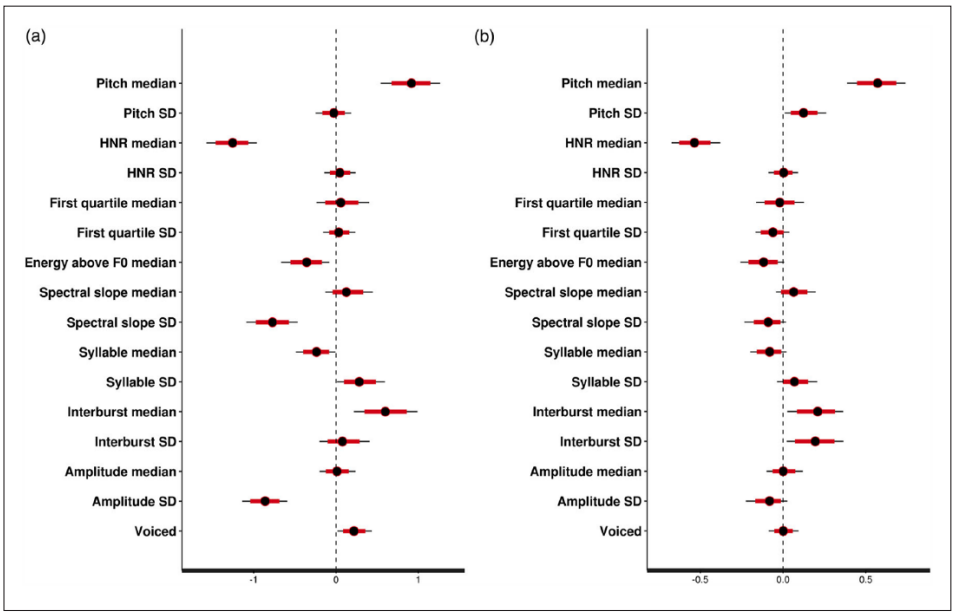

Figure 4. Standardized beta coefficients on a logit scale with 80% and 95% confidence interval (CI) in multiple logistic regression. The outcome variable is (a) the status of each sound as authentic or (b) the real/fake subjective judgment provided by participants in the behavioural experiment. Both 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

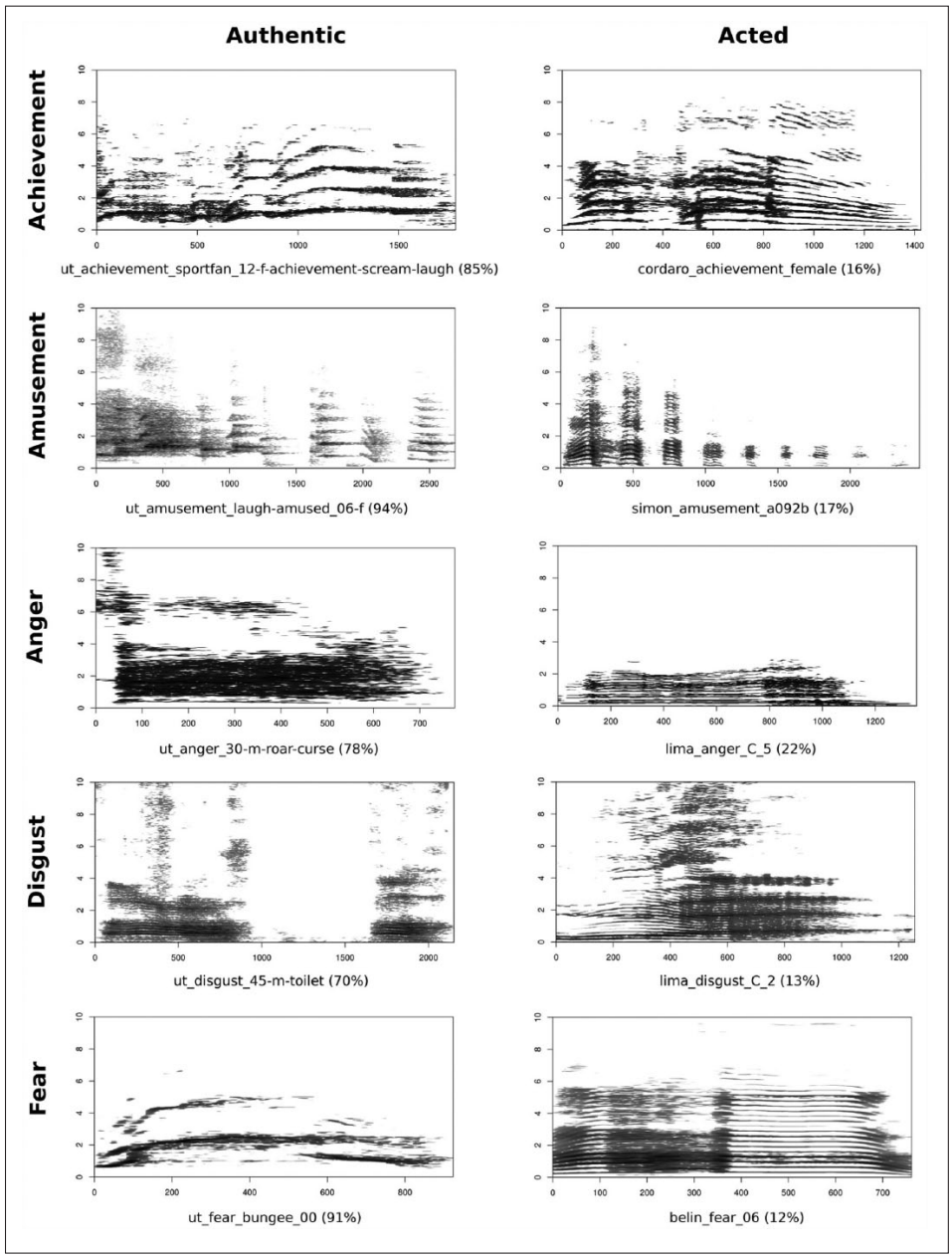

Figure 5. Continued.

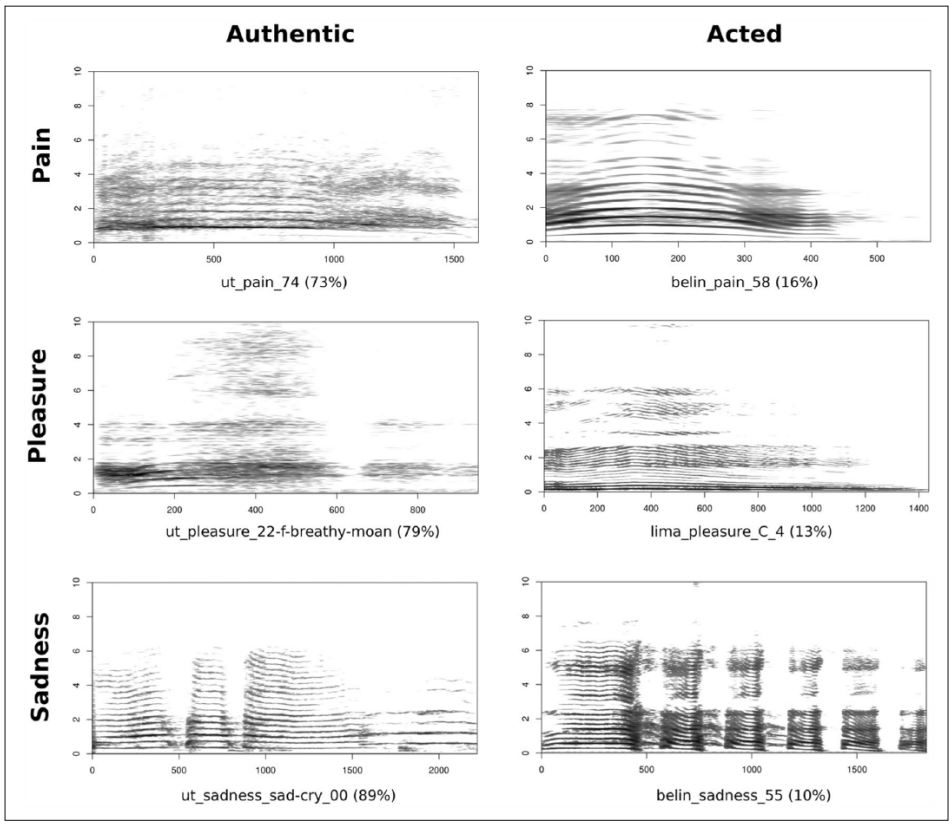

Figure 5. Contrast pairs of sounds with high and low perceived authenticity. Shown: emotion, spectrogram (time in ms, frequency in kHz), file name, and proportion of participants who responded "real".

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.

Paper III

J Nonverbal Behav (2018) 42:53–80
https://doi.org/10.1007/s10919-017-0267-y

ORIGINAL PAPER

Human Non-linguistic Vocal Repertoire: Call Types and Their Meaning

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

Published online: 30 September 2017

Abstract Recent research on human nonverbal vocalizations has led to considerable progress in our understanding of vocal communication of emotion. However, in contrast to studies of animal vocalizations, this research has focused mainly on the emotional interpretation of such signals. The repertoire of human nonverbal vocalizations as acoustic types, and the mapping between acoustic and emotional categories, thus remain underexplored. In a cross-linguistic naming task (Experiment 1), verbal categorization of 132 authentic (non-acted) human vocalizations by English-, Swedish- and Russian-speaking participants revealed the same major acoustic types: laugh, cry, scream, moan, and possibly roar and sigh. The association between call type and perceived emotion was systematic but non-redundant: listeners associated every call type with a limited, but in some cases relatively wide, range of emotions. The speed and consistency of naming the call type predicted the speed and consistency of inferring the caller's emotion, suggesting that acoustic and emotional categorizations are closely related. However, participants preferred to name the call type before naming the emotion. Furthermore, nonverbal categorization of the same stimuli in a triad classification task (Experiment 2) was more compatible with classification by call type than by emotion, indicating the former's greater perceptual salience. These results suggest that acoustic categorization may precede attribution of emotion, highlighting the need to distinguish between the overt form of nonverbal signals and their interpretation by the perceiver. Both within- and between-call acoustic variation

Electronic supplementary material The online version of this article (doi:10.1007/s10919-017-0267-y) contains supplementary material, which is available to authorized users.

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

Rasmus Bååth
rasmus.baath@gmail.com

Tomas Persson
tomas.persson@lucs.lu.se

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

can then be modeled explicitly, bringing research on human nonverbal vocalizations more in line with the work on animal communication.

Keywords Emotion · Non-linguistic vocalizations · Semantic spaces · Cross-linguistic naming study · Triad classification task

Introduction

Emotion is an essential part of being human and a matter of great theoretical and clinical significance. It has justifiably attracted a lot of attention in psychology and neuroscience, including research on facial expressions (Ekman et al. 1969; Izard 1994), prosody (Banse and Scherer 1996), and non-linguistic vocalizations (Belin et al. 2008; Lima et al. 2013). This abiding interest in nonverbal communication has shed light on how affective states can be expressed without words; on the other hand, the most obvious level of analysis, namely the surface form of the signals themselves, has received far less attention.

The relative neglect of alternative, non-affective categories in nonverbal communication may prove a liability, because such categories are both intuitively appealing and useful for research. For example, when researchers analyze the differences between authentic and posed laughter (Bryant and Aktipis 2014; Lavan et al. 2015), evolutionary adaptive value of crying (Provine et al. 2009), or unique acoustic signatures of screaming (Arnal et al. 2015), they implicitly refer to these sounds as acoustic categories that are somehow different from each other and from other sounds. Using the terminology common in animal research, laughter, vocal crying, and screaming are treated as distinct vocalizations, or "call types". Is this classification justified? In what sense is laughter a call type? What other call types do humans have? A systematic analysis of these issues is the goal of this study.

We begin by justifying the applicability of the concepts and methods of ethology to the study of human non-linguistic vocalizations. We then review the available evidence on the types of vocalizations in human vocal repertoire and present the results of two perceptual experiments that contrast the classification of non-linguistic sounds in terms of emotion and in terms of acoustic categories. Our key objective is to test the hypothesis that acoustic categories (such as a laugh, a scream, etc.) are salient to listeners and not equivalent to affective states.

Non-linguistic Vocalizations from an Ethological Perspective

Despite some important exceptions (Oller and Griebel 2008; Watson et al. 2015), the acoustic structure of animal vocalizations is largely determined on a genetic level, so that all members of a species produce essentially the same vocalizations (Owren et al. 2011; Wheeler and Fischer 2012). In contrast, human language is not only unusually flexible and powerful as a communicative tool (Devitt and Sterelny 1999), but it is also entirely dependent on a socially transmitted, culture-specific code: We are not born speaking English or Javanese. At the same time, the privileged status of language should not blind us to the fact that many of the sounds humans produce in everyday life are non-linguistic (Provine 2012). There is mounting evidence (reviewed below) that non-linguistic sounds such as laughter are more similar to the vocalizations of other mammals than they are to human language. This evidence comes from neurological research on vocal production and psychological research on vocal perception.

To begin with production, it is well established that the vocal flexibility associated with mastery of language does not preclude the existence of separate, phylogenetically older neural networks responsible for the production of non-linguistic vocalizations (Ackermann et al. 2014; Jürgens 2009). Aphasic patients with lesions in motor cortex (Jürgens 2009) as well as congenitally deaf (Scheiner et al. 2006) and even unencephalic (Newman 2007) infants may laugh and moan just like typical infants. This is possible because non-linguistic vocalizations are controlled by dedicated circuits deep in the brain stem, whereas speech relies on a separate pathway leading from motor cortex directly to laryngeal motoneurons (Jürgens 2009). This separate neural control mechanism explains why it is hard to laugh or cry at will (Provine 2012)—these vocalizations are normally not under direct volitional control. The currently available neurological evidence is not sufficiently detailed to determine precisely how many species-typical vocalizations humans have and what these vocalizations are. However, since the neural machinery controlling vocalizing in mammals is known to be evolutionarily stable (Ackermann et al. 2014), at least some human vocalizations should have direct analogs in the calls of other mammals, and some likely candidates are being investigated (see below).

Moving on to perception, numerous studies have demonstrated that listeners can extract a lot of useful information from sounds that contain little or no phonemic structure. At least eight (Belin et al. 2008; Lima et al. 2013) and perhaps as many as 14–16 (Cordaro et al. 2016; Simon-Thomas et al. 2009) affective states can be correctly identified if a person is instructed to portray them without resorting to language. In addition, listeners can discriminate between authentic and posed emotion (Anikin and Lima 2017; Bryant and Aktipis 2014) or judge whether two people laughing together are friends or strangers (Bryant et al. 2016). Vocalizations of pain (Belin et al. 2008) and physical effort (Anikin and Persson 2017) are also easily recognizable. The information available from non-linguistic vocalizations is thus rich and not strictly limited to emotion.

Recognition accuracy in cross-cultural studies tends to be slightly higher when the speaker belongs to the same group (Elfenbein and Ambady 2002; Koeda et al. 2013; Laukka et al. 2013), demonstrating that even non-linguistic sounds have some culture-specific component. Nevertheless, listeners in even the most isolated communities with little exposure to Western media recognize the emotion expressed by non-linguistic vocalizations of Westerners at above-chance levels (Cordaro et al. 2016; Sauter et al. 2010). The signal system involving laughs and moans is thus much more universal than the expressions of a given language. This raises the question of what the meaningful units of this signal system might be. Are they word-like, as in language?

To answer this question, we need to understand what makes non-linguistic sounds meaningful. A commonly used method is to investigate the communicative significance of particular acoustic features. For example, pitch, intensity, and duration increase with arousal in both human (Scheiner et al. 2002) and animal (Briefer 2012) vocalizations; harsh, noisy sounds are perceived as more aggressive compared to tonal sounds (Anikin and Persson 2017; August and Anderson 1987); authentic vocalizations have more acoustic variability than posed vocalizations (Anikin and Lima 2017; Lavan et al. 2015), and so on. However, it must still be determined how all this potentially available acoustic information is processed. One possibility is that listeners go from acoustic features directly to a model of the speaker's emotional state and intentions, perhaps mapping sounds to discrete emotional states (e.g., Ekman's basic emotions, 1992) or dimensions such as valence and arousal (Briefer 2012; Russell 1980). Alternatively, the interpretation of a vocalization could be mediated by its acoustic classification: first we recognize that we hear a laugh and

then decide whether this is a laugh of genuine amusement, mere social politeness, an "evil" laugh, etc. (Provine 2001).

If acoustic classification indeed mediates interpretation of vocalizations, acoustic categories should be highly salient. In several studies of non-linguistic vocalizations (Anikin and Persson 2017; Gendron et al. 2014a) participants sometimes hesitated to attribute any particular emotion to the caller, while confidently naming the sound (e.g., as a laugh or a scream). In the visual domain, a similar dissociation has been observed between naming a facial expression and interpreting its emotional significance (Boster 2005; Gendron et al. 2014b). The receiver of a communicative display, such as a vocalization or a facial expression, may thus recognize and classify the signal itself (e.g., as a laugh or a scowl) without attributing any particular emotion to the signaler. As a result, descriptions of signals in terms of their surface form (a laugh, a scowl) and in terms of their meaning (merriment, annoyance) are complementary rather than redundant. A possible theoretical interpretation is that the identification of a communicative display precedes its attribution to a particular social or emotional cause. According to this view, sometimes known as the Identification-Attribution model, these two processes may reflect a neurological division of labor between different systems (Sperduti et al. 2014; Spunt and Lieberman 2012).

There are thus some indications that affective states may not be the only, or even the most appropriate, categories for describing the repertoire of nonverbal communicative displays. In the case of non-linguistic vocalizations, there is also a natural alternative to emotional categories, namely acoustic categorization of sounds in terms of call types.

Human Call Types

Despite its central significance in studies of animal communication, the concept of call type has no generally accepted definition. It refers to distinguishable acoustic units ("calls") that together comprise a species' vocal repertoire, but the exact nature and number of such units depend on whether the main interest is in production or perception, on the chosen method of classification, on the extracted acoustic variables, and so on (Fischer et al. 2016; Kershenbaum et al. 2014). Primate vocalizations are particularly challenging to categorize, because they tend to grade into each other acoustically (Marler 1976; van Hooff and Preuschoft 2003), and because they possess a high amount of within-call variability, complicating the task of identifying discrete vocalizations with objective statistical measures (Fischer et al. 2016; Wadewitz et al. 2015). Following the ethological tradition, we provisionally define call types as distinct species-typical vocalizations whose basic spectral-temporal structure is innate (not learned).

What ultimately makes vocalizations "distinct" is their unique neurological production mechanism. In practice, however, animal researchers seldom have access to this information, so they have little choice but to record a large number of vocalizations and deduce the underlying call types, normally by means of comparing the acoustic structure and typical context in which each sound occurs (Kershenbaum et al. 2014). Some distinctions that are salient to the animals themselves may be lost in the process, because the human ear or the analytic technique misses them (lumping), and some spurious distinctions may be found between what is actually a single vocalization (splitting). Likewise, it is unclear to what extent perceptual distinctions, whether they are made by the researcher or by the animal itself, correspond to call types defined by their unique production mechanism. The task of studying human vocal behavior is further complicated by the fact that species-typical vocalizations such as laughter coexist with language and semi-linguistic interjections (such as *urgh*, *ouch*, etc.). The silver lining for researchers working with human

sounds is that they have more methodological options at their disposal: unlike animal subjects, human participants can be asked to label the stimuli verbally or to classify them in some other way, providing direct access to perceptual categories distinguished by the listeners.

The research on production and perception of human vocalizations reviewed in the previous section indicates that some vocalizations such as laughter are innate—that is, their acoustic form and, to some extent, meaning are predetermined by our genetic endowment. As a result, researchers are increasingly looking for the evolutionary roots of such vocalizations, usually by comparing them with the vocal repertoire of other primates (Provine 2001; Sauter et al. 2010; Scheumann et al. 2014). By definition, the unit of analysis in such phylogenetic reconstructions is an acoustic category rather than an emotion, and the two best-known examples are laughter and vocal crying.

Laughter presumably originated in mammalian social play (van Hooff and Preuschoft 2003; Provine 2001). Acoustically, this vocalization is recognizable above all by its distinct rhythm with approximately five syllables per second (Bryant and Aktipis 2014; Provine 2001). Unlike the ingressive–egressive laughter of the great apes, humans laugh with several syllables produced on a single exhalation (Provine 2001). Nevertheless, acoustic and contextual similarities are sufficiently strong to claim that laughter is a vocalization that humans share with other great apes (Ross et al. 2009) and perhaps even with rats (Panksepp 2007). Vocal crying is another human vocalization with clear evolutionary parallels. Several studies have indicated that crying in humans is related to mother-infant separation or distress calls, which are common in many mammalian species (Lingle et al. 2012; Newman 2007; Provine 2012). In contrast to laughter, crying consists of longer voiced syllables repeated at intervals approximately corresponding to respiratory cycles (Provine 2012). The sound of crying is typically tonal, with a pronounced harmonic structure, but it may also include noisy episodes (Lingle et al. 2012). This variation within the same basic acoustic template (within-call variation) is highly informative in cries of human infants (Scheiner et al. 2002) as well as animals (Lingle et al. 2012).

Laughter and vocal cry are thus two call types whose species-typical nature in humans is widely accepted and whose evolutionary origins are relatively clear. But to complete the puzzle, we have to learn what other call types, if any, the human vocal repertoire includes. Naming studies offer a powerful method for identifying perceptually salient acoustic categories and their meaning, and we utilized this technique in addition to performing acoustic analysis (Experiment 1). However, a linguistic approach is not without its pitfalls (see the Introduction to Experiment 2), and therefore we also performed a triad classification study, which allowed us to investigate the categorization of non-linguistic vocalizations without using any verbal labels (Experiment 2). It is worth reiterating that perceptual studies can only reveal the categories distinguished by listeners, which may or may not correspond to the underlying "true" call types (i.e., vocalizations with unique, genetically determined neurological production mechanisms and evolutionary histories). Clustering based on acoustic measurements is likewise not guaranteed to produce an "objective" classification, because call types may be graded and because the choice of acoustic variables affects the outcome. In this regard, human acoustic research is not very different from the studies of vocal communication in other mammals, and the same caution is needed when interpreting its results.

To the best of our knowledge, no study has systematically analyzed the repertoire of human non-linguistic vocalizations from this acoustic perspective, only the emotional states that they convey. As a result, there is little empirical data on our chosen research questions:

1. What acoustic categories do listeners distinguish in the wide variety of human non-linguistic vocalizations?
2. To what extent is this acoustic categorization language-specific?
3. How closely does acoustic categorization map onto emotional categorization?
4. What cognitive model best describes the relation between acoustic and emotional categorization of vocalizations?

Source of Sounds

Our research questions require that we compare acoustic and emotional categorizations of non-linguistic vocalizations. In particular, we would like to learn whether these classifications are relatively independent or redundant (e.g., whether each acoustic category closely corresponds to a single emotion), and whether one of them precedes the other. This task calls for a novel approach to collecting the audio material. Vocalizations in most existing corpora are either elicited from people who are verbally instructed to portray a particular emotion, or they are induced by an experimental manipulation (Scherer 2013). For a project aiming to describe the repertoire of non-linguistic vocalizations and investigate their association with emotion, this type of material presents three problems:

(1) There is evidence that listeners can distinguish between authentic and acted vocalizations (Anikin and Lima 2017; Bryant and Aktipis 2014; Gervais and Wilson 2005). This raises concerns about the latter's ecological validity, suggesting that voluntarily produced vocalizations in some cases may deviate from the natural, spontaneous form.

(2) Listeners can extract more information from vocalizations produced by members of the same cultural group, indicating that there is important cultural variation in human non-linguistic vocalizations (Elfenbein and Ambady 2002; Koeda et al. 2013; Laukka et al. 2013). This may be problematic if the research interest concerns species-specific, rather than culture-specific, vocalizations.

(3) Acted vocalizations are typically elicited by providing participants with short vignettes or asking them to imagine a scenario targeting a particular emotion (Scherer 2013), and the recordings are then validated in a multiple-choice task, often preserving only a subset with the highest recognition rate. Each vocalization in the final corpus is thus designed to be a maximally transparent vehicle for the expression of a single emotional state. This excludes sounds—presumably abundant in real life—that accompany a complex, mixed emotional experience (e.g., a blend of fear, anger, and pain experienced by someone in a fight) as well as vocal expressions not typically associated with affect (e.g., grunts of physical effort or the trembling whine of a person freezing at a bus stop).

To avoid these limitations of most available corpora, the ideal source of sounds for the current project would be a large corpus of observational material, recorded in culturally diverse locations and not tied to particular emotional states. To our knowledge, no such "perfect" collection of human vocalizations exists. As a reasonable compromise, we chose to work with the observational corpus compiled from social media and validated by Anikin and Persson (2017), which contains 260 authentic vocalizations from a wide variety of contexts. It has the advantage of containing many intense and potentially hard-to-fake (Anikin and Lima 2017) vocalizations associated with acute fright, injury, genuinely funny incidents, etc. This makes it more likely that the available material extends to extreme and

socially inappropriate vocalizations. Many of these sounds may be associated with mixed emotional states (Anikin and Persson 2017), making them more realistic objects for investigating the mapping between acoustic and emotional categorizations compared to actor portrayals of discrete emotions. This corpus also goes beyond the traditional range of contexts in emotion research and includes vocalizations of pain and physical effort (for a list of contexts and audio files, see Electronic Supplementary Materials).

Experiment 1

In this cross-linguistic naming study, participants heard non-linguistic vocalizations from real-life interactions and chose one or more verbal labels to describe each sound in terms of its acoustics (e.g., a laugh, a moan, etc.) and emotion (e.g., amusement, pleasure). To our knowledge, naming studies have not been used in this manner to compare categorizations of emotional displays in different languages. The research on facial expressions (Ekman et al. 1969; Izard 1994) is different in that it focused on cross-cultural recognition of particular emotions, rather than on the categorization of facial behavior in each language. More relevant to our purpose, there is a growing body of cross-linguistic research in domains other than emotion, such as color (Berlin and Kay 1991), body parts (Enfield et al. 2006), locomotion (Malt et al. 2010), and verbs of breaking-cutting (Majid et al. 2008). The principal technique, known as the Nijmegen method (Slobin et al. 2014), is to elicit free descriptions of events or objects. The more often participants apply the same name to two stimuli, the more similar these two stimuli are assumed to be. A low-dimensional representation of these similarities together with lexical information may be referred to as a conceptual space, semantic space, or semantic map (on terminological distinctions, see Zwarts 2010). Languages are compared in terms of the overall structure of their respective semantic spaces as well as the extensions and prototypical core meanings of particular words (Zwarts 2010).

Semantic spaces may include both gradients and discontinuities. Where important natural discontinuities exist, languages are likely to make a categorical distinction. For instance, speakers of different languages agree on the exact transition point between walking and running, demonstrating a clear categorical distinction between these two modes of locomotion (Malt et al. 2010). In contrast, the distinctions are more likely to be language-specific in domains containing gradients with no abrupt discontinuities. For example, within each of the two basic gaits of walking and running, there is a continuum carved up differently by different languages (Slobin et al. 2014). Despite this general rule, continuous domains may also have natural attractors, so that categories in different languages may have the same best exemplars. For instance, while the range of hues falling under the local term for "red" varies considerably across languages, people in most societies agree on what constitutes a good example of pure red. It is therefore generally accepted that focal colors are universal, probably because of the physiology of human vision (Berlin and Kay 1991; Lindsey and Brown 2009).

By applying this linguistic method combined with acoustic analysis to non-linguistic vocalizations, we aimed to address the first two research questions, namely to identify the most salient call types distinguished by listeners and to compare this categorization in different languages. If some human vocalizations are species-typical, as is often suggested (Provine 2001; Ross et al. 2009; Sauter et al. 2010; Scheumann et al. 2014), we hypothesized that they should be recognized cross-culturally as distinct perceptual categories. The

semantic spaces of sound names should thus have comparable global configurations in different languages, although the number of subdivisions within each major category and the extensions of different terms could be language-specific.

In addition to naming each sound, we also asked participants to interpret it emotionally. This allowed us to explore the mapping of call types to emotions and to address research question 3, namely to test whether: (a) there is a close correspondence between the perceived call type and the perceived emotion, or (b) acoustic and emotional categorizations are relatively independent (non-redundant).

Finally, to shed some light on the cognitive processes involved in the interpretation of non-linguistic vocalizations (research question 4), we tested whether there would be any preference to perform the acoustic and emotional categorization of vocalizations in a particular order, and whether these naming decisions would differ in speed, subjective certainty, and consistency. The Identification-Attribution model predicts that the surface form of the communicative signal—its call type—should be identified first, followed by a more elaborate interpretation in terms of the feelings and goals of the vocalizer. Alternatively, acoustic and emotional categorizations could represent two independent processes that run in parallel rather than sequentially. In this case we should not find a consistent temporal relationship or a strong correlation between the ease of categorizing a particular sound by acoustic type and by emotion.

In pilot tests we initially followed the Nijmegen method (Slobin et al. 2014) and elicited free-text descriptions of each sound. With this design, sounds are classified in an inductive manner: each participant creates their own categories for classifying the stimuli. Our participants volunteered a manageable number of sound names, but emotion names contained many synonyms, and there was a tendency to provide complex descriptions of the hypothetical context in which vocalizing took place instead of monolexemic labels (cf. Boster 2005). We strove to keep the two naming tasks compatible and therefore opted to provide participants with a list of monolexemic sound names and emotion names that were commonly used by participants in the pilot study.

By analogy with Berlin and Kay's (1991) technique for eliciting basic color terms, we were less interested in polylexemic descriptions, very low-frequency words, terms that are mostly applicable to animal but not human sounds, and recent foreign loans. To make sure the list of sound names was comprehensible, we also checked the frequencies of all potential sound names in English, Swedish, and Russian, whether or not these words were actually used by participants in the pilot study. All high-frequency words were included in the labels (see Electronic Supplementary Materials). Eventually we chose 16 sound names in English, but in Swedish and Russian this would have required including some uncommon words, so we reduced the number of sound names to 12. The list of emotion labels in all three languages included 16 terms (see Fig. 3 for a complete list of labels for each language).

Materials and Methods

Stimuli

The experimental stimuli consisted of 132 authentic non-linguistic vocalizations (63 by men, 69 by women and children), which were selected by stratified random sampling from a larger, previously validated corpus (Anikin and Persson 2017). This corpus was compiled from online videos of people engaged in a variety of emotionally charged and easily interpretable activities, such as cleaning a blocked toilet or eating exotic foods (disgust),

playing with distorting web cameras or watching a friend take a spectacular tumble (amusement), lifting heavy weights (effort), and so on, for a total of nine categories: amusement, anger, disgust, effort, fear, joy, pain, pleasure, and sadness. Strictly speaking, these categories are contextual-emotional, since we only know in what context the vocalization was emitted, not the "true" affective state of the caller. The sounds were on average 2.2 ± 1.8 s in duration. The callers were primarily English speakers, but we tried to avoid language-specific emblems such as *ouch*, *yuck*, etc. For the most part, the tested sounds are thus free from any phonemic structure.

Participants

Participants ($N = 64$) were mono- or bilingual speakers of Swedish ($n = 20$), English ($n = 19$), or Russian ($n = 25$). The Swedish- and English-speaking participants were recruited among students and junior staff at Lund University and tested in person, ensuring that every participant rated all 132 stimuli. Russian participants were recruited and tested online, resulting in some incomplete reports (18 out of 25 Russian participants completed over 85% of trials).

Procedure

The experiment was performed in a web browser (See Electronic Supplementary Materials, Figure A1). Participants chose one or more suitable sound names and emotion names from a list of alternatives. This task is different from the inductive categorization in the pilot tests, since participants chose among a limited number of provided categories. They could change their minds and correct their answers as many times as needed, until they clicked the *Next* button and moved on to the next sound. It took 30–40 min to rate 132 sounds.

To assess the facility of naming acoustic types and emotions, participants could have been asked to do these two tasks sequentially, in random order. However, responses were generally slow (mean total time for both tasks 25 s), making it hard to know which processes might be responsible for differences in response times. We therefore opted to present both sound names and emotion names on the same screen, which allowed us not only to measure response times, but also to evaluate individual preferences for starting by naming either the sound or the emotion. To control for the general tendency to start with the left-hand side of the screen, for half of the participants in each language sound names were on the left-hand side, and emotion names were on the right-hand side of the screen. For the other half of participants, this order was reversed.

Statistical Analysis

All analyses were performed in R (R Core Team 2016).

Semantic Space of Sound Names In order to construct semantic spaces representing the perceptually salient acoustic categories and dimensions along which they are distinguished, we calculated pairwise Euclidean distances between all stimuli based on how often participants chose the same name for two sounds. This was done separately for each language, after which the resulting distance matrices were averaged across languages. Distance matrices were analyzed using principal components analysis (PCA) and multi-dimensional scaling (MDS). We defined a cluster as a group of stimuli with the same most commonly

chosen name. Centroids were calculated by taking a weighted mean of the coordinates of all sounds in a cluster, using as weights a product of (1) the average subjective certainty with which each sound was named by participants and (2) the proportion of the most common sound name out of all sound names applied to the same stimulus by different participants. This ensured that cluster centroids were close to the most representative sounds in each category.

We also performed affinity propagation clustering of the semantic distance matrix using *apcluster* R package (Bodenhofer et al. 2011). To find optimal clustering solutions, we varied the parameter q (sample quantile of the preference with which a data point becomes a centroid), which modulates the propensity of clustering algorithm for splitting or lumping. We then examined the quality of the resulting clustering solution by measuring (1) the average Silhouette Index, which is a measure of compactness and purity of clusters, and (2) the similarity of the clustering solution to the clusters defined by the most common name of each sound chosen by the participants (cf. Gamba et al. 2015).

Analysis of Acoustic Data All acoustic measurements were taken from the original acoustic analysis of the corpus as reported in Anikin and Persson (2017) and Anikin and Lima (2017). They included measures (median and standard deviation) of amplitude, fundamental frequency (pitch), distribution of energy in the spectrum, harmonics-to-noise ratio, proportion of voiced frames, and several temporal measures, such as the number, spacing and regularity of syllables. We aimed to define the acoustic space that would optimally preserve the structure of the semantic space of sound names. To do this, we chose a subset of acoustic variables and their weights iteratively, trying to maximize the correlation between the acoustic distance matrix (Euclidean distances between stimuli based on a weighted linear combination of acoustic predictors) and a reference distance matrix derived from the participants' judgments.

A subset of twelve acoustic predictors listed and explained in Table 1 proved optimal for maximizing the correlation with the semantic distance matrix (based on sound names in all three languages). In practice, the weights of acoustic variables did not have to be adjusted much to achieve optimal correlation with any of the other explored distance matrices (Table 2, first column): Cronbach's alpha for weights optimized for different targets was 0.95; 95% CI [0.91, 0.99]. We then employed two classification algorithms to predict the chosen sound names in each language. The more easily interpreted multinomial regression was trained on the first two principal components of the acoustic matrix in order to visualize the acoustic space in each language (Fig. 2), while the more powerful Random Forest classifier, which builds and cross-validates a large "forest" of decision trees (Breiman 2001), was trained on the 12 individual predictors to estimate the extent to which objective acoustic measurements were sufficient to predict the perceived acoustic type.

Relation Between Call Types and Emotions Contingency tables describing co-occurrence of sound names and emotion names were analyzed using Random Forest. This allowed us to estimate to what extent we could predict the perceived emotion knowing the chosen sound name(s) of a sound.

To compare the speed with which participants named the acoustic type and emotion of each stimulus, we recorded the delay between sound onset and (1) choosing the first sound name, (2) choosing the first emotion name, and (3) clicking the *Next* button to proceed to the next sound. Response times greater than 60 s were occasionally ($\sim$ 4% of trials) recorded among Russian participants, who took the test online without supervision.

Presumably, such long delays were related to technical problems or participants taking a break, and they were removed from the analysis of response times. Time measures were log-transformed due to a right skew in their distribution and analyzed using a Gaussian model with two random effects: sound and participant. This and other linear models were fit using Markov chain Monte Carlo in the Stan computational framework (Stan Development Team 2014).

Subjective certainty in the chosen answer was indicated separately for emotion name and sound name as, "*Don't know*", "*Unsure*", or "*Sure*". It was analyzed using ordinal logistic regression, again with two random effects. The consistency of participants' choices was operationalized as normalized entropy of all names chosen for a particular stimulus by all participants who had rated it, separately for sound names and for emotion names:

$$\text{entropy} = -\text{sum}(\log_2(\text{a}/\text{sum}(\text{a})) * \text{a}/\text{sum}(\text{a}))/\log_2(\text{number\_alternatives}) * 100\%,$$

where *a* was a vector of the same length as the number of alternative answers (number_alternatives, which was either 12 or 16) consisting of the number of times each sound name or emotion name was chosen. Because both the number of alternatives and the total number of responses per term varied, entropy was normalized to range from 0 to 100%. The distribution of entropy of 132 sounds was approximately normal, and therefore it was analyzed using Gaussian models.

The sounds, *R* scripts, raw data, additional tables and graphs can be accessed at http://cogsci.se/publications.html.

Results

Perceptually and Acoustically Distinct Call Types

In each language, we identified the most common name for each of 132 sounds and constructed a language-specific semantic space, in which the relative distance between any two stimuli depends on how often they were described with the same word. In all three languages, the first three principal components explained $> 80\%$ of variance in the resulting distance matrix, suggesting that three-dimensional solutions were adequate. Figure 1 (top panel) shows the semantic spaces of sound names for English, Swedish, and Russian. Each text label represents a single sound, labeled with its most commonly chosen name. The closer two sounds are in the graph, the more often they were given the same name by different participants. In addition, for each sound name the central location of stimuli with this name—cluster centroid—is shown in bold letters. For example, the centroids for the English words scream and shriek are close to each other, indicating that this distinction was not particularly consistent.

Based on visual inspection, semantic spaces of sound names are remarkably similar for all three languages: one dimension separates moan-like from scream-like sounds, while two more dimensions separate laughing and crying from all other sounds. More formally, the distance matrices for English, Swedish and Russian sound names are strongly correlated: $r > 0.8$ for all three pairs of languages (see Table 2).

As shown in the cladograms in the bottom panel of Fig. 1, in all three languages the most fundamental distinction was made between laughing, crying, and the remaining vocalizations. Beyond these three major groups, the order of separation between clusters was more language-specific. The languages also differed in the depth of classification: English appears to have the richest sound vocabulary with at least ten consistently labeled

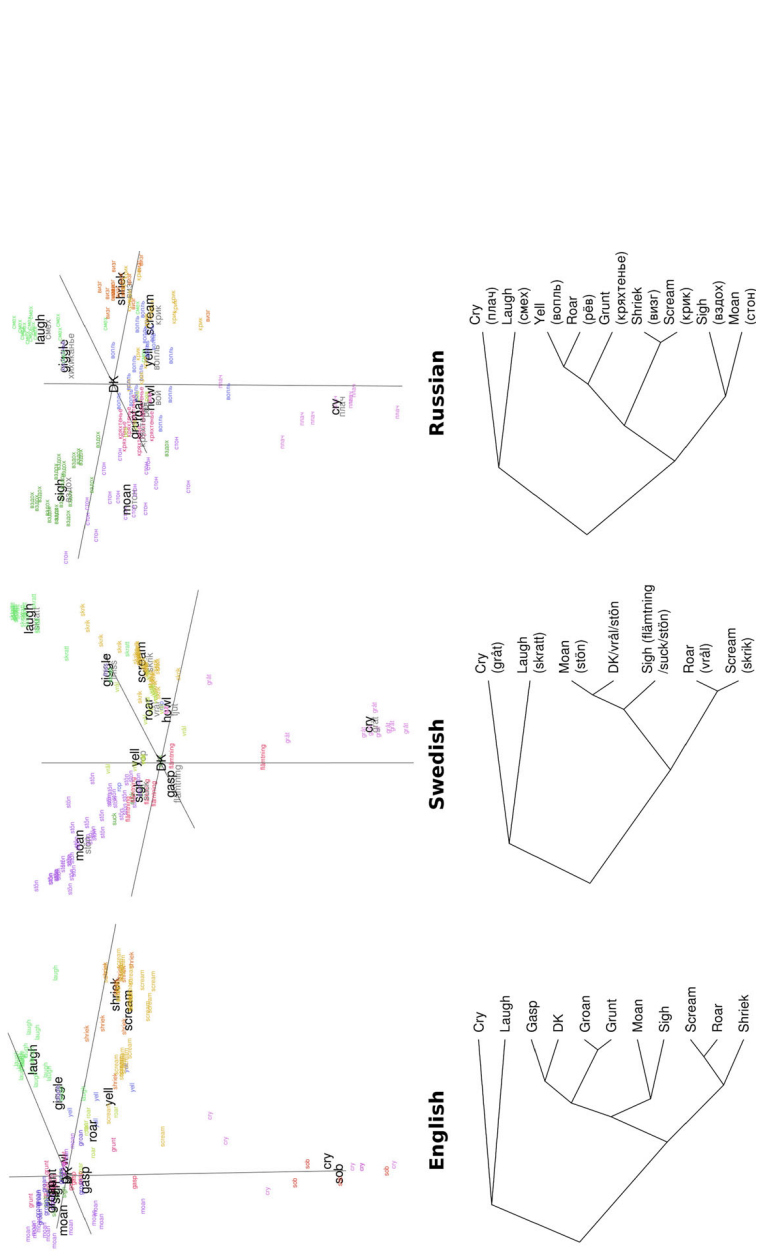

Fig. 1 Top panel: semantic space representing naming distinctions in English, Swedish, and Russian. Text labels are positioned in prototypicality-adjusted cluster centroids. Bottom panel: cladograms of the major call types. Affinity propagation clustering with q selected manually (0.5 for English, 0.3 for Swedish, 0.45 for Russian). *DK* don't know

Table 1 Variables used to construct the acoustic space and their weights optimized for maximum correlation between acoustic and semantic spaces

Variable	Interpretation	Weight	Loadings	
			PC1	PC2
Amplitude, median	Median root square amplitude (loudness)	0.61	0.13	−0.17
Proportion of voiced frames	How much of the sound is voiced	1.28	0.24	−0.41
Pitch, median	Fundamental frequency or perceived pitch (manually checked)	1.89	0.75	0.19
Pitch, SD		0.79	0.21	0.1
First quartile, median	First quartile of spectral energy distribution	1.42	0.53	0.04
First quartile, SD		0.74	0.15	0.16
Spectral entropy, SD	SD of the entropy of spectral energy distribution	0.9	0.03	0.16
Interburst interval, median	Time between vocal bursts (amplitude peaks)	0.58	0.01	−0.03
Interburst interval, SD		1.73	0.04	−0.07
Number of bursts	Total number of amplitude peaks per sound	1.61	−0.11	0.81
Syllable length, median	Length of continuous vocal segments	0.77	0	−0.16
Syllable length, SD		0.48	0.04	−0.08

acoustic types, compared to as few as six in Swedish and seven or eight in Russian. The exact number is hard to determine, since measures of clustering quality indicated several valid clustering solutions. The cladograms in Fig. 1 are merely one possible interpretation of the major call types based on the naming data in these three languages.

We also performed acoustic analysis to determine how subjective categorization of non-linguistic vocalizations related to objective acoustic differences between these sounds. Since the semantic spaces of sound names were so similar in English, Swedish, and Russian, we averaged the corresponding distance matrices from all three languages and used this averaged matrix to find an acoustic space of non-linguistic vocalizations that would represent, as faithfully as possible, the acoustic distinctions observed by speakers of these languages. As shown in Table 1, a subset of 12 weighted acoustic variables maximized the correlation between acoustic and semantic distance matrices ($r = 0.50$).

In other words, we asked the following question: What acoustic characteristics do we have to measure in order to separate the sounds into the same groups as did our participants when they named the sounds? The matrix of the chosen 12 (scaled and weighted) acoustic variables had only two strong principal components, which together explained 64% of variance. The first principal component correlated primarily with median pitch and the second with the number of vocal bursts (Table 1; Fig. 2). Based on the available acoustic measurements, it appears that participants distinguished between call types primarily based on their pitch, the number and irregularity of syllables, the balance between voiced and unvoiced parts, and some spectral characteristics.

It is also interesting to determine to what extent the classification of sounds into call types can be reproduced using objective acoustic measurements. Adjusted Rand Index demonstrates a much higher agreement of the actual naming with a clustering solution based on distances in the averaged semantic space (0.46, 0.48, and 0.49 for English,

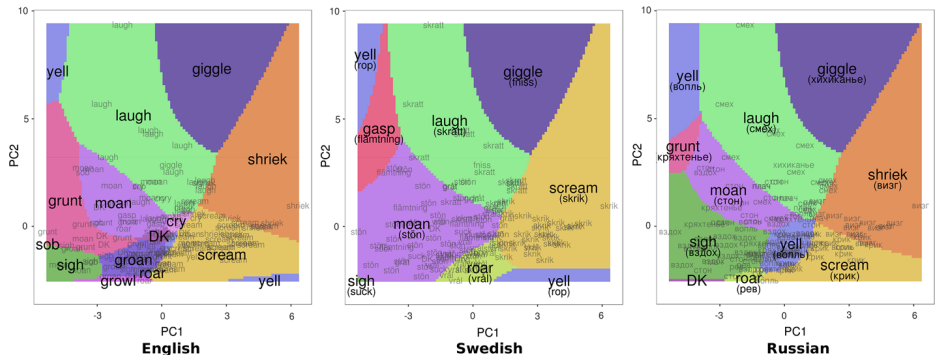

Fig. 2 Acoustic models for classifying vocalizations based on sound names chosen by English, Swedish, and Russian participants. Shaded areas show the acoustic class predicted by a multinomial regression model using two first principal components of 12 acoustic features (see Table 1). Small labels show the position of individual stimuli and their call type

Swedish, and Russian, respectively) than with a clustering solution based on distances in the acoustic space (0.14, 0.15, and 0.13). The reason is that acoustically the stimuli are highly graded. As can be seen in the scatterplot in Fig. 2, sounds form a single cloud, with no clear clusters and a lot of overlap between call types. This contrasts with the relatively well-separated clusters in Fig. 1. Using the 12 acoustic variables listed in Table 1, a Random Forest classifier correctly predicted the chosen sound name approximately 40% of the time in English, 62% in Swedish, and 54% in Russian. We also repeated Random Forest classification after pooling sound names into six major categories (laugh, cry, scream, moan, sigh, and roar) plus one residual unclassified "other" category. With these seven categories, classification accuracy was approximately 60% for all three languages.

Our findings thus indicate that participants classified sounds into call types more consistently than could be expected given the available acoustic measures. This result should be treated with some caution, however, since several call types were represented by only a few sounds (e.g., gasp, howl, etc.). The most common types, such as laughs and screams, also had high recognition rates in Random Forest models (75% and better), suggesting that classification accuracy by acoustic models might improve with a larger training sample.

How Do Call Types Map onto Emotion?

To explore the correspondence between naming the sound and naming the speaker's emotion, we analyzed contingency tables of sound and emotion names (Fig. 3). For example, the cell in the top left corner for English shows that a sound was simultaneously labeled *scream* and *anger* in 29 individual trials, whereas the combination of *scream* and *fear* was more common (191 trials). A Chi square test performed on this table proved that these acoustic-emotional classifications were not independent (English: $\chi^2 = 8568$, $df = 256$; Swedish: $\chi^2 = 7761$, $df = 192$; Russian: $\chi^2 = 7102$, $df = 192$; $p < 10^{-15}$ for all three). However, the association between naming the acoustic type of a vocalization (e.g., a scream) and naming its emotion (e.g., fear) was far from perfect. Based on the chosen sound name, a Random Forest classifier correctly predicted the chosen emotion name approximately 60% of the time in English, 50% in Swedish, and 60% in Russian. Knowing what speakers called a sound thus provided roughly half the information needed to predict its perceived emotion.

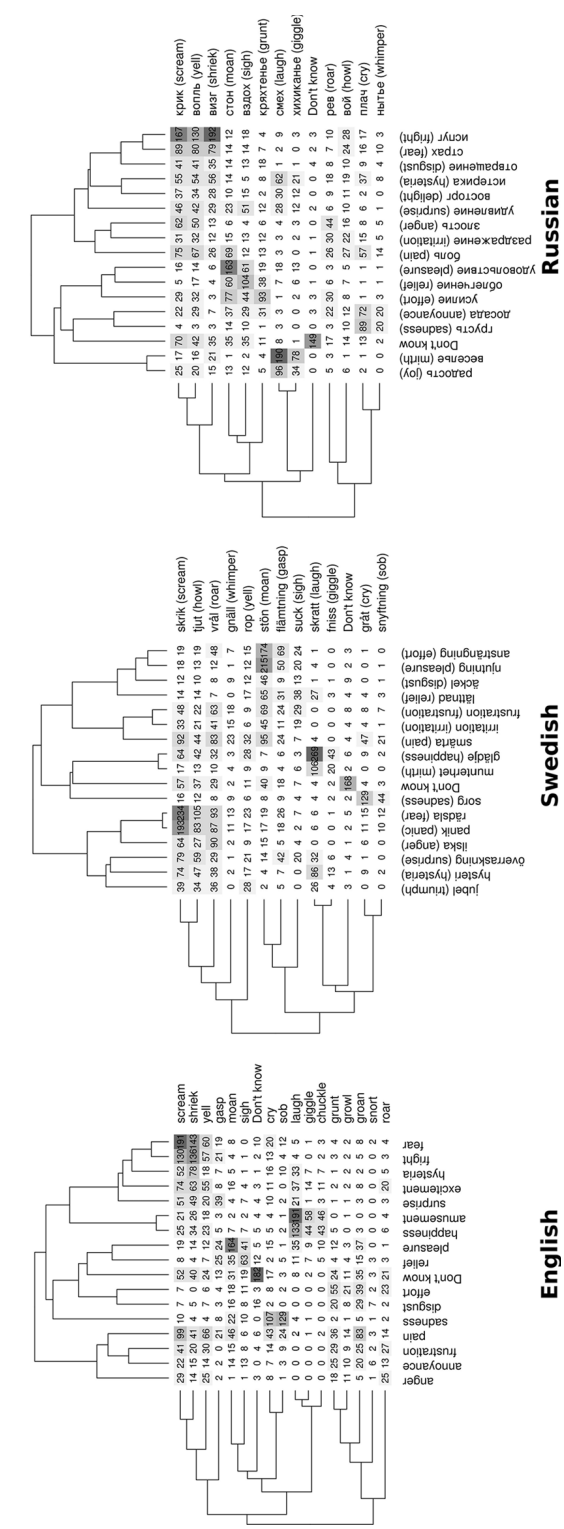

Fig. 3 Co-occurrence of sound names (vertical) and emotion names (horizontal) in English, Swedish, and Russian. Each cell shows the number of times a participant chose a particular sound name and emotion name to describe the same sound. The labels are ordered by hierarchical clustering, so that similar terms are placed close together, resulting in the dendrograms shown next to each table

Of course, perfect correspondence is less likely when there are more emotion names than sound names, as was the case in Swedish and Russian. However, the association between call type and emotion was similarly imperfect in English, which had 16 sound names and 16 emotion names. Furthermore, the lack of one-to-one mapping is not only due to the presence of close synonyms among the available verbal labels. For example, when a participant classified a sound as a scream, the perceived emotion varied widely and included quite distinct contexts, such as fear, pain, delight, surprise, etc. Moans, grunts, and sighs also varied considerably in their emotional interpretation. In contrast, laughing and crying were more closely associated by participants with a particular emotional state (amusement/joy and sadness, respectively; see Fig. 3).

Ease and Consistency of Naming the Sound Versus Naming the Emotion

Participants started by naming the emotion in $\sim$ 3% of trials when emotion names were on the right, but they started by naming the sound in $\sim$ 27% of trials when sound names were on the right: odds ratio = 28, 95% CI [7, 141]. If the relative position of sound names and emotion names on the screen was the only factor affecting the order of responses, the probability of answering left-to-right should have been the same regardless of whether sound names or emotion names were on the left. Instead, we observed a bias to name the sound before naming the emotion.

Median time needed to name both the sound and its emotion was 14 s, and median time needed to choose the first of these names was 5 s. Controlling for the order in which the two blocks were presented on the screen, it took 850 ms [740, 960] longer to choose the first emotion name versus the first sound name. This observation confirms that participants preferred to name the sound before naming its emotion.

Subjective certainty in the given answers, measured on a scale of 1–3 (*Don't know*—*Unsure*—*Sure*), was on average 2.76 [2.75, 2.78] for sound names and 2.54 [2.52, 2.56] for emotion names. The proportion of "*Sure*" ratings was higher for sound names, while the proportions of "*Don't know*" and "*Unsure*" ratings were higher for emotion names (Fig. 4). The results were similar for all three languages (not shown). Furthermore, there were 7.4% of trials in which participants named the sound but not the emotion, whereas the reverse pattern of naming the emotion but not the sound occurred in only 1.7% of trials: odds ratio = 5.2 [4.2, 6.5]. Participants thus named the sound with more certainty than they named the emotion.

Normalized entropy was considerably lower for sound names than for emotion names (49 vs. 59%, difference = 9.6% [7.7, 11.6], Cohen's $d = 0.76$). Low entropy means that participants mostly agreed on what to call a particular stimulus, whereas high entropy means that different participants chose many different terms for the same stimulus. Our

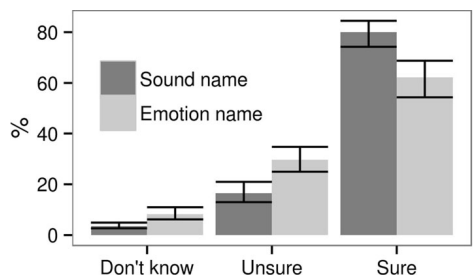

Fig. 4 The probability of expressing different levels of certainty in the chosen sound names and emotion names for all three language groups combined. Median of the posterior distribution and 95% CI

results thus suggest that sound names were chosen more consistently than emotion names. In Swedish and Russian, this could be due to the smaller number of available alternatives: 12 sound names versus 16 emotion names. However, we corrected for the number of alternatives and used normalized entropy. Moreover, in English there were equal numbers of sound names and emotion names (16 of each), but the entropy of sound names was still 7.3% [5.0, 9.5%] lower.

We expected that it would be easier to name the call type than to name the emotion only for those sounds that were the most distinct acoustically (e.g., laughs), while for other sounds it would be easier to name the emotion than to name the call type. However, the average certainty in the given sound names was higher than the average certainty in the given emotion names for all 40 (12 + 12 + 16) sound names in the three languages and for all but one emotion (disgust). Sound names were also chosen faster than emotion names for all call types in all languages except *flämtning* (gasp) in Swedish, *snort* in English, and *рёв* (roar) in Russian. In addition, there was a strong positive correlation between the speed of naming each sound and naming its emotion: $r = 0.75$, 95% CI [0.62, 0.85]. Similarly, the certainty in the choice of sound name (averaged per sound) correlated with the certainty in the choice of emotion name: $r = 0.75$, 95% CI [0.63, 0.86]. There was also a positive correlation between the entropy of sound names and emotion names: $r = 0.59$ (95% CI [0.44, 0.72]). The speed, certainty and consistency of naming a particular sound were thus strongly correlated with the speed, certainty and consistency of naming the emotion that it expressed.

To summarize, English-, Swedish-, and Russian-speaking participants in Experiment 1 demonstrated a high level of agreement when classifying non-linguistic vocalizations into approximately six major call types, which could also be defined in terms of objectively measured acoustic features. More fine-grained classification into acoustic subtypes was generally less consistent both across and within languages. The classifications of a sound in terms of its acoustic type and emotion were neither totally independent nor redundant: some call types were strongly associated with a single emotion, while others were perceived to express a variety of states. It seemed more natural to name the sound before naming its emotion, apparently for most call types and emotions. However, these two processes were not independent: sounds that were easy to classify acoustically were also easy to interpret in terms of the caller's emotion, while acoustically unnameable sounds remained emotionally opaque.

Experiment 2

Cross-linguistic naming studies, such as the one above, have their limitations. One problem is that the availability of verbal labels in a language is not a prerequisite for distinguishing categories of stimuli. For example, Yucatec Maya does not possess two separate words for disgust and anger, but there is evidence that speakers still perceive the corresponding facial expressions as two distinct categories (Sauter et al. 2011). Often modifiers allow speakers to make subtle distinctions despite a paucity of basic lexemes (Malt et al. 2010). In other cases the abundance of language-specific lexical distinctions may exaggerate the apparent complexity and culture-specificity of a cognitive domain and obfuscate its underlying universality. For example, similarities between household utensils based on direct non-linguistic comparisons are more consistent across languages compared to similarities derived from verbal labeling of such objects (Ameel et al. 2005; Malt et al. 1999).

In other words, the presence or absence of a linguistic distinction in several languages can be suggestive, but in itself it can neither prove nor falsify the universality of the corresponding conceptual distinction. It is therefore desirable to obtain language-independent evidence. Moreover, in Experiment 1 participants were forced to choose among 12 or 16 pre-given labels, further restricting the possible patterns of classification. Given these limitations, we also tested the same 132 sounds in another experiment, avoiding verbal labels altogether and aiming to obtain an estimate of "naked" perceived similarity between stimuli.

To do this, we used the triad classification task, which is an established tool for studying the categorization of multidimensional stimuli (Alvarado 1996; Raijmakers et al. 2004). Participants in a triad task are presented with three stimuli at a time and select two that are the most similar. These decisions can be used to estimate the perceived "distances" between stimuli. Since in Experiment 1 we discovered that call types were highly salient to listeners, we hypothesized that this distance matrix would be more compatible with the distance matrix calculated in Experiment 1 based on the chosen sound names, rather than with the distance matrix based on emotion names. Participants' choices in a triad task depend on the instructions: they have to be told on what basis they are supposed to compare the stimuli in each triad. We loaded the dice against the hypothesis and specifically asked participants to choose based on the similarity of underlying emotional states, not the similarity of acoustic characteristics.

The triad task has previously been applied to the classification of emotional vocalizations: Green and Cliff (1975) tested 11 sounds, one for each emotion. However, it is impossible to discover an alternative clustering with so few stimuli. Besides, Green and Cliff worked with artificial and speech-like material (recited letters of the alphabet) rather than natural vocalizations. Our study is thus the first to use the triad task for label-free classification of human vocalizations.

Materials and Methods

Stimuli

We used the same 132 sounds as in Experiment 1.

Participants

Participants in the triad task were recruited on the campus of Lund University or online, through advertisements and personal contacts. All participants who performed at least ten out of forty-two trials were included in the analysis ($N = 241$). The experiment was available in three languages: Swedish ($n = 156$ participants), English ($n = 77$) and Russian ($n = 8$). Since the Russian sample was too small to construct a distance matrix, we only present the results for the Swedish and English samples.

Procedure

The experiment was written in html/javascript and made available online. Participants performed the test in common rooms at the university or at home. It took approximately 10–15 min to complete the entire test (132 sounds in 42 triads), although incomplete tests

were also accepted. All data was completely anonymous and the online test could be interrupted at any time.

A standard version of the triad classification task (Raijmakers et al. 2004) was used. Participants were presented with three sounds at a time, and they could replay each sound as many times as they needed before indicating which two sounds in the triad were emotionally more similar. Just like Nijmegen method of free-text labeling in the pilot version of Experiment 1, categorization in the triad task is inductive, in the sense that the nature of categories is not predetermined and their number is not limited. The instructions, visible throughout the experiment, specifically asked to choose based on the emotional state of the caller. Each of 132 sounds was presented once in random order.

Statistical Analysis

The output of the triad task was analyzed using a Bayesian model. The model assumes that each sound is embedded in a d-dimensional space and that for every triad the participant's choice is a function of the relative distances between the three sounds. The pair of sounds with the smallest distance is the one most likely to be chosen by the participant. To find the posterior distribution of the embedding in d-dimensional space, the model was fit using Markov chain Monte Carlo in the Stan computational framework (Stan Development Team 2014).

Since dimensionality was hard-coded in the generative model, we explored models with different numbers of dimensions and estimated how well each described the actual responses of participants. Watanabe-Akaike Information Criterion (WAIC), which is an approximation to leave-one-out cross-validation, was used as a measure of overall fit (Watanabe 2010). In addition, we calculated the correlation between the distance matrices based on linguistic labeling in Experiment 1 (either sound names or emotion names, averaged across three languages) and the distance matrix in Experiment 2 estimated by a generative model with d dimensions, separately for Swedish and English (Fig. 5).

Results and Discussion

The first step was to determine how many dimensions were necessary to represent the configuration of stimuli corresponding to the distinctions made by participants in the triad task. A three-dimensional model achieved optimal correlation with the distance matrices from Experiment 1 based on naming both the sound and its emotion for both English and Swedish data (Fig. 5). WAIC suggested that three dimensions were optimal for Swedish and two or three for English; we therefore focused on three-dimensional models.

For both English and Swedish, the distance matrix from the triad task was more similar to the distance matrix from Experiment 1 based on sound names ($r = 0.69$ for English and 0.73 for Swedish data) than to the distance matrix from Experiment 1 based on emotion names ($r = 0.50$ and 0.57, respectively; Table 2). A visual inspection of the configuration of stimuli that best represented similarity judgments made by participants in the triad task (Fig. 6) confirmed that this configuration was qualitatively similar to the semantic space of sound names in Fig. 1. Once again, laughs and cries formed clearly separated clusters, while the remaining sounds were spread out in a cloud from sighs and moans to screams. The main difference between this configuration and semantic spaces in Experiment 1 was that the clusters were less compact in the triad task. The reason may be that participants in Experiment 1 had to choose among a few available verbal labels, whereas similarity judgments in the triad task were unrestrained, allowing more subtle distinctions.

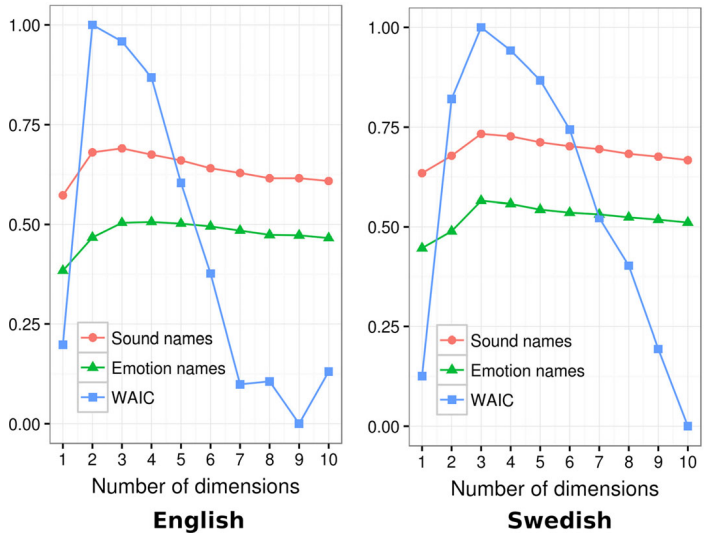

Fig. 5 Model fit as a function of its dimensionality for the triad classification task. Shown: Pearson's correlation with distance matrices from Experiment 1 based on naming the sound or emotion and normalized negative WAIC (larger is better)

Since so few Russian-speaking participants took part in the triad classification task, no comparison could be made for this language. Moreover, the English group in the online-based triad task is not guaranteed to consist entirely of native speakers (in contrast to Experiment 1, where participants were tested in person and were native speakers), potentially limiting the compatibility of English data from the two studies. Despite these limitations, Experiment 2 has demonstrated that, overriding the explicit instructions to choose based on emotion, participants made similarity judgments that were more compatible with verbal classification of stimuli into acoustic categories than into emotional categories. This convergent evidence highlights the perceptual salience of acoustic categories and confirms that linguistic labeling in Experiment 1 provided valid information about the underlying cognitive representation of non-linguistic vocalizations.

General Discussion

To investigate the relation between acoustic categories (e.g., laughter or moan) and perceived emotion, we analyzed acoustically 132 sounds from a corpus of authentic non-linguistic vocalizations (Anikin and Persson 2017) and compared their verbal classification by acoustic category and emotion by native speakers of English, Swedish, and Russian (Experiment 1). We found strong parallels across all three languages in the distinguished acoustic categories, which was further confirmed using a nonverbal classification test (Experiment 2). In line with acoustic research in other primates (Fischer et al. 2016), human vocalizations appear to be highly graded, and all sounds apart from laughing and crying can be roughly aligned along a single dimension, which acoustically corresponds to pitch. Based on the analyzed sample of sounds and languages, the conceptual space of non-linguistic vocalizations thus appears to be three-dimensional, and the most salient acoustic categories are: laughing, crying, screaming, and moaning. We suggest that these categories

Table 2 Pearson's correlations between the distance matrices in Experiments 1 and 2

	Acoustic analysis	Sound names EN	Sound names SV	Sound names RU	Sound names EN + SV + RU	Emotion names EN	Emotion names SV	Emotion names RU	Emotion names EN + SV + RU	Triad classif. EN	Triad classif. SV
Sound names EN	0.47										
Sound names SV	0.46	0.85									
Sound names RU	0.48	0.8	0.8								
Sound names EN + SV + RU	**0.5**	0.94	0.95	0.92							
Emotion names EN	0.34	0.68	0.65	0.62	0.69						
Emotion names SV	0.33	0.65	0.64	0.58	0.67	0.87					
Emotion names RU	0.43	0.69	0.68	0.73	0.74	0.8	0.77				
Emotion names EN + SV + RU	**0.38**	0.72	0.7	0.68	**0.75**	0.95	0.95	0.91			
Triad classif. EN	0.5	0.62	0.68	0.63	0.69	0.45	0.46	0.52	0.5		
Triad classif. SV	0.48	0.67	0.71	0.67	0.73	0.52	0.52	0.55	0.57	0.75	
Triad classif. EN + SV	**0.52**	0.69	0.74	0.69	**0.75**	0.51	0.51	0.57	**0.56**	0.96	0.91

Bold values indicate the most relevant information

Acoustic analysis = distance matrix based on 12 acoustic variables with weights optimized for maximum correlation with each target distance matrix (see Figure S2), thus producing the highest achievable correlation

Sound names EN/SV/RU: distance matrix based on English/Swedish/Russian sound names

Sound names EN + SV + RU: averaged distance matrix for sound names in all three languages

Emotion names EN/SV/RU: distance matrix based on English/Swedish/Russian emotion names

Emotion names EN + SV + RU: averaged distance matrix for emotion names in all three languages

Triad classif. EN/SV: distance matrix based on the triad classification task in the English/Swedish sample

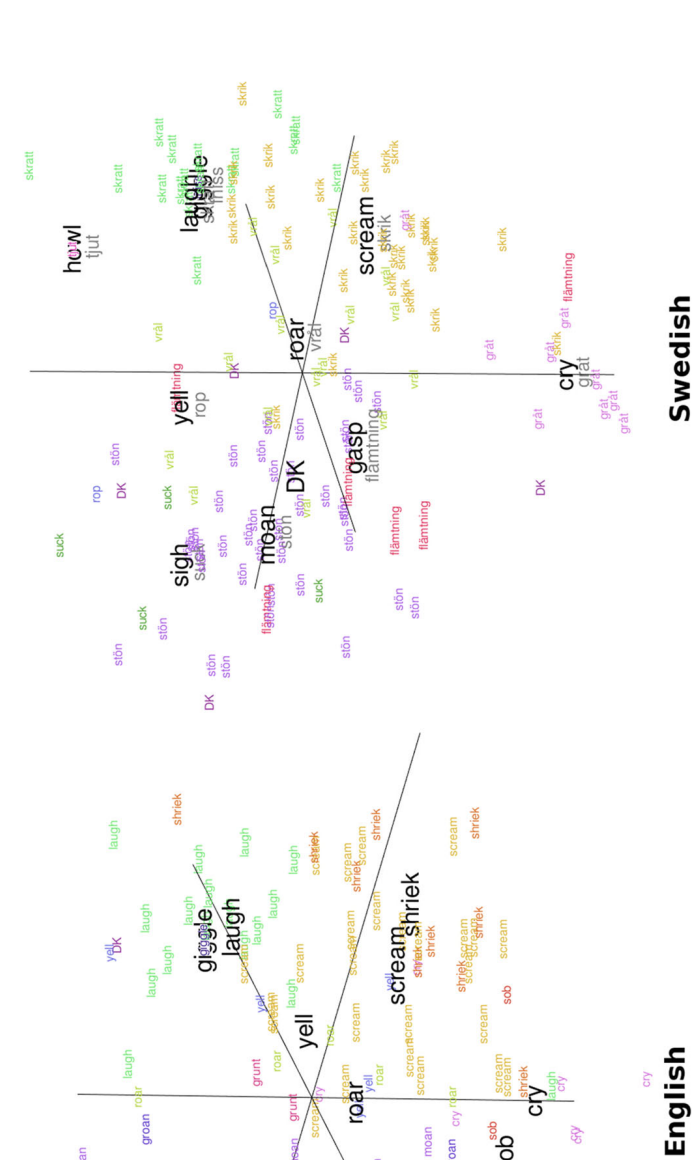

Fig. 6 Three-dimensional configuration of stimuli in the triad task. Each sound is labeled and color-coded by its most common name in Experiment 1 (cf. Fig. 1). Prototypicality-adjusted centroids are shown in large bold font. Origin lies in the center of gravity of each cloud. *DK* don't know

may correspond to species-typical call types—innate vocalizations that are produced and recognized in all cultures. Roaring and sighing are two more candidate call types, but the evidence in their case is less conclusive.

This list of perceptually distinct vocalizations can only be regarded as preliminary, since it critically depends on the range of tested sounds. For example, there were no completely voiceless sounds in the corpus, likely influencing the apparent semantic extension of the word *sigh*. Moreover, we only examined three languages from the Indo-European family, and the variation in sound-related vocabulary may become larger if more distantly related languages are compared. However, even with only three languages, it is already clear that the number of "basic" (i.e., perceptually and acoustically distinct) call types is considerably smaller than the number of available lexemes for acoustic categories in the general vocabulary.

To understand why this is so, it is helpful to distinguish between extension and connotation of sound names. For example, the words for breathy sounds in English, Swedish, and Russian appear to differ primarily in their extensions: only English has a consistent ingressive–egressive distinction at the level of basic lexemes (*sigh* versus *gasp*), the Russian *вздох* (sigh) apparently allows for relatively more voicing, etc. There are also sound names that differ primarily in their connotations, such as *groan*/*moan* in English or *rop*/*vrål* in Swedish. These words were often applied to the same sound by different participants, depending on which emotion they perceived. They are therefore not fully synonymous, but it may still be unwarranted to claim that they represent two different vocalizations, since these semantic distinctions are neither acoustically robust nor consistent across languages. Finally, words like *laughing*/*giggling*/*chuckling*, *crying*/*sobbing*, and *screaming*/*shrieking*/*yelling*, as well as the equivalent terms in Swedish and Russian, are close synonyms that overlap in both extension and connotation in all three languages. In such cases, the likely interpretation is that these words refer to subtypes of what is perceptually a single vocalization.

As a result, we are left with only a handful of cross-culturally recognized and acoustically definable call types, perhaps as few as four to six. This relatively small number may come as a surprise, considering the larger number of lexemes for non-linguistic vocalizations and of emotions that can be correctly detected based on vocal cues (9–16 emotions in Cordaro et al. 2016; 14 emotions in Simon-Thomas et al. 2009). The number of call types we identified also falls far short of that ascribed to the great apes. Estimates vary, but gorillas may have about 16 call types (Fossey 1972), bonobos 12–19 (Bermejo and Omedes 2000; de Waal 1988), chimpanzees 13–24 (Goodall 1986; Marler 1976), and orangutans 32 (Hardus et al. 2009). An intriguing possibility is that these estimates of the size of vocal repertoire in apes are inflated, because within-call acoustic variation is easily mistaken for distinct call types. For example, by simply varying the amount of nonlinearities such as subharmonics and deterministic chaos, a nearly tonal vocalization can be made bark-like and almost unrecognizable as an instance of the same call (Fitch et al. 2002). Once the production mechanism of each call is better understood, some ape vocalizations may thus be reclassified as variations of the same basic type.

The relatively small number of identified human call types does not contradict the well-established fact that a rich variety of affective states can be recognized from non-linguistic vocalizations. Even a few distinct vocalizations may still be sufficient for expressing a wide range of meanings, provided that within-type acoustic variation is meaningful (Scheiner et al. 2002; Wadewitz et al. 2015). For instance, the exact manner of laughing may tell the listener as much as the fact that this is a laugh rather than, say, a grunt. Consistent with this explanation, the distinction between tonal and noisy sounds did not

appear to contribute to the categorization of sounds by call type in this study, whereas this acoustic parameter is of major importance for the categorization of the same sounds by emotion (Anikin and Persson 2017). Relatively tonal and noisy vocalizations of the same basic acoustic type may thus be associated with different emotions. The expressive range of each call type may be further enhanced by contextual information and integration of sound with input from other sensory modalities. For example, visible tears make crying less ambiguous and enhance the impression of sadness (Provine 2012; Provine et al. 2009).

It is also quite possible that humans possess more call types than we have identified, but these vocalizations lack monolexemic labels, at least in the investigated Indo-European languages. These call types may also fail to be consistently distinguished by participants and acoustic models, perhaps because the boundaries between them are blurred. In fact, our acoustic analysis revealed that most call types were highly graded, complicating their clustering based on the extracted acoustic features and limiting the accuracy with which acoustic models could predict the sound name in each language. A possible objection is that the acoustic characteristics we measured do not describe the sounds comprehensively. However, even fewer acoustic variables sufficed for machine learning algorithms to achieve accuracy on a par with human raters when classifying the original corpus by emotion (Anikin and Persson 2017). A more serious limitation of the current research is that our sample of 132 sounds may not be large enough or comprehensive enough to be considered representative of the range of non-linguistic vocalizations people produce. Our analysis needs to be extended, using a larger and more diverse collection of vocalizations, ideally recorded from an even broader range of contexts and from several cultural groups.

The interpretation we favor is that humans do possess species-typical vocalizations, but these are graded and further masked by the great variety of culturally learned non-linguistic vocalizations. Only the most salient and involuntary vocalizations remain universal and distinct enough to be perceived categorically in all cultures, with laughter being the paradigmatic example. We did not test for categorical perception per se, but the compact clustering of laughing and crying in the naming task, and particularly in the triad classification task, strongly suggests that at least these two vocalizations are perceived as qualitatively different from all other sounds, which is in line with previous studies of these two vocalizations (Lingle et al. 2012; Provine 2012). The fact that the separation between acoustic types made by participants was more consistent than might be expected based on acoustic measurements also implies their categorical perception, which can be verified in future studies. Ultimately, it would be also be illuminating to analyze the neurological and physiological processes involved in the production of each call type putatively identified in perceptual studies. This would both verify the validity of suggested acoustic categories and determine whether their universality is due to innateness or some other processes driving cross-cultural convergence.

In addition to identifying the major call types and their meaning, it is important to specify a cognitive model of the relation between sound and emotion classification by the listener. As a step in this direction, we compared the two tasks—naming the sound and naming its emotion—in terms of decision time, preferred order, subjective certainty, and consistency. Naming a sound acoustically (as a laugh, a scream, etc.) was associated with faster responses, greater certainty and higher consistency across participants compared to naming its emotion (Experiment 1). Intriguingly, this was the case for practically all analyzed vocalizations, not only for some particular classes. Furthermore, asked to compare the sounds based on the underlying affective state of the caller, participants still appeared to think largely in terms of acoustic categories (Experiment 2). At the same time, there was a close relation between the ease of acoustic and emotional interpretations. If a

sound could not be named, its emotion could not be determined, and vice versa: sounds that were easy to name were also more easily and consistently interpreted in terms of the underlying emotion.

A parsimonious explanation for these observations is that every vocalization is initially categorized acoustically and then interpreted in terms of the caller's emotion or intention, in accordance with the identification-attribution model (Spunt and Lieberman 2012). This task is streamlined when the vocalization belongs to a common and acoustically well-defined category, such as laughing or screaming. This would explain the strong correlation between the ease of naming the sound and the ease of naming its emotion: the identification of a particular call type carries useful information for the receiver, since each call type is associated with only a restricted range of emotions. Nevertheless, the association between call type and emotion is not redundant; instead, it turns out to be considerably more complex than might have been expected.

This calls for a complementary approach to the study of non-linguistic vocalizations—one mindful of acoustic types as such, rather than solely their potential for expressing emotion. We hope that this approach may provide a more comprehensive and phylogenetically informed account of vocal behavior, shedding new light on human nonverbal communication and bringing it more in line with research on vocal communication in other animals.

Acknowledgements We would like to thank Can Kabadayi, Judith Hall, and two anonymous reviewers for their useful comments on the manuscript. We are also grateful to the many participants who volunteered their time to rate the sounds.

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*(06), 529–546.

Alvarado, N. (1996). Congruence of meaning between facial expressions of emotion and selected emotion terms. *Motivation and Emotion, 20*(1), 33–61.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language, 53*(1), 60–80.

Anikin, A., & Lima, C. F. (2017). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *The Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2016.1270976.

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

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology, 25*(15), 2051–2056.

August, P. V., & Anderson, J. G. (1987). Mammal sounds and motivation-structural rules: A test of the hypothesis. *Journal of Mammalogy, 68*(1), 1–9.

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

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.

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.
Bermejo, M., & Omedes, A. (2000). Preliminary vocal repertoire and vocal communication of wild bonobos (*Pan paniscus*) at Lilungu (Democratic Republic of Congo). *Folia Primatologica, 70*(6), 328–357.
Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. *Bioinformatics, 27,* 2463–2464.
Boster, J. S. (2005). Emotion categories across languages. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 187–222). Oxford: Elsevier.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology, 288*(1), 1–20.
Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior, 35*(4), 327–335.
Bryant, G. A., Fessler, D. M., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., et al. (2016). Detecting affiliation in colaughter across 24 societies. *PNAS, 113*(17), 4682–4687.
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.
de Waal, F. B. (1988). The communicative repertoire of captive bonobos (*Pan paniscus*), compared to that of chimpanzees. *Behaviour, 106*(3), 183–251.
Devitt, M., & Sterelny, K. (1999). *Language and reality: An introduction to the philosophy of language* (2nd ed.). Oxford: Blackwell.
Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200.
Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, 164*(3875), 86–88.
Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203–235.
Enfield, N. J., Majid, A., & Van Staden, M. (2006). Cross-linguistic categorisation of the body: Introduction. *Language Sciences, 28*(2), 137–147.
Fischer, J., Wadewitz, P., & Hammerschmidt, K. (2016). Structural variability and communicative complexity in acoustic communication. *Animal Behaviour*. doi:10.1016/j.anbehav.2016.06.012.
Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour, 63*(3), 407–418.
Fossey, D. (1972). Vocalizations of the mountain gorilla (*Gorilla gorilla beringei*). *Animal Behaviour, 20*(1), 36–53.
Gamba, M., Friard, O., Riondato, I., Righini, R., Colombo, C., Miaretsoa, L., et al. (2015). Comparative analysis of the vocal repertoire of Eulemur: A dynamic time warping approach. *International Journal of Primatology, 36*(5), 894–910.
Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014a). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*(4), 911–920.
Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014b). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion, 14*(2), 251.
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.
Goodall, J. (1986). *The chimpanzees of Gombe: Patterns of behavior*. London: Harward University Press.
Green, R. S., & Cliff, N. (1975). Multidimensional comparisons of structures of vocally and facially expressed emotion. *Perception and Psychophysics, 17*(5), 429–438.
Hardus, M. E., Lameira, A. R., Singleton, I., Morrogh-Bernard, H. C., Knott, C. D., Ancrenaz, M., et al. (2009). A description of the orangutan's vocal and sound repertoire, with a focus on geographic variation. In S. A. Wich (Ed.), *Orangutans: Geographic variation in behavioral ecology and conservation*. New York: Oxford University Press.
Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*(2), 288–299.
Jürgens, U. (2009). The neural control of vocalization in mammals: A review. *Journal of Voice, 23*(1), 1–10.
Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., et al. (2014). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews, 91*(1), 13–52.
Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Frontiers in Psychology, 4,* 105. doi:10.3389/fpsyg.2013.00105.

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., et al. (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*. doi:10.1007/s10919-015-0222-8.

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.

Lindsey, D. T., & Brown, A. M. (2009). World Color Survey color naming reveals universal motifs and their within-language diversity. *PNAS, 106*(47), 19785–19790.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology, 58*(5), 698–726.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109*(2), 235–250.

Malt, B. C., Gennari, S. P., & Imai, M. (2010). Lexicalization patterns and the world-to-word mapping. In B. C. Malt & P. Wolf (Eds.), *Words and the mind: How words capture human experience*. New York: Oxford University Press.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40*(2), 230–262.

Marler, P. (1976). Social organization, communication and graded signals: The chimpanzee and the gorilla. In P. Bateson & R. Hinde (Eds.), *Growing points in ethology*. Oxford: Cambridge University Press.

Newman, J. D. (2007). Neural circuits underlying crying and cry responding in mammals. *Behavioural Brain Research, 182*(2), 155–165.

Oller, D. K., & Griebel, U. (Eds.). (2008). *Evolution of communicative flexibility: Complexity, creativity, and adaptability in human and animal communication*. Cambridge, MA: MIT Press.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology, 73*(6), 530–544.

Panksepp, J. (2007). Neuroevolutionary sources of laughter and social joy: Modeling primal human laughter in laboratory rats. *Behavioural Brain Research, 182*(2), 231–244.

Provine, R. R. (2001). *Laughter: A scientific investigation*. New York: Penguin books.

Provine, R. R. (2012). *Curious behavior: Yawning, laughing, hiccupping, and beyond*. Cambridge: Harvard University Press.

Provine, R. R., Krosnowski, K. A., & Brocato, N. W. (2009). Tearing: Breakthrough in human emotional signaling. *Evolutionary Psychology, 7*(1), 52–56.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Raijmakers, M. E., Jansen, B. R., & van der Maas, H. L. (2004). Rules and development in triad classification task performance. *Developmental Review, 24*(3), 289–321.

Ross, M. D., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology, 19*(13), 1106–1111.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *PNAS, 107*(6), 2408–2412.

Sauter, D. A., LeGuen, O., & Haun, D. (2011). Categorical perception of emotional facial expressions does not require lexical categories. *Emotion, 11*(6), 1479.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2002). Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice, 16*(4), 509–529.

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

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

Scheumann, M., Hasting, A. S., Kotz, S. A., & Zimmermann, E. (2014). The voice of emotion across species: How do human listeners recognize animals' affective states? *PLoS ONE, 9*(3), e91192.

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.

Slobin, D. I., Ibarretxe-Antuñano, I., Kopecka, A., & Majid, A. (2014). Manners of human gait: A crosslinguistic event-naming study. *Cognitive Linguistics, 25*(4), 701–741.

Sperduti, M., Guionnet, S., Fossati, P., & Nadel, J. (2014). Mirror neuron system and mentalizing system connect during online social interaction. *Cognitive Processing, 15*(3), 307–316.

Spunt, R. P., & Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *Neuroimage, 59*(3), 3050–3059.

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

Van Hooff, J. A. R. A. M., & Preuschoft, S. (2003). Laughter and smiling: The intertwining of nature and culture. In F. B. M. de Waal & P. L. Tyack (Eds.), *Animal social complexity: Intelligence, culture, and individualized societies*. Cambridge, MA: Harvard University Press.

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing vocal repertoires—Hard vs. soft classification approaches. *PLoS ONE, 10*(4), e0125785. doi:10.1371/journal.pone.0125785.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross-validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research, 11,* 3571–3594.

Watson, S. K., Townsend, S. W., Schel, A. M., Wilke, C., Wallace, E. K., Cheng, L., et al. (2015). Vocal learning in the functionally referential food grunts of chimpanzees. *Current Biology, 25*(4), 495–499.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews, 21*(5), 195–205.

Zwarts, J. (2010). Semantic map geometry: Two approaches. *Linguistic Discovery, 8*(1), 377–395.

Paper IV

Behavior Research Methods (2019) 51:778–792
https://doi.org/10.3758/s13428-018-1095-7

Soundgen: An open-source tool for synthesizing nonverbal vocalizations

Andrey Anikin [1]

Published online: 27 July 2018

Abstract
Voice synthesis is a useful method for investigating the communicative role of different acoustic features. Although many text-to-speech systems are available, researchers of human nonverbal vocalizations and bioacousticians may profit from a dedicated simple tool for synthesizing and manipulating natural-sounding vocalizations. *Soundgen* (https://CRAN.R-project.org/package=soundgen) is an open-source R package that synthesizes nonverbal vocalizations based on meaningful acoustic parameters, which can be specified from the command line or in an interactive app. This tool was validated by comparing the perceived emotion, valence, arousal, and authenticity of 60 recorded human nonverbal vocalizations (screams, moans, laughs, and so on) and their approximate synthetic reproductions. Each synthetic sound was created by manually specifying only a small number of high-level control parameters, such as syllable length and a few anchors for the intonation contour. Nevertheless, the valence and arousal ratings of synthetic sounds were similar to those of the original recordings, and the authenticity ratings were comparable, maintaining parity with the originals for less complex vocalizations. Manipulating the precise acoustic characteristics of synthetic sounds may shed light on the salient predictors of emotion in the human voice. More generally, *soundgen* may prove useful for any studies that require precise control over the acoustic features of nonspeech sounds, including research on animal vocalizations and auditory perception.

Keywords Nonverbal vocalizations · Animal vocalizations · Formant synthesis · Parametric synthesis · Voice synthesis · Open source · Emotion

An important goal for research on acoustic communication is to determine how the particular characteristics of a produced sound affect its meaning. For example, acoustic correlates of different affective states can be identified by comparing recordings that were obtained in different contexts (Briefer, 2012; Hammerschmidt & Jürgens, 2007) or that are perceived as expressing different emotions (Banse & Scherer, 1996; Sauter, Eisner, Calder, & Scott, 2010). Correlation is not causation, however: To determine which acoustic features actually influence the perceivers, methodologically the most powerful approach is to modify the signal, one feature at a time (Scherer, 2003). Speech synthesis is a diverse and mature field (Schröder, 2009), but fewer options are available to researchers who wish to synthesize or modify human nonverbal vocalizations, such as laughs and screams, or sounds produced by nonhuman animals. For instance, it would be easier to elucidate the contested role of nonlinear phenomena in pant-hoots of chimpanzees (Riede, Arcadi, & Owren, 2007) or to determine what acoustic characteristics help listeners discriminate between spontaneous and volitional laughs (Anikin & Lima, 2018; Bryant & Aktipis, 2014) if there were a simple way to synthesize these sounds and then manipulate their acoustic properties. This is the context in which *soundgen* (https://CRAN.R-project.org/package=soundgen) was developed as an open-source tool designed specifically for the manual, fully controlled synthesis and manipulation of nonverbal vocalizations.

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

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

What is soundgen?

Soundgen is an open-source library that contains tools for analyzing, manipulating, and synthesizing sounds. Its main function for sound synthesis, *soundgen()*, can generate one or more syllables with voiced and unvoiced segments. The control parameters refer to acoustically transparent and

perceptually meaningful characteristics such as amplitude envelope, intonation, and various aspects of voice quality. The input code is sparse, so that an entire vocalization or even multiple syllables can be created with a single short command. For example, the intonation of the entire vocalization can be specified with a few values of the fundamental frequency (f_0): one at 0 ms, another at 300 ms, and a third at 1,000 ms, producing a smooth f_0 contour that passes through these anchor points. Under the hood, soundgen creates a combined harmonic-noise excitation source (Erro, Sainz, Navas, & Hernaez, 2014; Gobl & Ní Chasaide, 2010; Stylianou, 2001) and then filters it to imitate the effects of the vocal tract, enhancing certain frequencies in the spectrum (formants).

Soundgen can be installed, used, and modified freely; it is distributed under the GPL-2/GPL-3 license as an R package available for Windows, Mac OSX, and GNU/Linux platforms. R is a popular general-purpose programming language with excellent support for sound processing thanks to a number of dedicated packages that *soundgen* imports and builds upon, particularly *tuneR* (Ligges, Krey, Mersmann, & Schnackenberg, 2016) and *seewave* (Sueur, Aubin, & Simonis, 2008). After installing R (https://www.r-project.org/), and preferably RStudio (https://www.rstudio.com), both of which are open-source, sounds can be generated from the command line using the *soundgen()* function. There is also an interactive graphical user interface (GUI), namely a Web app launched with the *soundgen_app()* function (Fig. 1).

Additional documentation is available on the project's homepage at http://cogsci.se/soundgen.html. The so-called "vignette" on sound synthesis, which is published and regularly updated together with the package, provides an illustrated, step-by-step manual on using the *soundgen()* function. Several demos, including the R code for creating dozens of human vocalizations (laughs, roars, screams, moans, etc.) and animal sounds, are available on the project's website. There are also numerous examples of both human and animal vocalizations built into the package itself and available through the interactive app.

How does soundgen compare to the alternatives?

In many research applications it is not necessary to synthesize a vocalization from scratch, but simply to take an existing recording and modify some of its acoustic characteristics. Some operations are trivial: for example, the amplitude envelope of a recording can be adjusted with any audio editor. Other acoustic manipulations are more challenging and require specialized software. For instance, the effect of f_0 on perceived dominance and mating preferences in humans has been investigated by manipulating f_0 experimentally, often together with formant frequencies (e.g., Fraccaro et al., 2013; Puts, Gaulin, & Verdolini, 2006). However, this manipulation also affects the average distance between neighboring formants (formant dispersion), making it difficult to determine whether it was f_0 or formants that caused the observed effect. It is also possible, although more technically challenging, to avoid this confound by manipulating f_0 and formant dispersion independently of each other (Kawahara, Masuda-Katsuse, & De Cheveigne, 1999; Reby et al., 2005; Taylor, Reby, & McComb, 2008). For example, Reby et al. scaled formants in the roars of male red deer without changing f_0 and demonstrated the important role of formant dispersion for exaggerating apparent body size during roaring contests.

Morphing is an interesting special case of manipulating existing recordings, in which certain acoustic features of sound A are gradually adjusted to match those of sound B, producing several hybrid sounds that combine characteristics of both A and B. A popular choice for such work is to use the STRAIGHT algorithm, which can morph broadly similar sounds on the basis of several manually specified landmarks that identify corresponding parts of the two spectrograms (Kawahara et al., 1999). For example, this technique was used to demonstrate categorical perception of macaque vocalizations by human listeners (Chakladar, Logothetis, & Petkov, 2008), to study acoustic features enabling individual recognition in macaques (Furuyama, Kobayasi, & Riquimaroux, 2017), and to prepare morphs of emotional vocalizations for a neuroimaging study (Salvia et al., 2014).

The techniques described above require that the researcher should prepare the stimuli in advance. For real-time manipulation, a promising tool is the *David* software (Rachman et al., 2018), which can modify the intonation and some aspects of voice quality and feed the new auditory stream back to the speaker with only a short delay. This technique was used to demonstrate that covert manipulation of a speaker's voice suggestive of particular emotional states induces the same emotion in the speaker, as predicted by the self-perception framework (Aucouturier et al., 2016). When a slightly longer delay is acceptable—for example, for voice manipulation during interaction instead of sensorimotor feedback—a powerful technique to use is frequency warping (Erro, Navas, & Hernáez, 2013). By stretching and warping the spectral envelope on the basis of a user-defined nonlinear function, it can achieve complex manipulations of the filter in real time, for example, raising the frequencies of a few individual formants to produce the impression that the speaker is smiling (Arias et al., 2018).

The greatest advantage of sound manipulation over synthesis is that it preserves subtle characteristics of the original recording. The results can be highly naturalistic, and as long as it is feasible to achieve the desired manipulation without resynthesizing the entire sound, this is the preferred method. However, some acoustic features cannot be modified so easily

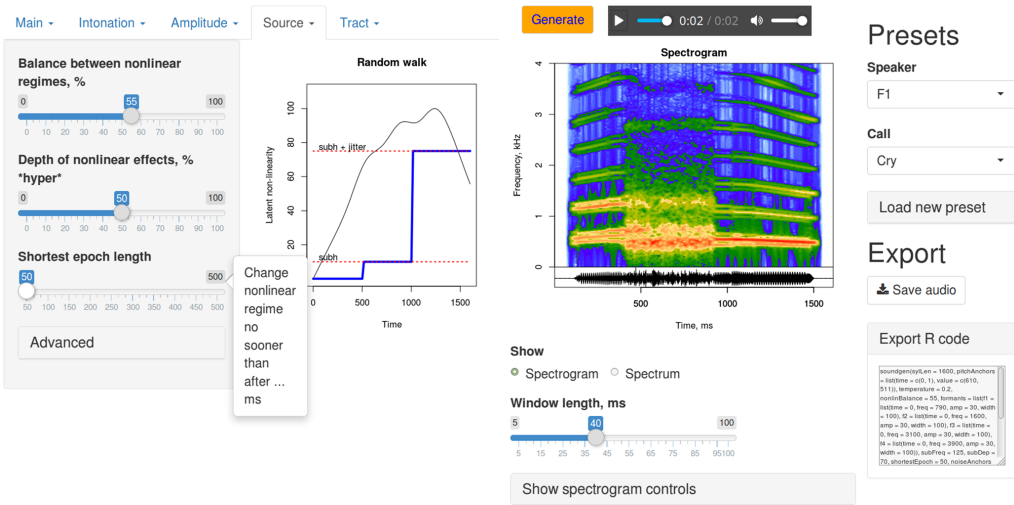

Fig. 1 Graphical user interface for *soundgen*

(e.g., the harmonic structure of the glottal source) or require complex transformations that may degrade the quality of the manipulated sound (e.g., frequencies of individual formants). Because of these limitations, sometimes it is preferable to synthesize a vocalization de novo, which requires an algorithm for generating a waveform from a list of control parameters—a so-called "vocoder."

Any parametric text-to-speech engine includes a vocoder, and any vocoder can potentially be used to generate nonverbal vocalizations as well as speech (e.g., Klatt & Klatt, 1990; Kreiman, Antoñanzas-Barroso, & Gerratt, 2010; Morise, Yokomori, & Ozawa, 2016). However, text-to-speech platforms usually prioritize efficiency and high-fidelity output (Tokuda et al., 2013; van den Oord et al., 2016; Wu, Watts, & King, 2016) over control of individual acoustic features (but see Birkholz, Martin, Xu, Scherbaum, & Neuschaefer-Rube, 2017; Drugman, Kane, & Gobl, 2012). As a result, they are not always optimal for academic acoustic research, particularly when working with non-speech sounds. One feature of the vocoders developed for speech synthesis is that they require a separate list of parameter values for each short frame. In other words, control parameters even for a short vocalization add up to a large matrix, which is unwieldy to work with manually. In addition, many control parameters resist simple adjustments. For example, the parameters that control voice quality do so by changing the shape of glottal pulses, and there is no simple way to predict which changes in control parameters will achieve the desired change in the output spectrum (Kreiman, Garellek, Chen, Alwan, & Gerratt, 2015).

Because of this complexity and opacity, vocoders are normally controlled by statistical models rather than manually; popular choices are hidden Markov models (HMM; Tokuda et al., 2013; Zen, Tokuda, & Black, 2009) or neural networks (Juvela et al., 2016; Ling et al., 2015; Wu et al., 2016). An even more "black-box" approach to statistical parametric synthesis is to build neural networks that directly generate raw waveforms (van den Oord et al., 2016). Whatever the exact statistical model, first it must be trained on a large annotated corpus, and the purpose of such tools is to take text as input and produce audio as output.

Text-to-speech systems are thus not the most appropriate tools if the input categories are longer than phonemes (e.g., vocalizations) or if the goal is to synthesize nondiscrete sounds (e.g., continuous modulated tones for psychophysical experiments). There have been a few attempts to synthesize nonverbal vocalizations, usually laughter, by adapting existing text-to-speech engines based on concatenative (Campbell, 2006), articulatory (Lasarcyk & Trouvain, 2007), or HMM parametric synthesis (Haddad, Cakmak, Sulír, Dupont, and Dutoit, 2016), but this research is scarce. An interesting alternative developed for resynthesis and manipulation of short verbal fragments is the UCLA Voice Synthesizer (Kreiman et al., 2010). This is an integrated analyzer–vocoder that can be analyze, reproduce, and then manipulate a recording, although it is optimized for working with speech and does not offer the researcher an easy solution for generating a completely new vocalization based on a high-level acoustic description.

Unlike integrated platforms for parametric speech synthesis, which consist of a vocoder and a statistical model to operate it, *soundgen* is essentially only a vocoder, giving the user direct access to acoustic controls. As a result, its control

parameters must be kept transparent and sparse—defined at a low temporal resolution with a few anchor points. Another difference between *soundgen* and the vocoders used for text-to-speech conversion is that the latter are designed to synthesize speech, so their settings are optimized for human voice, more specifically for the voice quality typically encountered in ordinary conversation. In particular, several mathematical models can approximate the shape of glottal pulses (Fant, Liljencrants, & Lin, 1985; Rosenberg, 1971; Shue & Alwan, 2010), but they are less accurate for nonmodal phonation (Gobl & Ní Chasaide, 2003) and may not be suitable for nonverbal vocalizations such as high-pitched screams or noisy roars. It is even less certain whether the same parametric models of glottal pulses will capture the excitation source in nonhuman animals, who display a variety of sound production mechanisms, from the glottal whistles of rodents to the syringeal phonation of birds (Goller, 2016). In fact, even if a vocoder closely matches the shape of glottal pulses, it does not guarantee that the perceptually relevant spectral characteristics will be captured successfully. This is why it has been suggested that auditory perception in humans is better modeled in the frequency than time domain (Doval & d'Alessandro, 1997; Kreiman et al., 2015). Furthermore, despite all the diversity of sound sources in the animal world, the source-filter model (Fant, 1960) still holds across species (Goller, 2016; Taylor & Reby, 2010). Capitalizing on the flexibility and general applicability of spectral-domain modeling, *soundgen* creates individual harmonics instead of individual glottal pulses (it does support pulse-by-pulse synthesis, but this is not the default method), making it straightforward to directly adjust the perceptually relevant characteristics of the glottal source for any mode of phonation and potentially across species.

In this sense *soundgen* is more similar to the Mammalian Vocal Synthesizer implemented in the MiRo biomimetic robot (Moore, 2016), which specifically aims to model voices of different mammals and has inspired some of *soundgen*'s features. The main difference is that *soundgen* offers more flexible control, making it more suitable for perceptual experiments that involve precise acoustic manipulations. On the other hand, the Mammalian Vocal Synthesizer is easier to control, and it runs in real time, making it a better choice for interactive systems such as social robots.

Implementation details

A brief summary of the fundamental principles of voice synthesis with *soundgen* is provided below. The actual control parameters are described fully from a user's perspective in the package documentation, particularly in the so-called "vignette" on sound synthesis.

Excitation source Although early vocoders used a train of rectangular impulses as a source of excitation, this produced a relatively flat source spectrum with too much high-frequency energy, which made the voice sound buzzy (Cabral, Renals, Richmond, & Yamagishi, 2007). Modern vocoders solve this problem by using more realistic excitation sources that resemble actual glottal pulses produced by vibrating vocal folds (Fant et al., 1985; Klatt & Klatt, 1990). In contrast to this time-domain model of the glottal excitation source, *soundgen* generates a separate sine wave for each harmonic, covering one voiced syllable instead of one glottal pulse at a time. At each time point the frequency of each sinusoidal component is an integer multiple of f_0, which can vary within one syllable. This approach is similar to tonal synthesis as implemented in the *seewave* R package and described in detail by Sueur (2018). For each harmonic h in $\{1, 2, \ldots\}$, the waveform $w_h(t)$ for harmonic h is synthesized as a sine wave with time-varying frequency $h * f_0(t)$ and amplitude $a_h(t)$:

$$w_h(t) = a_h(t) * sin\left(2 * pi * h/s * \sum_1^t f_0(t)\right),$$

where t is an integer index of the synthesized point, $f_0(t)$ is the instantaneous fundamental frequency at time t, $a_h(t)$ is the instantaneous amplitude of harmonic h at time t, and s is the number of points per second of audio (sampling rate). The phase of each harmonic is set to zero. The final waveform is then given by the sum of all harmonics. The number of synthesized harmonics is determined by the sampling rate: no harmonics are synthesized above the Nyquist frequency (half the sampling rate) to avoid aliasing. The relative strength of harmonics relative to f_0 is governed by a family of *rolloff* parameters, the most important of which gives the rate of exponential decay in the amplitude of upper harmonics: with each octave above f_0, the power of harmonics decreases by *rolloff* dB. Rolloff can also be automatically adjusted depending on f_0, and the source spectrum may assume more complex shapes than a simple exponent, providing more flexibility with the excitation source.

In addition to the harmonic component, *soundgen* generates broadband noise with a spectrum that is flat up to a certain threshold (by default 1200 Hz) and has an adjustable linear spectral slope of *rolloffNoise* dB/kHz in higher frequencies (Johnson, 2011; Stevens, 2000). Noise is created in the frequency domain by drawing from a uniform distribution, multiplied by the rolloff function and the vocal tract transfer function (see below), and then converted to a waveform via inverse short-time Fourier transform (STFT). The noise component is then modulated with its own amplitude envelope, which is specified independently of the amplitude envelope of the voiced component.

The manner in which the harmonic and noise components are added up depends on whether or not they should be filtered with the same vocal tract transfer function. The default behavior is to assume that the noise originates close to the glottis and

passes through the same vocal tract, as when an animal is breathing. If the source of the obstruction lies further from the glottis, its filter is different from that of the harmonic component (Stevens, 2000). To synthesize such non-glottal noises as hissing, *soundgen* can handle a separate filter function for its noise component. It is also possible to synthesize voiceless sounds without any harmonic component.

Filter The sound changes as it passes through the vocal tract: Some frequencies, known as "formants," are amplified, whereas other frequencies may be dampened (Fant, 1960). The spacing between formant frequencies depends on the length of the vocal tract from the source of excitation (glottis in mammals, syrinx in oscine birds) to the opening through which the air escapes. In *soundgen* an entire vocalization is first synthesized without formants, and then it is filtered through the vocal tract transfer function—a matrix that specifies a scale coefficient for each frequency bin and each time step in the spectrogram of an unfiltered waveform. This process involves taking an STFT of the generated waveform, multiplying the resulting spectrogram by the vocal tract transfer function, and then performing inverse STFT to transform the signal back to a waveform.

The transfer function is determined by time-varying frequencies, amplitudes, and bandwidths of several formants, which are either specified by the user or estimated from the length of the vocal tract. If the user provides the frequencies of the first few formants, these are used to estimate the apparent vocal tract length (VTL) using the regression method described in Reby et al. (2005). Additional formants are then added above the user-specified ones, with frequencies determined according to the uniform tube model (Stevens, 2000):

$$F_n = \left(2^{*}n{-}1\right)/2^{*}d,$$
$$d = c/\left(2^{*}VTL\right),$$

where F_n is the nth formant, d is formant dispersion, and c is the speed of sound in warm air (35,400 cm/s). If only VTL is specified, a neutral schwa [ ] with equidistant formants is produced. Formant frequencies are adjusted according to the degree of mouth opening using a formula adapted from Moore (2016):

$$\Delta F = \left(m{-}0.5\right)^{*} c/\left(4^{*}\,\mathrm{VTL}\right),$$

where ΔF is the change in formant frequency and m is the degree of mouth opening (0 = *closed*, 1 = *fully open*, 0.5 = *default neutral position, no adjustment*). When the mouth is completely closed, the sound is also nasalized by increasing the bandwidth of the first formant to 175 Hz and creating a new zero-pole pair in the vicinity of the first formant, as described in Hawkins and Stevens (1985). These settings are presumably specific to human voice, so it is not recommended to use the closed mouth feature for animal vocalizations; the corresponding formant transitions can be specified manually instead. The effects of sound radiation through the lips or the nose are controlled by two separate parameters instead of embedding them in the source spectrum, which makes it easy to adjust the settings for nonhuman biological sounds.

Unless specified by the user, formant bandwidths are estimated from frequency using an empirical formula derived from human phonetic research, namely the TNF-63 approximation (Tappert, Martony, & Fant, 1963) corrected below 500 Hz to increase bandwidth at low frequencies (Khodai-Joopari & Clermont, 2002). Once time-varying formant frequencies and bandwidths have been determined, the vocal tract transfer function is calculated in the frequency domain by using a standard all-pole model if there are only formants, or a zero-pole model if there are also antiformants (Stevens, 2000). The only modification of these models in *soundgen* was to enable more flexible control over the strength of individual formants. Mathematical details of this algorithm are beyond the scope of the present article; the relevant code can be found in the function *getSpectralEnvelope()*.

Other control parameters Both the source of excitation and the filter can be modified in many ways using a number of control parameters, some of which are mentioned below. Most of these parameters are vectorized, so that the amount and quality of each effect can vary over time. A vibrato can be added as a sinusoidal modulation of f_0 with adjustable frequency and depth. Variation in f_0 can also be stochastic (jitter), again with time-varying depth and frequency—from slow, vibrato-like random fluctuations to very rapid pitch jumps that can be used to simulate harsh voices in roars and noisy screams. Attack at the beginning and end of voiced fragments can be specified separately from the overall amplitude envelope, and rapid stochastic amplitude modulation can be added to simulate pulse-to-pulse variation in glottal pulses (shimmer). Low-frequency amplitude modulation with adjustable depth, frequency, and shape is useful for making trill-like sounds.

A special subroutine in *soundgen* is devoted to nonlinear effects, namely subharmonics (or sidebands) and deterministic chaos (Wilden, Herzel, Peters, & Tembrock, 1998). Subharmonics are created by generating additional harmonics in the excitation source, which corresponds to introducing an additional fundamental frequency (g_0) at an integer ratio to f_0. Chaos is simulated by adding strong jitter and shimmer. The parts of vocalization affected by nonlinear phenomena can be specified explicitly by the user or determined stochastically. A random walk is generated; its bottom part corresponds to fragments with no nonlinear effects, the middle part to subharmonics, and the highest part to both subharmonics and chaos (cf. Fitch, Neubauer, & Herzel, 2002). This makes it possible to generate sounds with unpredictable transitions between different regimes of nonlinear phenomena.

Soundgen contains a number of high-level hyperparameters that affect multiple acoustic features at once. For example, f_0

and formant frequencies can be adjusted in a coordinated manner with the *maleFemale* parameter. The most important hyperparameter, *temperature*, adjusts the amount of stochastic variation in f_0 contour, voice quality, and most other control parameters. A natural vocalization is seldom completely static, and this stochastic behavior offers an easy way to introduce some variability without manually coding every irregularity. If *temperature* is above zero, calling the *soundgen()* function repeatedly with the same settings does not produce identical output every time. This is helpful when the purpose is to create a number of authentic-sounding and similar, but not identical, vocalizations. When the goal is to generate a sound with high precision (e.g., when synthesizing multiple modifications of the same basic vocalization for perceptual testing), stochastic behavior is not desirable, and *temperature* should be set to a small positive value (setting it to exactly zero disables the addition of new formants above the user-specified ones and is not recommended).

Validation experiment

To validate *soundgen* as a tool for synthesizing human nonverbal vocalizations, a number of laughs, screams, moans, and other sounds were synthesized aiming to approximately reproduce the original recordings. The similarity of the synthetic stimuli to their originals was then assessed in a perceptual experiment by means of comparing their ratings on three continuous scales—valence, arousal, and authenticity—as well as their classification by emotion.

Method

Stimuli Sixty authentic human nonlinguistic vocalizations were chosen from a previously published corpus (Anikin & Persson, 2017) and reproduced with *soundgen* 1.1.1. The selection criteria were (1) a minimum amount of background noise, echo, clipping, or other acoustic impurities, (2) a high degree of consensus on what acoustic type (laugh, scream, and so on) the sound represented in a previous cross-linguistic naming study (Anikin, Bååth, & Persson, 2018), and (3) high perceived authenticity, as reported by Anikin and Lima (2018). To constrain the acoustic complexity of the stimuli, longer bouts were truncated. The average duration was 1.5 ± 0.7 s ($M \pm SD$), range 0.3 to 3.4 s (Table 1). All sounds were then normalized for peak amplitude and down-sampled to a rate of 22050 Hz.

Control parameters were chosen manually, on the basis of an iterative visual comparison of the spectrogram with the target. The difficulty of this task varied depending on the complexity of the target, from a few minutes of work for simple moans or screams to a few hours for some laughs. Whenever possible, the entire sound was synthesized with a single command (see Table 2 for an annotated example). Some of the more complex polysyllabic vocalizations were synthesized one segment at a time and then concatenated. The *temperature* parameter was usually kept positive—that is, most sounds were synthesized in a stochastic mode. For example, for polysyllabic vocalizations the length of syllables and pauses between them was deliberately allowed to vary at random and deviate from the original. The synthetic sounds were thus not meant to be exact replicas of the originals, but only approximations. Highly stochastic sounds, such as screams consisting of segments with various vocal regimes (tonal, subharmonics, deterministic chaos) were generated several times with the same settings, and the copy closest to the original was retained for testing.

Table 1 Characteristics of the original recordings in Experiment 1

Acoustic type	Number of sounds (M/F)	Duration, s Mean [range]
Cry	10 (4/6)	2.1 [1.3, 2.7]
Gasp	5 (2 /3)	1.6 [1.1, 2.6]
Grunt	5 (3/2)	0.5 [0.3, 0.8]
Laugh	10 (5/5)	1.9 [1.0, 3.4]
Moan	10 (5/5)	1.7 [0.7, 3.1]
Roar	10 (6/4)	1.4 [0.6, 3.2]
Scream	10 (2/8)	1.4 [0.4, 3]
TOTAL	60 (27/33)	1.5 [0.3, 3.4]

Supplementary materials for this article, including all sound files, R code for their generation, raw data from the validation experiment, and R scripts for statistical analyses, can be downloaded from http://cogsci.se/publications/anikin_2018_sova.html. Note that *soundgen* has been extensively upgraded since the time of the validation experiment. It may therefore be preferable to use the more up-to-date demos from the project's website, rather than the code from the supplementary materials, as templates for sound synthesis.

Procedure The validation experiment included three separate tasks, which were performed by different groups of participants. In Tasks 1 and 2, participants rated 60 synthetic (Task 1) or real (Task 2) vocalizations on three scales: valence ("How pleasant or unpleasant is the experience?"), arousal ("How high is the level of energy, alertness?"), and authenticity ("Does the person sound natural, like in real life?"). Prior to testing, each participant was informed which type of sounds, human or synthetic, they would hear, ensuring that authenticity ratings would not be interpreted as binary guessing between "real" and "fake" sounds. To minimize the correlation between ratings on the three scales, only one scale was displayed in each block. Each sound was thus presented three times. The order of blocks and of sounds within each block was randomized for each participant.

Table 2 R code for generating moaning in *soundgen* version 1.1.1 and above (stimulus #38)

```
# Begin call to soundgen()
s = soundgen(
  nSyl = 3,                                # number of voiced syllables
  sylLen = 520,                            # average length of syllables (ms)
  pauseLen = 400,                          # average pause between syllables (ms)
  pitchAnchors = c(270, 210),              # f0 drops from 270 to 210 Hz in each syllable
  vibratoDep = .5,                         # add vibrato, half a semitone deep
  rolloff = c(-20, -30),                   # rolloff drops from -20 to -30 dB/oct in each syllable
  formants = c(900, 1400, 3100, 3750,
               4900, 6200, 6800),          # formants F1 to F7 (Hz)
  mouthAnchors = c(.5, .6),                # mouth opens slightly across syllables
  noiseAnchors = data.frame(
    time = c(0, 350, 520, 530, 700, 830),  # time anchors (ms) at which noise amplitude is defined
    value = c(-55, -40, -45, -60, -40, -60)  # loudness (dB) of turbulent noise at each time anchor
  ),
  rolloffNoise = c(0, 0, -5),              # time-varying slope of noise spectrum
  temperature = 0.15,                      # general level of stochasticity (0.15 is very high)
  tempEffects = list(formDrift = 2),       # higher-than-default stochasticity in random drift of formants
  samplingRate = 22050,                    # desired sampling rate of output (Hz)
  play = T, plot = T, osc = T              # play and plot the output, show oscillogram
)

# Play, plot, export
playme(s, 22050)
spectrogram(s, 22050)
seewave::savewav(s, f = 22050, filename = 'moans.wav')
```

For more examples, see demos on the project's homepage: http://cogsci.se/soundgen.html

In Task 3, participants indicated the emotion portrayed by each sound and rated their confidence in this classification. There were ten emotional categories to choose from: *amusement*, *anger*, *disgust*, *effort*, *fear, pain*, *pleasure*, *sadness*, *surprise*, and *other/neutral/don't know*. To avoid presenting the same sound twice (the original and the synthetic version), this task was split into two subtasks, each with 60 unique sounds (30 human + 30 synthetic). Each subtask was performed by a different sample of participants. Sounds could be repeated as many times as needed. The average completion time was 10–15 min.

Participants Out of the 106 participants included in the final analysis, 47 were volunteers contacted via online advertisements, and 59 were recruited via https://www.prolific.ac and received £1 or £1.5, depending on the task. An informal comparison of the data from volunteers and paid participants revealed no systematic differences, and their responses were pooled. Each participant performed only one task. The numbers of responses per sound were as follows: ratings of synthetic sounds 19.6 (range 19.0 to 20.3), ratings of human sounds 21.7 (range 21.0 to 22.0), forced choice classification of emotion for mixed human and synthetic sounds 30.7 (range 28.0 to 35.0).

Data were collected via the Internet. Participants were informed about the goal of the study and agreed with the conditions of confidentiality by clicking the active link after reading the instructions. No personal or demographic information was collected. Online experiments are increasingly being used for academic research (Hewson, Vogel, & Laurent, 2016), but the responses may be noisy compared to those from face-to-face testing. To ensure data quality, all submissions were first manually checked for fraud (e.g., clicking through stimuli very fast and without varying the responses). Once data collection was complete, a second round of verification was performed by means of correlating the ratings provided by each participant with the median ratings per stimulus aggregated from all responses. Participants were excluded from the analysis on the basis of the following criteria: (1) correlation with global median < 0.3 on any two scales, or (2) correlation with global median < 0 on either valence or arousal scale, or (3) proportion of emotion classification corresponding to the most commonly chosen emotion per stimulus < 0.3. This identified five participants, typically with very short response times, who were removed from further analysis. In addition, all trials with response time under 1 s (0.4% of the data) were excluded, since they presumably represented technical glitches.

Statistical analysis Except when otherwise stated, all analyses were performed on unaggregated, trial-level data using mixed models. To account for non-independence of observations, all models included random intercepts per participant and per stimulus. Ratings on continuous scales were not normally distributed, and they were modeled with beta distributions. Bayesian models were created in Stan computational framework (http://mc-stan.org/) accessed with *brms* package (Bürkner, 2017). To improve convergence and guard against overfitting, mildly informative regularizing priors were used for all regression coefficients. Fitted values are reported as the median of the posterior distribution and 95% credible interval (CI). Intraclass correlation coefficients were calculated with the *ICC* package (Wolak, Fairbairn, & Paulsen, 2012).

Results

Valence, arousal, and authenticity ratings The reliability of valence ratings was moderate: The intraclass correlation coefficients (ICCs) were .61 for human sounds and .54 for synthetic sounds. The average per-stimulus valences of human and synthetic sounds were highly correlated: Pearson's r = .88, $F(1, 58) = 207.1$, $p < .001$. A more nuanced analysis with mixed models revealed that human and synthetic sounds were rated similarly on the valence scale for all call types except laughter (Fig. 2A), for which the real recordings were judged to be more positive than the synthetic sounds: by + 0.41 on a scale of – 1 to + 1, 95% CI [0.29, 0.52].

The consistency with which participants used the arousal scale was again moderate: ICCs = .42 for human and .55 for synthetic sounds. The average arousal ratings of human and synthetic sounds were highly correlated: r = .92, $F(1, 58)$ = 327.8, $p < .001$. There was a slight tendency for human sounds to have higher arousal ratings than the synthetic sounds for several call types (Fig. 2B), but the difference was statistically significant only for gasps: + 0.44, 95% CI [0.29, 0.60]. Interestingly, the valence and arousal scales were not completely orthogonal. Averaging per stimulus, there was a quadratic relationship between valence and arousal ratings, $F(2, 117) = 43.4$, $p < .001$, $R^2 = .43$: Sounds with either very positive or very negative valence tended to have high arousal ratings.

There was little agreement among participants about the authenticity of individual sounds, with ICCs of only .15 for human sounds and .27 for synthetic sounds. The correlation between the average authenticity ratings for human and synthetic sounds was also weaker than in the case of valence and arousal ratings, although it was still significantly higher than would be expected by chance: Pearson's r = .33, $F(1, 58)$ = 6.9, p = .01. As expected, the real recordings were overall judged to be more authentic than the computer-generated sounds: + .22 on a scale of – 1 to + 1, 95% CI [.07, .36]. It is worth emphasizing that before the experiment, participants were informed which type of sounds, human or synthetic, they would hear. Nevertheless, the difference in authenticity was statistically significant only for laughs (.74 [.57, .89]), roars (.33 [.17, .49]), and cries (.27 [.09, .44]), but not for the remaining four call types (Fig. 2C). In fact, the average authenticity of 22 out of the 60 synthetic sounds was higher than the authenticity of the original recordings.

Considering the experimental nature of the algorithms used for adding subharmonics and chaos, it was important to check the authenticity of sounds containing these nonlinear vocal phenomena. Out of 60 stimuli, six were synthesized with pitch jumps, another six with subharmonics, two with biphonation (the originals contained ingressive whistles), 23 with chaos, and 23 without nonlinear effects. The differences in authenticity between the real and synthetic versions were similar for sounds with subharmonics versus no nonlinear effects (.12 [– .05, .28], on a scale of – 1 to + 1), chaos versus no nonlinear effects (– .09 [– .19, .01]; a marginal advantage for sounds with chaos), and pitch jumps versus no nonlinear effects (.03 [– .13, .19]). More stimuli would need to be synthesized to test this further, but at least there was no obvious disadvantage of synthetic sounds with nonlinear effects, in terms of their perceived authenticity.

Recognition of emotion A new sample of participants classified a mixture of the same 60 real and 60 synthetic sounds by emotion. The stated certainty of this classification was high for all sounds (Fig. 2D), but it was .12 (95% CI [.09, .15]) higher for the original recordings than for their synthetic reproductions. This difference was small but consistent across most call types. Synthetic sounds were also slightly more likely to be placed in the residual category of *Other/neutral/don't know*: In 19.1% of trials for synthetic sounds, and only 6.6% for human sounds, predicted difference = 12.4%, 95% CI [10.3, 14.6]. The normalized Shannon entropy of the counts of emotional labels applied to a particular sound (hereafter, "emotion vectors"; see Anikin et al., 2018) was 14.4% (95% CI [7.5, 21.0]) higher, on a scale of 0% to 100%, for synthetic versus human sounds. This suggests that there was slightly less agreement among the listeners about the emotion portrayed by synthetic sounds than with the original recordings.

The emotions associated with each call type were broadly similar for both real and synthetic vocalizations (Fig. 3). Looking at individual stimuli rather than call types, the correlation between the matrices of classification counts for real and synthetic sounds was high: r = .82, $\chi^2(531) = 4{,}644.5$, $p < .001$. Common measures of interrater agreement, such as Fleiss's kappa, were not strictly appropriate: Although this was a forced choice classification task, the categories overlapped semantically and were not exclusive. For example, if laughs were classified as *amusement* by some participants and *pleasure* by others (as, indeed, sometimes happened), this might not be a sign of genuine disagreement among the raters,

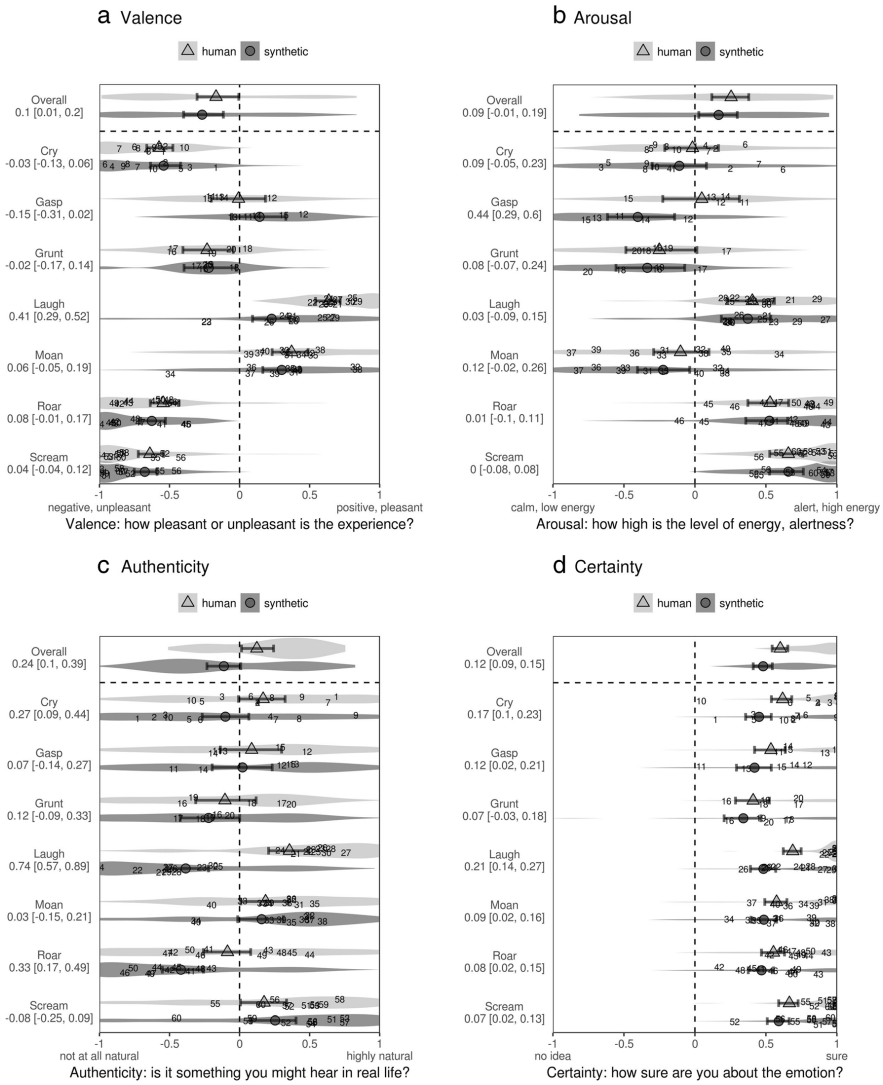

Fig. 2 Ratings of 60 human and 60 synthetic nonlinguistic vocalizations. Violin plots show the distributions of individual ratings for each call type (the "overall" category is aggregated per stimulus), with individual stimuli marked by indices from 1 to 60. Solid points with error bars show fitted values per call type: the median of the posterior distribution with 95% CI. Contrasts between real and synthetic sounds per call type are shown as axis labels

but kappa would drop. A more appropriate method might be to compare the classifications of each real and synthetic sound by correlating their emotion vectors. Correlations close to 1 would indicate that the distributions of responses were nearly identical for the original and synthetic versions of a particular sound, whereas low correlation would indicate systematic differences in the ways these sounds were categorized by participants. The average observed correlation between the emotion vectors of human and synthetic sounds was high ($r = .77$), as compared to the correlation expected by chance ($r = .16$, estimated by permutation). As is shown in Fig. 4, the correlations of emotion vectors were over .75 for almost all cries,

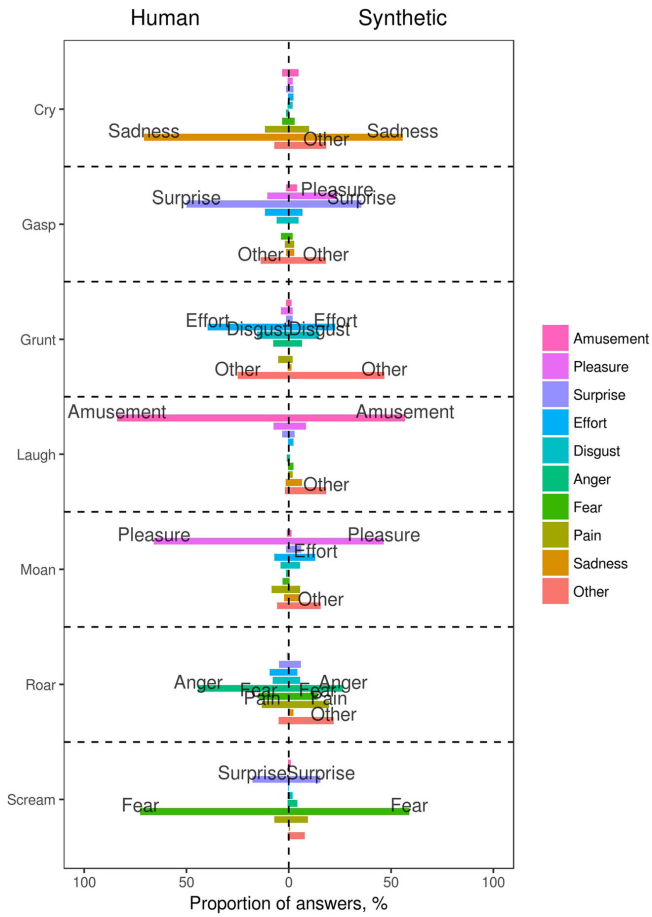

Fig. 3 Forced choice classification of sounds in terms of their underlying emotion: Proportions of responses averaged per call type. Assuming that the synthetic versions are functionally equivalent to the original recordings, the two halves of the figure should be mirror images of each other. All bars over 12% high are labeled, to simplify reading the graph

laughs, and screams, but they were more variable in the remaining four call types, suggesting that a few synthetic stimuli differed from the original recordings in terms of perceived emotion.

Discussion of the validation experiment

The validation experiment was designed to investigate whether parametric voice synthesis, as implemented in the open-source R library *soundgen*, is sufficiently flexible and precise to reproduce human nonverbal vocalizations in such a way that the perceived valence, arousal, authenticity, and emotion of the synthetic versions would be similar to those of the original recordings. It must be reiterated that synthetic sounds were not exact replicas of the originals, but stochastic generative models that aimed to preserve only the most salient and easily identifiable acoustic characteristics of the original. Even so, valence and arousal ratings of human and synthetic sounds were tightly correlated, demonstrating that the perceived affective meaning of synthetic vocalizations was very close to that of the original recordings. More importantly, synthetic vocalizations covered the entire available range on valence and arousal scales and were rated as consistently as the original human recordings. The validation study thus demonstrated that all kinds of human nonverbal vocalizations—high- or low-intensity, hedonistic or aversive—can potentially be synthesized with *soundgen*.

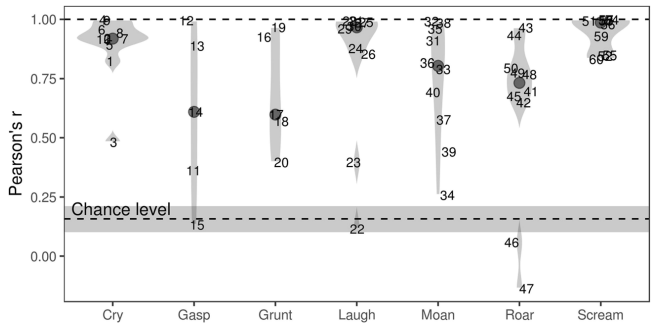

Fig. 4 Pearson's correlations between emotion vectors (counts of emotional labels applied to a particular sound) for real and synthetic vocalizations. Solid points mark the median for each call type, and violin plots show the distribution of values for individual stimuli, which are marked 1 to 60. The shaded area shows the correlation that would be expected by chance (median and 95% CI), which was estimated by permuting the dataset

In addition to ensuring that synthetic sounds were close to the originals in terms of the perceived valence and arousal of the speaker, it was important to ascertain that they sounded natural and not too machine-like. Participants in the validation study were told beforehand whether they would hear human or synthetic vocalizations and then had to rate them on naturalness (authenticity). The motivation for this design was the need to evaluate the relative authenticity of both human and synthetic sounds without making the task into an attempt to guess which sounds were synthetic and which were not—this would not be particularly meaningful, since many recordings contain extraneous clues to their nonsynthetic nature (traces of background noise, a slight echo, and so on). In addition, the aim was to synthesize natural-sounding vocalizations, not to trick the listeners into believing them to be real, which is also the established practices when testing the naturalness of synthetic speech (e.g., Erro et al., 2014; van den Oord et al., 2016).

In line with previous reports (Bänziger, Mortillaro, & Scherer, 2012; Lima, Castro, & Scott, 2013), authenticity ratings were highly variable for both human and synthetic sounds and presumably depended on how often similar sounds occurred in everyday life, how genuine the speaker's emotion appeared to be, and (for synthetic sounds) how convincing or "human-like" they sounded. Given the diversity of factors that may have affected authenticity rating of individual stimuli, they are not easy to interpret in themselves—the key metric is the difference in perceived authenticity within each pair of real and synthetic sounds. As it turned out, recordings of real people had a 10% advantage in terms of authenticity, but this difference strongly depended on the acoustic type: It was pronounced for laughs, moderate for cries and roars, and absent for gasps, grunts, moans, and screams. In other words, it was hardest to succeed in synthesizing the most acoustically complex, polysyllabic vocalizations such as cries and laughs. These vocalizations contain a rich gamut of unvoiced sounds and physiological noises (snuffling, spluttering, gurgling, wheezing, etc.), rapid formant transitions, episodes of biphonation, and a variety of transients that make it difficult for the operator to read the relevant acoustic parameters off a spectrogram and to control the synthesis manually. In addition, laughs and cries often contain a variety of syllables—they are not at all as repetitive as suggested by the conventional *Ha-ha-ha*. As a result, they have to be painstakingly synthesized in multiple steps, sometimes one syllable or even one acoustic "layer" at a time, which is time-consuming for the researcher.

Despite their lower authenticity ratings compared to the real recordings, most synthetic laughs and cries were readily recognized and correctly labeled in terms of the underlying emotion. It is therefore still possible to synthesize them, particularly if only the most authentic-sounding stimuli are retained for testing. Nevertheless, acoustically simpler vocalizations, such as moans and screams, represent much easier targets for synthesis with *soundgen*. Most synthetic vocalizations of these call types were judged to be highly authentic by the raters, although in a few cases the emotion they expressed was different from the emotion expressed by the original recording. This may partly be explained by the inherently ambiguous nature of such vocalizations as moans, grunts, and gasps (Anikin et al., 2018). In fact, the consistency with which only human (not synthetic) sounds were classified by emotion also varied across call types: The correlation of emotion vectors estimated by permutation was on average high ($r > .95$) for laughs, cries, and screams, whereas for gasps, grunts, and roars it was lower ($r \sim .80$). When there is no obvious emotion category to which to assign a sound, responses become noisier, so it is less likely that the classification decisions will be exactly the same for an original recording and its synthetic version.

There is also a second potential explanation for differences in emotional classification of certain real and synthetic sounds with high authenticity ratings. Since vocalizations like grunts and moans can potentially express a wide range of meanings,

even relatively minor acoustic variations might suffice to shift the interpretation from one emotion to another. For example, Gasps 11 and 15 both had low correlations of emotion vectors between the original and synthetic versions (Fig. 2C), but for entirely different reasons. Gasp 11 had a low authenticity rating and a higher proportion of *Don't know* responses than did the original recording, indicating that it was not synthesized very successfully. In contrast, the synthetic Gasp 15 had above-average authenticity ratings and was consistently classified as *Pleasure*, whereas the original recording was variously classified as *Surprise*, *Don't know*, or *Effort* (the actual emotion expressed by the speaker was pleasant surprise). Subtle acoustic changes may thus suffice to cause a considerable change in the meaning of an inherently ambiguous vocalization, and identifying the responsible acoustic characteristics is exactly the kind of task that *soundgen* was designed for.

Taken together, the results of the validation experiment suggest that most of the synthetic sounds preserved the essential acoustic characteristics of the original recordings to the extent that listeners exposed to human and synthetic sounds drew the same inferences about the mental state of the caller, as measured by either continuous (valence, arousal) or categorical (emotion) outcomes. However, the most successful synthetic stimuli in the validation experiment were relatively short and simple, whereas complex and polysyllabic vocalizations, such as bouts of laughing or crying, were often rated as less natural, demonstrating that there are limits to the level of complexity that *soundgen* (or its operator) can handle.

Suggested applications

Soundgen is not designed for text-to-speech conversion. Speech consists of very rapid, highly variable formant transitions and amplitude modulation. It is too complex to encode more than a few phonemes manually, which is why statistical modeling of the parameters controlling the vocoder is predominant in parametric text-to-speech systems (Schröder, 2009). However, *soundgen*'s explicit manual control becomes more appealing when the target vocalization is short or repetitive, since a reasonable synthetic approximation can be created rapidly, without having an annotated training corpus and without recalibrating the algorithm for each new species. More specifically, *soundgen* can be useful for the following applications:

1. *Synthesis of human nonverbal vocalizations, as in the validation experiment*. Naturally, the mission of parametric voice synthesis is not simply to recreate existing recordings, but to modify them in systematic ways or to generate novel sounds with desired acoustic properties (e.g., Gobl & Ní Chasaide, 2003). To explore some of these possibilities, in a follow-up study (Anikin, 2018) soundgen was used to manipulate two acoustic features that could previously be analyzed only indirectly—nonlinear vocal phenomena and the rolloff of harmonics in the source spectrum. This study provided the first evidence that perceptual effects of nonlinear phenomena depended on their type (subharmonics, chaos, or pitch jumps) and on the type of vocalization in which they occurred. It also shed new light on the effect of source spectrum on the perceived level of arousal and, for relatively ambiguous vocalizations, their valence. Such acoustic manipulations would be difficult, if not impossible, to perform without resynthesizing the sound. It would be equally difficult to elucidate the perceptual effects of such relatively subtle acoustic features on the basis of a traditional acoustic analysis of recorded vocalizations. In the future, it will also be interesting to test the perceptual consequences of manipulating other acoustic features that have previously been reported to correlate with affective states and intentions: intonation (Banse & Scherer, 1996; Ohala, 1984; Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001), formant frequencies and transitions (Puts et al., 2006; Reby et al., 2005; Wood, Martin, & Niedenthal, 2017), the duration and number of syllables (Briefer, 2012), and many others. Moreover, synthetic vocalizations can easily be morphed by interpolating between values of control parameters manually or with the help of the built-in *morph()* function, which makes it straightforward to create series of graded stimuli, test for categorical perception, and so on.
2. *Synthesis of animal vocalizations*. Although not yet formally tested, *soundgen* should be capable of creating high-quality synthetic versions of animal calls. In fact, many design features of *soundgen* (focus on nonlinear effects, excitation source defined in the frequency rather than time domain, etc.) were specifically intended to make the algorithm generalizable to nonhuman sounds. Sine-wave synthesis has been used to produce mammalian (DiMattina & Wang, 2006; Snowdon & Pola, 1978) and avian (Margoliash, 1983) calls since the 1970s, but without specialized software only simple, pure-tone vocalizations could be created. *Soundgen*'s closest modern relative is the Mammalian Vocal Synthesizer (Moore, 2016), which is intended for real-time generation of biologically plausible sounds in a biomimetic robot. Some examples of animal calls synthesized with *soundgen* are included in the preset library published with the package.
3. *Synthesis of voice-like stimuli for psychophysical experiments*. *Soundgen* may prove useful for research on psychophysics, auditory perception, and cross-modal associations, since it offers a straightforward way to create acoustically complex sounds with precisely controlled temporal and spectral characteristics. For example, it was

used to create the stimuli for research on cross-modal sound-color correspondences that required finely controlled acoustic manipulations, such as generating formant transitions without modifying the spectral centroid of synthetic vowels (Anikin & Johansson, 2018).

4. *Teaching of phonetics and bioacoustics*. The interactive app (Fig. 1) offers the advantage of immediately hearing and visualizing the effects of modifying individual acoustic features such as the glottal source, frequencies and bandwidths of individual formants, various nonlinear effects, and other acoustic features that are easier to understand with an interactive demonstration. This functionality of *soundgen* can be valuable in an educational setting.
5. *Integration with text-to-speech platforms, use in human–machine interaction*. Although suboptimal for synthesizing speech as such, *soundgen* or some of its subroutines can be adapted for introducing simple nonverbal vocalizations into speech synthesized with another engine. One way of adapting *soundgen* for this purpose would be to compile a library of presents for several common vocalizations, such as laughing or sighing, and to vary them on the basis of simple rules for modifying the acoustic parameters in accordance with the desired level of emotion intensity and valence. Nonverbal vocalizations as well as artificial, nonbiological sounds (Read & Belpaeme, 2016) are particularly interesting in the context of social robotics, since they have the potential to enrich interaction without an "uncanny valley" effect. *Soundgen* is an open-source project, so the relevant code can be easily adapted or translated to fit another platform, potentially facilitating the development of other tools. However, the principle of continuous sine-wave synthesis makes *soundgen* less suitable for the real-time generation of continuously concatenated short fragments than are vocoders that generate individual glottal pulses. In addition, *soundgen* is currently somewhat slower than real time (the exact speed depends on the nature of the synthesized sound). Computational efficiency would have to be improved for it to be used in interactive systems, such as social robots, but this slowness is not a drawback when the goal is to synthesize stimuli for testing.
6. *Creation of special sound effects*. *Soundgen* is primarily developed as a research tool, but it can be adapted for the task of creating a wide variety of nonrepetitive, parametrically controlled human and animal sounds, which can potentially be of interest for the gaming and entertainment industry.

Author note I am grateful to Dan Stowell, Jérôme Sueur, Nick Campbell, Rasmus Bååth, Robert Eklund, Roger Moore, Sarah Hawkins, and Tobias Mahlmann for their help with developing *soundgen* and to Tomas Persson and three anonymous reviews for commenting on an earlier version of this article.

References

Anikin, A. (2018). *The perceptual effects of manipulating nonlinear phenomena and source spectrum in human nonverbal vocalizations*. Manuscript submitted for publication.

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior*, *42*, 53–80.

Anikin, A., & Johansson, N. (2018). *Implicit associations between individual properties of color and sound*. Manuscript in preparation.

Anikin, A., & Lima, C. F. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology*, *71*, 622–641. https://doi.org/10.1080/17470218.2016.1270976

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

Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., & Aucouturier, J. J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, *14*, 1–12. https://doi.org/10.1109/TAFFC.2018.2811465

Aucouturier, J. J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., & Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, *113*, 948–953.

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

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*, 1161–1179.

Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., & Neuschaefer-Rube, C. (2017). Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. *Computer Speech & Language*, *41*, 116–127.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology*, *288*, 1–20.

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

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Cabral, J. P., Renals, S., Richmond, K., & Yamagishi, J. (2007). Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proceedings of the 6th ISCA Speech Synthesis Workshop* (pp. 113–118). Grenoble, France: International Speech Communication Association.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech and Language Processing 14*, 1171–1178.

Chakladar, S., Logothetis, N. K., & Petkov, C. I. (2008). Morphing rhesus monkey vocalizations. *Journal of Neuroscience Methods*, *170*, 45–55.

DiMattina, C., & Wang, X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *Journal of Neurophysiology*, *95*, 1244–1262.

Doval, B., & d'Alessandro, C. (1997). Spectral correlates of glottal waveform models: An analytic study. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97* (Vol. 2, pp. 1295–1298). Piscataway, NJ: IEEE Press.

Drugman, T., Kane, J., & Gobl, C. (2012). Modeling the creaky excitation for parametric speech synthesis. In *Thirteenth Annual Conference of the International Speech Communication Association* (pp. 1424–1427). Grenoble, France: International Speech Communication Association.

Erro, D., Navas, E., & Hernáez, I. (2013). Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech and Language Processing*, *21*, 556–566.

Erro, D., Sainz, I., Navas, E., & Hernaez, I. (2014). Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, *8*, 184–194.

Fant, G. (1960). *Acoustic theory of speech perception*. The Hague, The Netherlands: Mouton.

Fant, G., Liljencrants, J., & Lin, Q. G. (1985). A four-parameter model of glottal flow. *Department for Speech, Music and Hearing Quarterly Progress and Status Report*, *26*(4), 1–13.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, *63*, 407–418.

Fraccaro, P. J., O'Connor, J. J., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour*, *85*, 127–136.

Furuyama, T., Kobayasi, K. I., & Riquimaroux, H. (2017). Acoustic characteristics used by Japanese macaques for individual discrimination. *Journal of Experimental Biology*, *220*, 3571–3578.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*(1–2), 189–212.

Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 378–423). Singapore: Wiley-Blackwell.

Goller, F. (2016). Sound production and modification in birds—Mechanisms, methodology and open questions. In C. Brown & T. Riede (Eds.), *Comparative bioacoustics: An overview* (pp. 165–230). Sharjah, UAE: Bentham Science.

Haddad, K., Cakmak, H., Sulír, M., Dupont, S., & Dutoit, T. (2016). Audio affect burst synthesis: A multilevel synthesis system for emotional expressions. In *2016 24th European Signal Processing Conference (EUSIPCO)* (pp. 1158–1162). Piscataway, NJ: IEEE Press.

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, *21*, 531–540.

Hawkins, S., & Stevens, K. N. (1985). Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels. *Journal of the Acoustical Society of America*, *77*, 1560–1575.

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

Johnson, K. (2011). *Acoustic and auditory phonetics* (3rd ed.). Hoboken, NJ: Wiley-Blackwell.

Juvela, L., Wang, X., Takaki, S., Airaksinen, M., Yamagishi, J., & Alku, P. (2016). Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks. In *INTERSPEECH* (pp. 2283–2287). Grenoble, France: International Speech Communication Association.

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based *F*0 extraction: Possible role of a repetitive structure in sounds1. *Speech Communication*, *27*, 187–207.

Khodai-Joopari, M., & Clermont, F. (2002). A Comparative study of empirical formulae for estimating vowel-formant bandwidths. In *Proceedings of the 9th Australian International Conference on Speech, Science, and Technology* (pp. 130–135). Sydney, NSW: Australian Speech Science and Technology Association.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.

Kreiman, J., Antoñanzas-Barroso, N., & Gerratt, B. R. (2010). Integrated software for analysis and synthesis of voice quality. *Behavior Research Methods*, *42*, 1030–1041.

Kreiman, J., Garellek, M., Chen, G., Alwan, A., & Gerratt, B. R. (2015). Perceptual evaluation of voice source models. *Journal of the Acoustical Society of America*, *138*, 1–10.

Lasarcyk, E., & Trouvain, J. (2007). Imitating conversational laughter with an articulatory speech synthesis. In J. Trouvain & N. Campbell (Eds.), *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter* (pp. 43–48). Retrieved from www.coli.uni-saarland.de/conf/laughter-07/files/PROCEEDINGS_COMPLETE.pdf

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2016). tuneR: Analysis of music. Retrieved from http://r-forge.r-project.org/projects/tuner/

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

Ling, Z. H., Kang, S. Y., Zen, H., Senior, A., Schuster, M., Qian, X. J., … Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine, 32,* 35–52.

Margoliash, D. (1983). Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience*, *3*, 1039–1057.

Moore, R. K. (2016). A real-time parametric general-purpose mammalian vocal synthesiser. In *INTERSPEECH* (pp. 2636–2640). Grenoble, France: International Speech Communication Association.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, *99*, 1877–1884.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F_0 of voice. *Phonetica*, *41*, 1–16.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*, 283–296.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., … Aucouturier, J.-J. (2018). DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods, 50,* 323–343. https://doi.org/10.3758/s13428-017-0873-y

Read, R., & Belpaeme, T. (2016). People interpret robotic non-linguistic utterances categorically. *International Journal of Social Robotics*, *8*, 31–50.

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W. T., & Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B*, *272*, 941–947.

Riede, T., Arcadi, A. C., & Owren, M. J. (2007). Nonlinear acoustics in the pant hoots of common chimpanzees (*Pan troglodytes*): Vocalizing at the edge. *Journal of the Acoustical Society of America*, *121*, 1758–1767.

Rosenberg, A. E. (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, *49*, 583–590.

Salvia, E., Bestelmeyer, P. E., Kotz, S. A., Rousselet, G. A., Pernet, C. R., Gross, J., & Belin, P. (2014). Single-subject analyses of magnetoencephalographic evoked responses to the acoustic

properties of affective non-verbal vocalizations. *Frontiers in Neuroscience*, *8*, 422. https://doi.org/10.3389/fnins.2014.00422

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

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

Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In J. Tao & T. Tan (Eds.), *Affective information processing* (pp. 111–126). London, UK: Springer.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001, September). *Acoustic correlates of emotion dimensions in view of speech synthesis*. Paper presented at the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark.

Shue, Y. L., & Alwan, A. (2010). A new voice source model based on high-speed imaging and its application to voice source estimation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5134–5137). Piscataway, NJ: IEEE Press.

Snowdon, C. T., & Pola, Y. V. (1978). Interspecific and intraspecific responses to synthesized pygmy marmoset vocalizations. *Animal Behaviour*, *26*, 192–206.

Stevens, K. N. (2000). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, *9*, 21–29.

Sueur, J. (2018). *Sound analysis and synthesis with R*. Heidelberg, Germany: Springer.

Sueur, J., Aubin T., Simonis C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*, *18,* 213–226.

Tappert, C. C., Martony, J., & Fant, G. (1963). Spectrum envelopes for synthetic vowels. *Speech Transmission Laboratory Quarterly Progress Status Report*, *4*, 2–6.

Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, *280*, 221–236.

Taylor, A. M., Reby, D., & McComb, K. (2008). Human listeners attend to size information in domestic dog growls. *Journal of the Acoustical Society of America*, *123*, 2903–2909.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, *101*, 1234–1252.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*. arXiv preprint. arXiv:1609.03499

Wilden, I., Herzel, H., Peters, G., & Tembrock, G. (1998). Subharmonics, biphonation, and deterministic chaos in mammal vocalization. *Bioacoustics*, *9*, 171–196.

Wolak, M. E., Fairbaim, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution*, *3*, 129–137.

Wood, A., Martin, J., & Niedenthal, P. (2017). Towards a social functional account of laughter: Acoustic features convey reward, affiliation, and dominance. *PLoS ONE*, *12*, e0183811.

Wu, Z., Watts, O., & King, S. (2016). Merlin: An open source neural network speech synthesis system. In *Proceedings of the 9th ISCA Speech Synthesis Workshop* (pp. 202–207). Grenoble, France: International Speech Communication Association. https://doi.org/10.21437/SSW.2016-33

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*, 1039–1064.

Paper V

BIOACOUSTICS
https://doi.org/10.1080/09524622.2019.1581839

Taylor & Francis
Taylor & Francis Group

The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations

Andrey Anikin

Division of Cognitive Science, Department of Philosophy, Lund University, Lund, Sweden

ABSTRACT

The communicative role of nonlinear vocal phenomena remains poorly understood since they are difficult to manipulate or even measure with conventional tools. In this study parametric voice synthesis was employed to add pitch jumps, subharmonics/sidebands, and chaos to synthetic human nonverbal vocalizations. In Experiment 1 (86 participants, 144 sounds), chaos was associated with lower valence, and subharmonics with higher dominance. Arousal ratings were not noticeably affected by any nonlinear effects, except for a marginal effect of subharmonics. These findings were extended in Experiment 2 (83 participants, 212 sounds) using ratings on discrete emotions. Listeners associated pitch jumps, subharmonics, and especially chaos with aversive states such as fear and pain. The effects of manipulations in both experiments were particularly strong for ambiguous vocalizations, such as moans and gasps, and could not be explained by a non-specific measure of spectral noise (harmonics-to-noise ratio) – that is, they would be missed by a conventional acoustic analysis. In conclusion, listeners interpret nonlinear vocal phenomena quite flexibly, depending on their type and the kind of vocalization in which they occur. These results showcase the utility of parametric voice synthesis and highlight the need for a more fine-grained analysis of voice quality in acoustic research.

ARTICLE HISTORY
Received 17 August 2018
Accepted 25 January 2019

KEYWORDS
Nonlinear phenomena; source spectrum; human nonverbal vocalizations; voice synthesis; emotion

Introduction

Nonlinear vocal phenomena, which can be roughly defined as various irregularities in the vibration of vocal folds (Riede et al. 2005), are common in the vocal repertoire of many avian (Fee et al. 1998) and mammalian species, including meerkats (Townsend and Manser 2011; Karp et al. 2014), whales (Tyson et al. 2007; Cazau et al. 2016), manatees (Mann et al. 2006), deer (Reby et al. 2005, 2016), canids (Wilden et al. 1998; Riede et al. 2000; Schneider and Anderson 2011), chimpanzees (Riede et al. 2007), and humans (Robb and Saxman 1988; Mende et al. 1990; Bachorowski et al. 2001; Facchini et al. 2005; Koutseff et al. 2018; Raine et al. 2018). Major advances have recently been made in providing mathematical (Mende et al. 1990; Herzel et al. 1995; Mergell and Herzel 1997; Tokuda et al. 2002; Herbst et al. 2013), physiological (Wilden et al. 1998; Titze 2008), and acoustic (Riede et al. 2000, 2005; Fitch et al. 2002) descriptions of these

CONTACT Andrey Anikin andrey.anikin@lucs.lu.se
Supplemental data for this article can be accessed here.

complex sounds, but their evolutionary significance and communicative roles are still debated (Fitch et al. 2002; Riede et al. 2007; Karp et al. 2014). To facilitate this research, an urgent task is to improve the experimental tools available for the measurement and manipulation of various nonlinearities in the voice.

I have recently presented *soundgen*, a novel tool for synthesizing nonverbal vocalizations (Anikin 2018), and argued that parametric voice synthesis could be a useful complement to more traditional correlational studies based on acoustic analysis of field recordings, as well as to interventional studies that modify recordings through sound editing. In this paper I report the results of two perceptual experiments in which *soundgen* was used to systematically manipulate various types of nonlinear vocal phenomena in synthetic human nonverbal vocalizations in order to explore their communicative significance.

Types of nonlinear vocal phenomena

During regular voiced phonation, the main oscillators – in most mammals, these are the left and right vocal folds – open and close in synchrony and at a relatively stable rate, describing a so-called limit cycle in the phase space (Wilden et al. 1998; Tokuda 2018). Phonation becomes less stable under certain conditions such as incomplete closure of the vocal folds with high pressure underneath them (subglottal pressure), or when fundamental frequency (f_0) approaches or crosses a formant – a frequency band amplified by the resonance of the vocal tract (Herzel et al. 1995; Wilden et al. 1998; Riede et al. 2000; Neubauer et al. 2004; Cazau et al. 2016; Tokuda 2018). Various acoustic irregularities can then occur, particularly if there is an underlying left-right asymmetry in the anatomical structure of oscillators (Herzel et al. 1995). The exact nomenclature of nonlinearities in acoustic signals varies across disciplines; in bioacoustics, it is common to distinguish pitch jumps, subharmonics, biphonation, and deterministic chaos (Wilden et al. 1998; Riede et al. 2000, 2007; Fitch et al. 2002; Mann et al. 2006; Tyson et al. 2007; Blumstein and Recapet 2009; Schneider and Anderson 2011; Tokuda 2018).

Pitch jumps, or frequency jumps, are abrupt discontinuities of f_0 associated with unstable phonation and described in the vocal repertoire of many animal species (Mann et al. 2006; Riede et al. 2007; Tyson et al. 2007; Schneider and Andersson 2011; Cazau et al. 2016). In humans, unwanted frequency jumps are an embarrassment to amateur singers to be avoided when switching between vocal registers (Wilden et al. 1998; Tokuda 2018), but they also occur in nonverbal vocalizations (Robb and Saxmann 1988). Overall, however, there are fewer reports and theoretical discussions of pitch jumps than other nonlinearities.

Subharmonics are observed when one vocal fold vibrates at exactly two or three times the frequency of the other (Riede et al. 2000; Fitch et al. 2002). The waveform remains periodic and describes a so-called 'folded limit cycle' in the phase space (Wilden et al. 1998). Subharmonics show on the spectrogram as one or more additional partials between harmonics of the fundamental frequency (Figure 1, right) and can be conceptualized as two harmonically related tones produced simultaneously. For example, period doubling ($g_0 = f_0/2$) sounds like two voices, one an octave lower than the other. As a result, subharmonics lower the apparent pitch (Fitch et al. 2002). In addition, the commonly

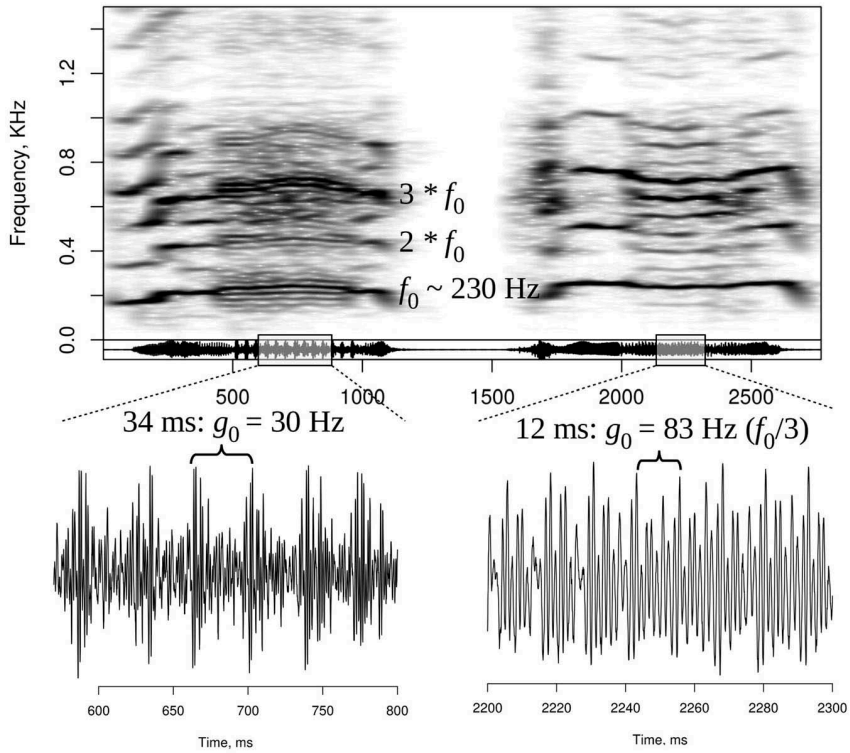

Figure 1. Subharmonics at $f_0/2$ (period doubling) followed by sidebands with slow amplitude modulation at a beat frequency of ~30 Hz (left) and subharmonics at $f_0/3$ (right). Spectrogram with a 150 ms Gaussian window and 90% overlap. Source: vocal demonstration by the author. AUDIO #1 in Supplements.

observed instability and irregularity of subharmonic regimes also creates an impression of vocal hoarseness or 'roughness' (Herzel et al. 1995; Fastl and Zwicker 2006).

Biphonation refers to a situation in which the left and right vocal folds or their parts vibrate at two different frequencies, f_0 and g_0. Biphonation in the narrow sense requires that f_0 and g_0 should not be harmonically related (i.e. f_0:g_0 should not be a rational number like 2:1, 4:3, etc.). The resulting waveform is not periodic, and the oscillating system describes a torus in the phase space (Wilden et al. 1998). If f_0 and g_0 vary independently, the spectrogram shows two nonparallel fundamental frequencies, which literally sounds like two voices. This is common in whale songs (Tyson et al. 2007) and can occasionally be seen in humans (Neubauer et al. 2004). Due to acoustic coupling, however, oscillators tend to partly synchronize (Fitch et al. 2002), so that f_0 and g_0 often vary in parallel or form integer ratios (Herzel et al. 1995). From this point of view, subharmonics can be regarded as a special case of biphonation with strong acoustic coupling and f_0 an integer multiple of g_0.

Another consequence of acoustic coupling in biphonation is that the spectrum contains not only *f*- and *g*-harmonics, but also their linear combinations $nf_0 + mg_0$, where n and m are integers (Wilden et al. 1998; Riede et al. 2000). If g_0 is low relative to f_0, the resulting spectrogram shows *sidebands* (Figure 1, left) – additional partials that

appear at $\pm g_0$, $\pm 2g_0$ and so on around each *f*-harmonic (Wilden et al. 1998; Reby et al. 2016). Looking at the waveform, the lower frequency g_0 shows as amplitude modulation of the carrier frequency f_0. The modulation frequency can be lower than the rate at which either vocal fold is vibrating (Mergell and Herzel 1997). Frequency modulation (FM), such as the vibrato found in opera singing, can also be seen as a form of biphonation (Wilden et al. 1998), and rapid FM produces sidebands around *f*-harmonics (Sueur 2018). Perceptually, sidebands make the voice rough and in this sense resemble subharmonics (Herzel et al. 1995; Riede et al. 2000; Audio #1 in Supplements).

The terminology used to describe biphonation-related phenomena is somewhat inconsistent, and their acoustic complexity can be further enhanced by the recruitment of additional oscillators, such as air sacs and extensions of the vocal folds known as 'vocal lips' (Fitch et al. 2002), as well as by combining regular phonation with glottal or nasal whistles (Neubauer et al. 2004; Reby et al. 2016). The present study only deals with relatively simple cases of period doubling/tripling and sidebands. In both cases a lower secondary frequency g_0 is added to the signal, and in the rest of the paper I refer to these manipulations simply as 'subharmonics'.

In addition to the limit cycle (normal phonation), folded limit cycle (subharmonics), and torus (biphonation), vocal folds can vibrate in a chaotic regime that mathematically corresponds to a strange attractor in the phase space (Wilden et al. 1998; Fitch et al. 2002; Herbst et al. 2013). *Deterministic chaos* looks superficially similar to broadband turbulent noise on the spectrogram (Figure 2) and also possesses a rough vocal quality, but it has residual harmonic structure. As a result, a roar is perceptually distinct from hissing even if the sound pressure and spectral envelopes of both recordings are comparable. Several nonlinear regimes are often found within one vocalization (Figure 2), and transitions between them are known as *bifurcations*. There are certain regularities in the typical sequence of bifurcations; for example, chaos is often preceded by subharmonics or biphonation (Wilden et al. 1998; Fitch et al. 2002).

Measuring and synthesizing nonlinear phenomena

The traditional – and still popular and useful – approach to characterizing vocal nonlinearities is based on a visual inspection of the spectrogram. Among quantitative indices, harmonics-to-noise ratio (HNR) is a straightforward measure of harmonicity or tonality (Boersma 1993) and a good predictor of human ratings of noisiness (Riede et al. 2005), but it requires accurate pitch tracking, which is problematic for noisy sounds. Crucially, HNR fails to discriminate between spectral noise caused by irregular phonation (e.g. chaos) and turbulence (e.g. aspiration noise) in voiced calls. In fact, vocalizations with more nonlinearities may have a higher, not lower, harmonicity. For example, Raine et al. (2018) report the seemingly paradoxical observation that the intensity of pain correlated with both the presence of nonlinearities and higher HNR. The likely explanation is that mild moans had a breathy voice quality, with weak harmonics and low HNR; screams of intense pain, on the other hand, were delivered in a bright voice with strong harmonics, making HNR higher despite episodes of chaos. In other words, HNR captures the overall level of spectral noise, but it is not a specific measure of nonlinear vocal phenomena.

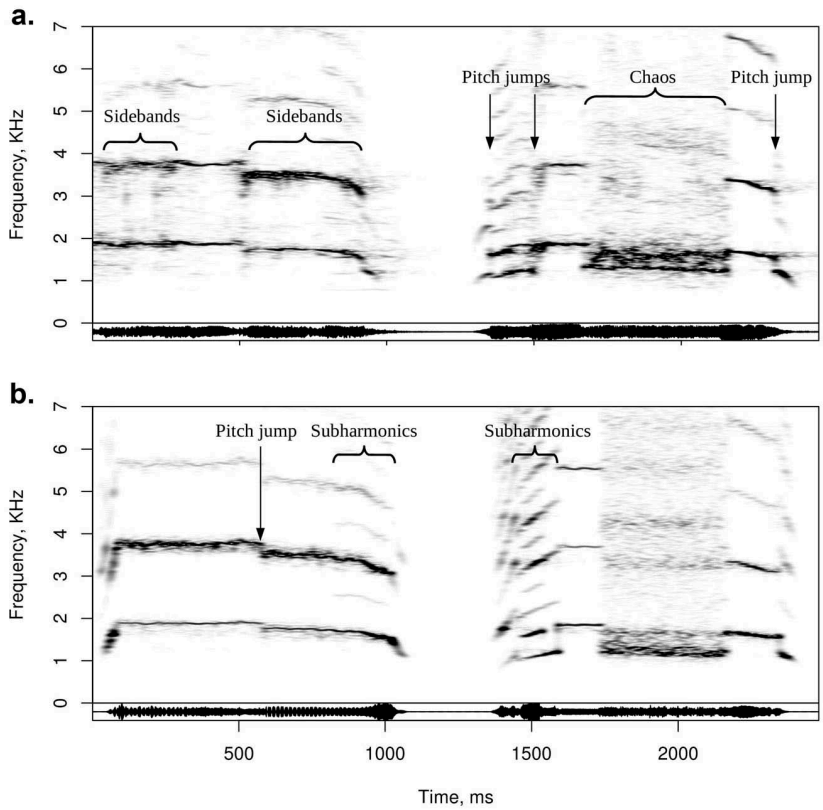

Figure 2. Various nonlinearities in recorded (A) and synthetic (B) versions of two human screams. Spectrogram with a 50 ms Gaussian window and 70% overlap. Source: an adult female at a haunted house attraction (file 213_ut_fear_29-f-scream.wav from Anikin and Persson 2017). AUDIO #2a and #2b in Supplements.

Unlike HNR, other measures describe specifically the voiced component, namely the cycle-to-cycle variability in frequency (jitter) and amplitude (shimmer), so in principle they can distinguish between chaos and a breathy voice. However, their accurate measurement depends on being able to detect individual glottal cycles, which in practice is only feasible in sustained vowels (Goudbeek and Scherer 2010). Specialized algorithms borrowed from nonlinear dynamics have also been applied to biological sounds (Tokuda et al. 2002; Tyson et al. 2007; Cazau et al. 2016), but these tools require advanced mathematical expertise and may not be superior to HNR at capturing perceptually relevant properties (Riede et al. 2005).

Manipulating or synthesizing nonlinear phenomena is even more problematic than measuring them. Pitch jumps are the easiest to synthesize, since they only require a rapid change in f_0. Subharmonics can be incorporated in models of individual glottal cycles (Alonso et al. 2015). Both subharmonics and sidebands can also be synthesized by adding rapid amplitude and/or frequency modulation to the signal (Sueur 2018; http://forumnet.ircam.fr/product/angus/). This is straightforward to do in *soundgen*, but a more flexible approach used in this study was to directly create extra partials around each *f*-harmonic, while forcing g_0 and f_0 to be at an integer ratio (Anikin 2018).

Chaos is the most challenging nonlinearity to synthesize. Biomechanical two-mass models of vocal folds (Herzel et al. 1995; Cazau et al. 2016) and even physical models of the vocal tract (Tokuda 2018) have been used to simulate chaotic behaviour and to investigate under what conditions chaos is likely to occur. However, by definition it is difficult to control the exact behaviour of the system once it has switched to a chaotic mode, so these models are less suitable for synthesizing sounds with the desired duration, strength, and dynamics of chaos as well as preserved traces of harmonics with the appropriate f_0 contour. On the other hand, parametric speech synthesis sometimes incorporates stochastic cycle-to-cycle variability in the amplitude (shimmer) and period (jitter) of glottal cycles (Alonso et al. 2015), which is also the principle behind the generation of 'chaos' in *soundgen*. It is worth emphasizing that, although *soundgen* technically uses jitter and shimmer to approximate chaos, they are orders of magnitude stronger than in ordinary phonation, and the result appears to be perceptually similar to deterministic chaos (Anikin 2018). This approach makes it possible to control both the timing of each episode and the strength of 'chaos' – that is, the amount of residual energy in harmonics.

The role of nonlinear phenomena in communication

The goal of measuring and manipulating nonlinear vocal phenomena is to learn more about how they contribute to acoustic communication. Of course, nonlinearities might simply be incidental to vocal production, but there are both theoretical and empirical reasons to believe that they provide listeners with useful information (Fitch et al. 2002; Riede et al. 2007; Karp et al. 2014). Unfortunately, while many studies report correlations between behaviour and some measure of spectral noise, such as HNR, it is not always clear how this relates to the presence of specific nonlinear phenomena. Casting the net wide to include previous observations on spectral noise of any kind in both animal vocalizations and human speech, nonlinear vocal phenomena appear to be associated with:

(1) *High arousal*. Rough and noisy vocalizations are produced in high-stakes contexts suggestive of high arousal understood as the level of general alertness, motivation, or intensity of emotional states (Fitch et al. 2002; Schneider and Anderson 2011; Briefer 2012). Noisy sounds may both serve as honest signals of urgency and advertise the caller's fitness (Fitch et al. 2002; Riede et al. 2007). This is corroborated by biomechanical modelling, which suggests that nonlinearities occur when f_0 and subglottal pressure are high, which is in turn indicative of an aroused physiological state (Herzel et al. 1995; Cazau et al. 2016). Nonlinearities also make a signal unpredictable, which prevents habituation in listeners (Blumstein and Recapet 2009; Townsend and Manser 2011; Karp et al. 2014). In humans, the perceived intensity of pain or distress correlates with the presence of nonlinearities in the cries of infants (Facchini et al. 2005; Koutseff et al. 2018) and adult actors (Raine et al. 2018), although it is not clear whether the effect is driven by high arousal, negative valence, or both.
(2) *High dominance*. Low pitch is generally associated with physical (Morton 1977) and social (Puts et al. 2006, 2007) dominance. Subharmonics, and possibly also chaos, lower the perceived pitch, exaggerating the apparent body size, and this

makes them suitable for displaying a dominant or aggressive attitude (Morton 1977; Ohala 1984; Fitch et al. 2002; Siebert and Parr 2003). Subharmonics and especially chaos also introduce spectral noise, which makes formant frequencies easier to detect (Fitch et al. 2002). Since formant dispersion is indicative of vocal tract length, nonlinearities help to advertize body size, as in roaring contests of male deer (Reby et al. 2005). In humans, HNR of a single word *Hello* affected the perceived social dominance of the speaker (McAleer et al. 2014).

(3) *Negative valence*. Compared to arousal, there is much less consensus about acoustic correlates of the positive or negative valence of emotional states (Briefer 2012). Spectral noise and nonlinearities have been linked to unpleasant contexts (Fichtel et al. 2001) and aggression (Morton 1977; August and Anderson 1987), although in the latter case it is not clear whether they express negative valence or dominance. In humans, more tonal laughs have been reported to be less positive (Lavan et al. 2016), although this is more likely to be related to the balance between voiced and unvoiced fragments than to nonlinear phenomena. As noted above, nonlinearities in vocalizations of human infants are associated with intense distress (Facchini et al. 2005; Koutseff et al. 2018), and rapid amplitude modulation in human screams is associated with greater perceived fear intensity (Arnal et al. 2015). The literature is thus consistent with at least two possibilities: either nonlinearities amplify the intensity of any expressed emotion (i.e. making laughs more positive and screams or cries more negative), or they make any vocalization more negative in valence.

The present study

This study was designed to exploit *soundgen*'s ability to add a controlled amount of specific nonlinear phenomena to synthetic vocalizations, offering an opportunity to test their perceptual effects directly, without acoustic confounds. The chosen material was human nonverbal vocalizations. Although nonlinear vocal phenomena were traditionally regarded as being of peripheral importance to speech and associated with voice pathology (Robb and Saxman 1988; Fitch et al. 2002), it is now established that nonlinearities are common in non-pathological voices. They are particularly well documented in infants (Robb and Saxman 1988; Mende et al. 1990; Facchini et al. 2005; Koutseff et al. 2018), but adult voices appear to be no different (Raine et al. 2018). In fact, various nonlinearities were found in about half of spontaneous human vocalizations collected by Anikin and Persson (2017). Since the stimuli in this study were synthetic replicas of these recordings, participants were presented with nonlinear vocal phenomena in a natural and familiar acoustic context.

Based on the evidence presented in the previous section, I hypothesized that listeners would associate nonlinear vocal phenomena with the speaker being alert (high arousal), assertive (high dominance), and experiencing something unpleasant (low valence). In order to test these hypotheses, listeners in a perceptual experiment rated the manipulated vocalizations on these three perceptual dimensions. Valence and arousal are standard dimensions used to describe the emotional component in both animal (Briefer 2012) and human (Belin et al. 2008; Lima et al. 2013) vocalizations. Dominance is less well established in emotion research, with different authors favouring

various dimensions such as control, power, potency, and so on (Fontaine et al. 2007; Goudbeek and Scherer 2010). In this study participants were asked whether the speaker sounded aggressively self-confident or submissive as a measure of perceived social dominance (Puts et al. 2006, 2007).

Experiment 1

Methods

Stimuli

The stimuli consisted of 144 modifications of 28 human nonverbal vocalizations, four prototypes from each of the following seven call types: cry, gasp, grunt, laugh, moan, roar, and scream. All stimuli were synthetic reproductions of the original recordings (Anikin and Persson 2017) created with *soundgen* 1.1.2 (Anikin 2018). The 28 prototype vocalizations were all from different individuals: 17 women and 11 men. These particular sounds were chosen based on the relatively high authenticity ratings of their synthetic versions in a previous study that validated *soundgen* (Anikin 2018). The duration of longer stimuli was reduced, limiting the range of duration to 0.22–2.7 s.

Each of 28 prototypes was synthesized in three to six versions, which differed only in the type and intensity of nonlinear effects (see Supplements). Nonlinear phenomena consisted of pitch jumps, subharmonics, chaos, and their combination at two intensities: low in 'mixed 1' and high in 'mixed 2' (Figure S1). When a particular effect could not be applied or the result sounded too contrived, this variation was skipped: for example, pitch jumps don't normally occur in gasps or short grunts.

Procedure

All data was collected online. Participants were told that they would hear synthetic versions of previously recorded human vocalizations. To minimize the correlation between scales, the experiment was divided into three blocks, one per scale (valence, arousal, and dominance). The order of blocks and sounds was randomized. Participants thus rated each sound three times, indicating how the speaker felt according to the following definitions:

> ***Valence*** *is high if the experience is pleasant, so the speaker is happy, pleased, relieved, etc. Valence is low if the experience is unpleasant, so the speaker is sad, afraid, in pain, etc.*
>
> ***Arousal*** *is high if the person is very energetic, alert, wide-awake. Arousal is low if the person is sleepy, relaxed, calm.*
>
> ***Dominance*** *is high if the speaker sounds assertive, self-confident, superior, perhaps aggressively so. Dominance is low if the person sounds submissive, uncertain, perhaps as someone who seeks reassurance or needs a hug.*

Prior to each block the upcoming scale was illustrated with two examples, one high and the other low on the upcoming scale. To avoid presenting very similar sounds repeatedly, the stimuli were divided into four batches rated by different samples of participants. As a result, each participant heard maximum two manipulated versions of the same prototype. The average completion time was about 9 minutes.

Participants

It is particularly important to ensure data quality in the context of online experiments. All submissions were therefore manually screened for obvious cheating. In addition, the responses of 11 participants poorly correlated with the global median ratings ($r < 0.3$ on any two scales or $r < 0$ on any one scale), presumably indicating that they had not attended to the task, so they were excluded from the analysis. A separate sensitivity analysis was performed to make sure that the exclusion of 11 out of 97 participants with noisy data did not affect the main conclusions from the study (Table S2). The final sample consisted of 86 participants, of whom 18 were unpaid volunteers contacted via online advertisements and 68 were recruited from https://www.prolific.ac for £1.50. Each sound was rated on average 20.4 times (range 17.3 to 24.7) on each scale. No demographic characteristics were collected; according to the statistics on https://www.prolific.ac/demographics, over 80% of participants are native English speakers, and about 75% are between 20 and 40 years of age.

Statistical analysis

Unaggregated, trial-level responses were analyzed using mixed models with random intercepts per participant, per stimulus, and per prototype (shared by all sounds that were modifications of the same original vocalization). Valence, arousal, and dominance ratings were modelled using the beta distribution. Mixed models were fit in Stan computational framework (http://mc-stan.org/) accessed with R package *brms* (Bürkner 2017). To improve convergence and guard against overfitting, regularizing priors were used for all regression coefficients. In models with multiple predictors or many levels of categorical predictors (such as call types), more strongly regularizing horseshoe priors (Carvalho et al. 2009) were employed to provide so-called 'shrinkage' of regression coefficients towards zero, thus providing an in-built correction for multiple comparisons. The reported estimates are thus slightly more conservative than they would be with non-Bayesian models.

Exploratory analysis of the effects of the strength and duration of nonlinear phenomena was performed with non-Bayesian mixed models, using likelihood ratio test to determine the significance of effects.

The stimuli, R code for their generation, experimental datasets, and scripts for statistical analysis can be downloaded from http://cogsci.se/publications.html.

Results

Valence and arousal ratings aggregated per stimulus were related quadratically ($F(2,141) = 20.0$, $p < .001$; $R^2 = .22$): sounds with both very positive and very negative valence ratings tended to be high on arousal. There was also a positive linear relationship between arousal and dominance ($F(1,142) = 29.9$, $p < .001$, $R^2 = .17$) and between valence and dominance ($F(1,142) = 12.4$, $p < .001$, $R^2 = .08$; Figure S3). Inter-rater reliability was moderate for valence ($ICC = .45$) and arousal ($ICC = .56$) and low for dominance ($ICC = .21$). Similar levels of inter-rater reliability were observed for valence and arousal ratings in a validation study with actual unmodified recordings of the same vocalizations (Anikin 2018). The imperfect agreement among raters is thus likely to be related to the intrinsic ambiguity of these vocalizations rather than to their synthetic

nature, but the dominance scale was clearly associated with less consistent responses than the valence and arousal scales.

Adding any nonlinear effects reduced the perceived valence of a synthetic vocalization by 0.18 points on a scale of −1 to +1 (95% CI [0.11, 0.24]) compared to having no nonlinear effects (Table 1). The negative effect of nonlinearities on perceived valence was weakest for subharmonics (−0.11 [−0.19, −0.02]) and strongest for mixed nonlinear effects (−0.23 [−0.3, −0.15]).

Harmonics-to-noise ratio (HNR) progressively decreased after adding subharmonics, chaos, and their combination at low ('mixed 1') and high ('mixed 2') levels of intensity, but not pitch jumps (Figure S2). As expected, noisier vocalizations were perceived as more aversive: as HNR dropped by 10 dB, the perceived valence was predicted to become 0.14 points lower [0.05, 0.24]. However, the negative effects of nonlinear phenomena on valence survived controlling for HNR (Table 1), except for the effect of subharmonics (−0.05 [−0.14, 0.04]). This suggests that chaos and pitch jumps made the vocalization appear more negative even after accounting for the change in HNR, while the negative effect of subharmonics on valence was primarily mediated by the accompanying drop in HNR.

Looking at individual call types, mixed nonlinear effects lowered the valence of all vocalizations except grunts (Table S1). The negative effect of pitch jumps on valence was driven exclusively by screams, which became 0.25 [0.10, 0.39] more negative after the addition of pitch jumps. Subharmonics lowered the perceived valence of moans by 0.28 [0.07, 0.46], but they had little or no effect on the valence of other vocalizations. Chaos had a pronounced negative effect on the valence of cries, roars, screams, and perhaps laughs, but not grunts and moans (Table S1). In other words, the extent to which pitch jumps, subharmonics, and chaos affected valence depended on the call type, but in all cases this effect was negative.

Nonlinear effects had no effect on arousal ratings either overall (+0.02 [−0.05, 0.08], Table 1) or for any call type (Table S1). The only marginal effect was a tendency for subharmonics to increase the perceived arousal by 0.08 [0.00, 0.16], which survived controlling for HNR (0.10 [0.00, 0.20]). HNR had no effect on arousal either by itself (0.02 [−0.04, 0.09]) or after controlling for the type of added nonlinear effects (0.06 [−0.05, 0.17]).

Table 1. The effect of adding nonlinear phenomena.

		Predicted difference: median [95% CI]*		
Model	Nonlinear phenomena	Valence	Arousal	Dominance
Without HNR	Any vs. none	**−0.18 [−0.24, −0.11]**	0.02 [−0.05, 0.08]	**0.05 [0.00, 0.11]**
	Pitch jumps vs. none	**−0.14 [−0.26, −0.03]**	−0.03 [−0.14, 0.08]	−0.02 [−0.11, 0.07]
	Subharmonics vs. none	**−0.11 [−0.19, −0.02]**	**0.08 [0.00, 0.16]**	**0.09 [0.01, 0.16]**
	Chaos vs. none	**−0.18 [−0.27, −0.09]**	0.05 [−0.04, 0.14]	0.03 [−0.05, 0.10]
	Mixed vs. none	**−0.23 [−0.30, −0.15]**	−0.01 [−0.08, 0.07]	**0.08 [0.02, 0.14]**
	Mixed level 2 vs. 1	−0.05 [−0.13, 0.03]	−0.01 [−0.10, 0.07]	0.04 [−0.03, 0.11]
	Chaos vs. subharmonics	−0.07 [−0.16, 0.02]	−0.02 [−0.12, 0.06]	−0.06 [−0.14, 0.02]
Controlling for HNR	Any vs. none	**−0.10 [−0.19, −0.03]**	0.04 [−0.04, 0.13]	−0.01 [−0.08, 0.06]
	Pitch jumps vs. none	**−0.14 [−0.24, −0.05]**	−0.04 [−0.15, 0.07]	−0.01 [−0.10, 0.08]
	Subharmonics vs. none	−0.05 [−0.14, 0.04]	**0.10 [0.00, 0.20]**	0.04 [−0.03, 0.12]
	Chaos vs. none	**−0.09 [−0.20, 0.00]**	0.09 [−0.02, 0.19]	−0.04 [−0.13, 0.04]
	Mixed vs. none	**−0.11 [−0.22, −0.01]**	0.04 [−0.07, 0.15]	−0.02 [−0.10, 0.07]
	Mixed level 2 vs. 1	−0.02 [−0.10, 0.06]	0.00 [−0.09, 0.08]	0.01 [−0.07, 0.08]
	Chaos vs. subharmonics	−0.04 [−0.13, 0.04]	−0.01 [−0.10, 0.08]	**−0.09 [−0.16, −0.01]**

* Cells in **bold** contain 95% CIs that exclude or mostly exclude zero. This is merely a visualization aid, not significance testing.

Dominance ratings became slightly higher after the addition of nonlinear effects (0.05 [0.00, 0.11]), particularly subharmonics (0.09 [0.01, 0.16]) or a mixture of subharmonics with pitch jumps and chaos (0.08 [0.02, 0.14]). However, neither chaos alone (0.03 [−0.05, 0.10]) nor pitch jumps alone (−0.02 [−0.11, 0.07]) had any effect on the perceived dominance, suggesting that the effect of nonlinear effects on dominance was driven specifically by subharmonics. On the other hand, this increase in dominance ratings was not strong enough to be significant when analyzing individual call types (Figure 3, Table S1) and disappeared after controlling for HNR (Table 1). A 10 dB drop in HNR caused dominance ratings to increase by 0.14 [0.05, 0.22]. Interestingly, the difference between the effects of subharmonics and chaos on dominance ratings became slightly more robust after controlling for HNR (−0.09 [−0.16, −0.01]). These findings suggest that rough voices were perceived as more dominant, but this was primarily due to subharmonics rather than chaos or some other type of spectral noise.

It is possible that not only the type of nonlinear phenomena but also their duration and strength might be salient to listeners. Valence ratings were affected by the duration of subharmonics (likelihood ratio test: $L = 4.5$, $df = 1$, $p = .03$) and arousal ratings by their strength ($L = 4.7$, $df = 1$, $p = .03$). Valence was affected by the duration ($L = 11.8$, $df = 1$, $p < .001$) but not strength of chaos, and arousal and dominance by neither. This exploratory analysis suggests that the effects of nonlinearities on valence mostly depended on the length of the affected vocal fragment, particularly in the case of chaos, although more stimuli would be needed to verify this finding.

Discussion

The aim of Experiment 1 was to evaluate the perceptual consequences of manipulating nonlinear vocal phenomena in several types of human nonverbal vocalizations. The findings are best seen as exploratory, but two key observations merit further investigation. First, although all nonlinearities made vocalizations more negative in valence, the effect of chaos was particularly strong; in contrast, only subharmonics increased the perceived arousal and dominance of the speaker. This suggests that different types of nonlinear phenomena may have distinct effects on listeners. Second, a commonly reported measure of harmonicity and spectral noise (HNR) did not fully account for the observed perceptual effects of nonlinear phenomena. From a listener's perspective, the experimental acoustic manipulations thus appeared to be both salient and highly specific. A second experiment was designed to verify and extend these findings.

Experiment 2

Experiment 1 had two main limitations. First, there were only four prototype stimuli per call type, making it difficult to determine whether the perceptual effects of nonlinear phenomena were similar for different types of vocalizations. Second, although dominance is an interesting dimension theoretically, participants did not find this scale very intuitive and provided less consistent ratings compared to valence and arousal. In the follow-up experiment it was therefore decided to make the following changes:

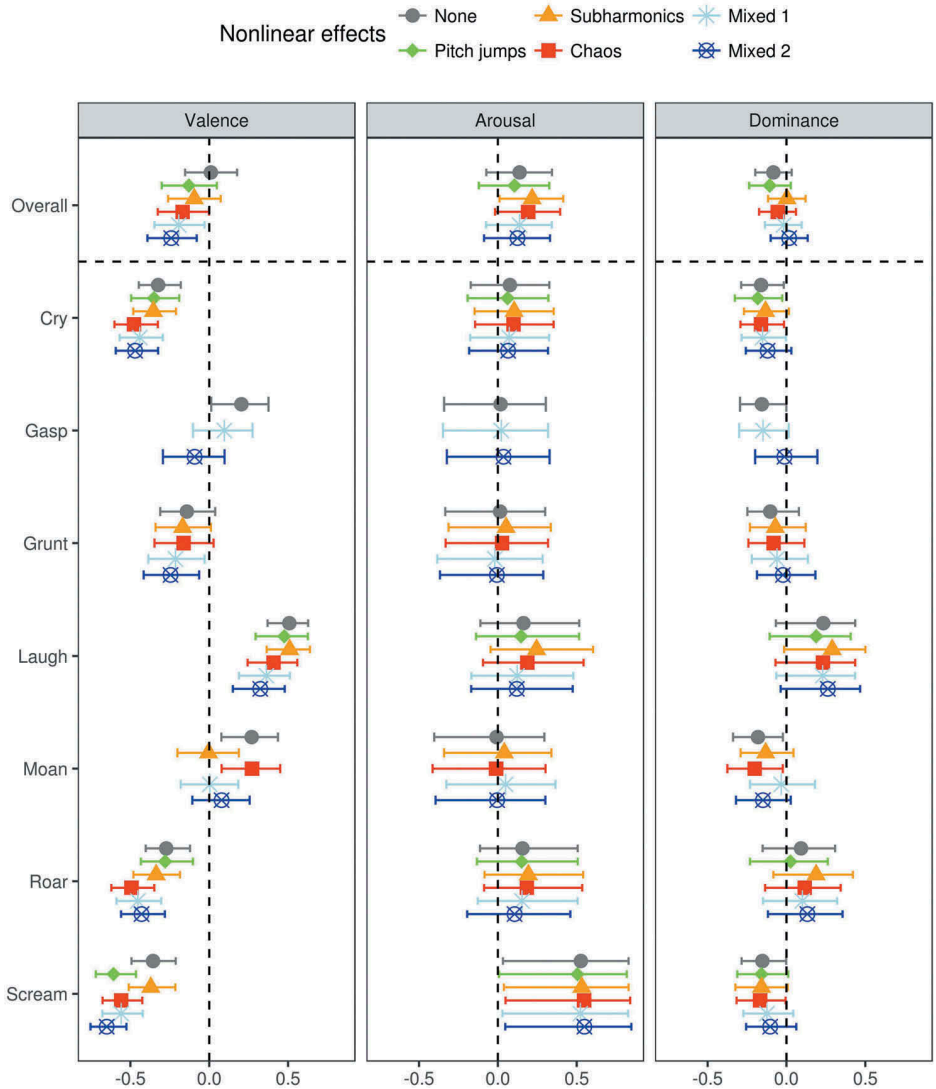

Figure 3. The effect of nonlinear phenomena on the ratings of valence, arousal, and dominance (range −1 to + 1). Median of posterior distribution and 95% CI. Mixed 1, mixed 2 = mixed nonlinear effects, low/high intensity.

(1) The number of call types was reduced to four, of which one was predominantly positive in valence (laughs), one negative (screams), and two more ambivalent (gasps and moans). In contrast, the number of stimuli per call type was increased.
(2) The outcome measure was the perceived intensity of several emotions instead of valence, arousal, and dominance scales. Three emotional labels were chosen for each call type: *Pleased/Hurt/Surprised* for gasps, *Amused/Evil (jeering)/Polite* for laughs (adapted from Szameitat et al. 2009; Wood et al. 2017), *Pleased/Hurt/Effortful* for moans, and *Pleased/Afraid/Aggressive* for screams. These labels do not cover the entire range of possible interpretations, but they correspond to the most common

classifications of acoustically similar vocalizations in earlier studies (Anikin 2018; Anikin et al. 2018). Moreover, the objective in Experiment 2 was not to identify the full range of emotions associated with each sound, but only to contrast the responses to the same stimulus after various acoustic manipulations.

(3) Experiment 1 suggested that the effects of different types of nonlinear phenomena were not identical. To clarify differences between pitch jumps, subharmonics, and chaos, these effects were always added separately, without mixing them.

Methods

Stimuli

The experimental stimuli were 212 synthetic replicas and modifications of 56 prototypes (Anikin and Persson 2017), including 10 gasps, 14 laughs, 15 moans or grunts, and 17 screams or high-pitched roars with duration ranging from 0.4 to 3.4 s (mean 1.25 s). Out of 56 prototypes, 30 were by women, 23 by men, and 3 by preadolescence children; 15 were also used in Experiment 1. Each prototype was replicated as faithfully as possible with *soundgen* 1.2.0 (Anikin 2018) and then re-synthesized in 3 to 5 modified versions (Table 2).

Procedure

The experiment was performed in a web browser and began with training, in which participants rated eight real recordings of human non-linguistic vocalizations: two laughs, two gasps, two screams, and two moans. This training was intended to familiarize the participants with the rating tool and stimuli and was followed by the main experiment; the average completion time was 15 minutes. The rating tool was a novel triadic scale designed specifically for this experiment, which generated ratings on three categories from a single click. As shown in Figure 4, three labels were placed in the corners of an equilateral triangle, and the weights of these three categories were related to the position of the marker within the triangle via a nonlinear transformation under the constraint that the three weights should sum to 100% (Table S3). The placement of emotion labels in the three corners was chosen randomly for each participant. Participants could see the bars indicating the weight of each category under the triangle, and pre-testing confirmed that this scale was intuitive to use. The code for running the triadic rating scale in html/javascript is available in the Supplements.

Table 2. Experimental manipulations of synthesized stimuli (N = 212, based on 56 prototypes).

Condition	Nonlinear effects	Number of stimuli (gasps/laughs/moans/screams)
O	None	56 (10/14/15/17)
Sb	Subharmonics	87* (16/21/17/33)
Ch	Chaos	56 (10/14/15/17)
PJ	Pitch jumps	13** (0/0/0/13)

* Only a single value of g_0 was synthesized for 25 out of 56 prototypes (with lower pitch), while for 31 out of 56 prototypes (with higher pitch) two different values of g_0 were synthesized: 31 * 2 + 25 = 87 stimuli with subharmonics.

** Only screams or roars.

Participants

Participants were recruited via https://www.prolific.ac and paid £1.50. To minimize exposure to multiple modified versions of the same prototype sound, the stimuli were divided into four batches and rated by four independent samples of participants, so that each person heard no more than 2 versions of the same prototype. Criteria for excluding participants were: (1) correlation with global median ratings <0.3 after averaging across all labels and call types or (2) correlation with global median <0 for any call type. This identified 19 participants, who were removed from further analysis. As in Experiment 1, a sensitivity analysis confirmed that the exclusion of 19 out of 102 participants with noisy data did not affect the main conclusions from the study (Table S5). Trials with response time under 2 s (0.7% of data) were also excluded. Data from the final sample of 83 participants provided on average 20 (range 19 to 22) responses per sound.

Statistical analysis

The triadic rating scale returns a vector of three weights that sum to one – a so-called simplex. It can be modelled with the Dirichlet distribution, but a more flexible approach is to use a redundantly-parameterized normal distribution that forces the means for the three categories to sum to one with a softmax transform (Gelman et al. 1996):

$$\mu_i = exp(\varphi_i)/(exp(\varphi_1) + exp(\varphi_2) + exp(\varphi_3))$$

for i in $\{1, 2, 3\}$, where μ_i is the mean of the normal distribution for the weight of each category and φ_i is normally distributed with a mean of zero. The φ parameters are not uniquely identifiable, making this parametrization redundant, but they do provide valid inference on μ. A corresponding Bayesian model was defined in Stan and extended to include a main effect of condition and two random intercepts (per participant and per prototype sound) with regularizing priors. A separate model was fit for each of four call types, since they had different outcome categories.

Results

The triadic rating scale used in this experiment is a novel tool, and prior to analyzing the results it was ascertained that participants interacted with it as intended (Figure S4). The marker's position inside the triangle (Figure 4) determined the weights of three response categories, which were indicated by the labels at each vertex. Since it was the weights that participants were asked to set, the main analysis focused on how acoustic manipulations affected the weights of different emotion categories, not the marker's coordinates. Figure 4 and Table 3 present the predictions of the model without considering harmonics-to-noise ratio (HNR). Similarly to Experiment 1, the same models were also re-built with HNR as a covariate (Table S4).

Gasps

As expected, both subharmonics and chaos increased the weight of the *Hurt* category, but the effect of chaos (+23.1%, 95% CI [17.0, 29.3] was more pronounced than the effect of subharmonics (+7.2% [2.0, 12.7]). The weight of the *Pleased* category dropped after the addition of nonlinear effects, and again chaos (−13.8% [−20.1, −7.8]) was more

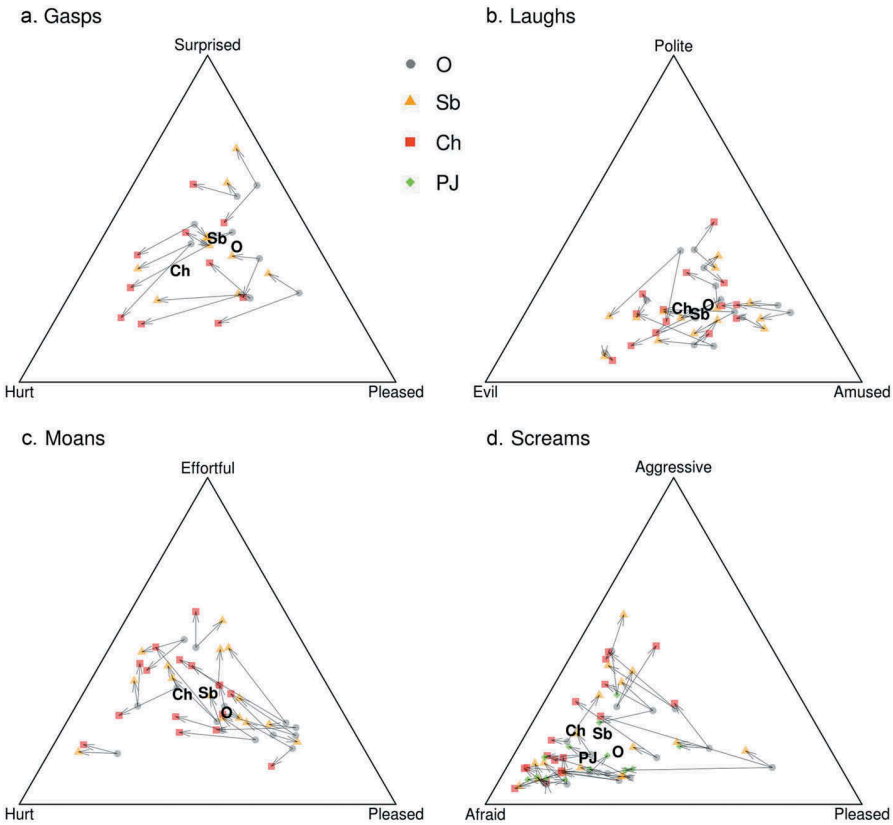

Figure 4. Mean coordinates representing the perceived emotion of different call types and acoustic manipulations. Labels in bold show the average for all sounds, while individual symbols and arrows show the effect of manipulations for each prototype sound. O = original, Sb = subharmonics, Ch = chaos, PJ = pitch jumps.

effective than subharmonics (−7.1% [−12.4, −1.6]). Chaos, but not subharmonics, also decreased the weight of the *Surprised* category (−9.2% [−17.8, −0.7]). Accounting for HNR did not substantively alter the observed effects of nonlinear phenomena on the perceived emotion (Table S4).

Laughs

The tested acoustic manipulations had little effect on the ratings of laughs. Chaos slightly shifted the interpretation of laughs from *Amused* (−7.7% [−13.1, −2.2]) to *Evil* (9.1% [1.7, 16.8]), but these effects disappeared after controlling for HNR (Table S4). It is therefore likely that the tendency to associate rough laughs with jeering was caused by the overall level of spectral noise rather than specifically by the presence of subharmonics or chaos.

Moans

In contrast to the weak findings for laughs, the effect of experimental manipulations on moans was very pronounced (Figure 4, Table 3). Chaos shifted the interpretation of moans from *Pleased* (−16.1% [−24.1, −8.5]) to *Hurt* (+10.5% [4.9, 15.9] or *Effortful*

Table 3. Contrasts between the effect of different acoustic manipulations on the weight of emotion categories for each call type: median of posterior distribution (%) and 95% CI.

	Gasps				Laughs		
Contrast	Hurt	Surprised	Pleased	Contrast	Evil	Polite	Amused
Sb vs. O*	**7.2 [2.0, 12.7]****	−0.1 [−7.9, 7.3]	**−7.1 [−12.4, −1.6]**	Sb vs. O	5.5 [−1.3, 12.3]	−2.8 [−7.7, 2.3]	−2.7 [−7.5, 2.2]
Ch vs. O	**23.1 [17.0, 29.3]**	**−9.2 [−17.8, −0.7]**	**−13.8 [−20.1, −7.8]**	Ch vs. O	**9.1 [1.7, 16.8]**	−1.4 [−6.9, 4.1]	**−7.7 [−13.1, −2.2]**
Ch vs. Sb	**15.8 [10.4, 21.6]**	**−9.1 [−17.1, −1.2]**	**−6.7 [−12.5, −1.1]**	Ch vs. Sb	3.6 [−3.3, 10.3]	1.4 [−3.6, 6.2]	**−5.0 [−9.8, −0.2]**
	Moans				Screams		
Contrast	Hurt	Effortful	Pleased	Contrast	Afraid	Aggressive	Pleased
Sb vs. O	2.9 [−2.0, 7.8]	**7.0 [2.0, 12.1]**	**−9.8 [−17.1, −3.1]**	Sb vs. O	2.0 [−1.7, 5.9]	**6.0 [2.3, 9.8]**	**−8.0 [−13.3, −2.7]**
Ch vs. O	**10.5 [4.9, 15.9]**	**5.6 [0.2, 11.0]**	**−16.1 [−24.1, −8.5]**	Ch vs. O	**9.5 [5.0, 13.6]**	**6.8 [2.6, 11.1]**	**−16.3 [−22.2, −10.3]**
Ch vs. Sb	**7.7 [2.2, 12.9]**	−1.4 [−6.8, 4.1]	−6.3 [−13.9, 1.6]	Ch vs. Sb	**7.5 [3.6, 11.0]**	0.8 [−2.9, 4.7]	**−8.2 [−13.5, −3.0]**
				PJ vs. O	**8.4 [3.7, 13.2]**	0.8 [−3.9, 5.5]	**−9.2 [−15.7, −2.5]**

* O = original, Sb = subharmonics, Ch = chaos, PJ = pitch jumps.
** Cells in **bold** contain 95% CIs that exclude zero. This is merely a visualization aid, not significance testing.

(+5.6% [0.2, 11.0]). Subharmonics had little or no effect on the weight of the *Hurt* category (2.9% [−2.0, 7.8]), but they shifted the weight from *Pleased* (−9.8% [−17.1, −3.1]) to *Effortful* (7.0% [2.0, 12.1]). Controlling for HNR weakened the effects of nonlinear phenomena in moans (Table S4).

Screams

Subharmonics in screams were associated with being less *Pleased* (−8.0% [−13.3, −2.7]) and more *Aggressive* (+6.0% [2.3, 9.8]). Chaos shifted the perceived emotion from *Pleased* (−16.3% [−22.2, −10.3]) to *Afraid* (+9.5% [5.0, 13.6]) or *Aggressive* (+6.8% [2.6, 11.1]). Pitch jumps were associated with being less *Pleased* (−9.2% [−15.7, −2.5]) and more *Afraid* (+8.4% [3.7, 13.2]), but not *Aggressive* (+0.8% [−3.9, 5.5]). Controlling for HNR removed the effects of nonlinear phenomena on the perceived level of aggression, but it did not account for the shift of interpretation from pleasure to fear (Table S4).

As in Experiment 1, an exploratory analysis was performed to test whether the perceptual effects of nonlinear vocal phenomena depended on their duration and strength. The strength of subharmonics in gasps was a negative predictor of being *Hurt* (L = 14.6, df = 1, p < .001 without a Bonferroni correction) and a positive predictor of being *Surprised* (L = 11.6, df = 1, p < .001). For laughs, the only significant finding was a positive effect of g_0 frequency on the weight of the *Amused* category (L = 13.1, df = 1, p < .001). No effects were discovered for moans or screams. The length of the episode with chaos had no effect on the weight of any emotion for any call type. More stimuli would be needed to improve the power of this analysis.

Discussion

Experiment 2 aimed to verify and nuance the findings from Experiment 1 by increasing the number of stimuli and using discrete emotions instead of the dimensions of valence, arousal, and dominance. The results strongly confirmed that nonlinear vocal phenomena, particularly chaos, were associated with aversive experiences, but with interesting differences between call types.

The results for laughs were inconclusive and not specific to any particular type of spectral noise. Adding any nonlinearities to screams shifted their interpretation from pleasure to fear. In addition, spectral noise of any kind, but not pitch jumps, made screams more aggressive. The most nuanced picture emerged for the relatively ambiguous vocalizations – gasps and moans. Nonlinearities turned a gasp or moan of pleasure into pain, but chaos was perceived as much more aversive than subharmonics. This difference was not fully explained by harmonicity (HNR), further suggesting that listeners distinguished between these two types of nonlinear vocal phenomena and did not base their judgments solely on overall vocal roughness. Chaos and particularly subharmonics also made moans more effortful.

General discussion

Experiments 1 and 2 tested the perceptual consequences of adding different combinations of nonlinear vocal phenomena to otherwise identical synthetic vocalizations. In contrast to correlational studies, with this approach the effects of specific acoustic

manipulations can be tested directly. The main trade-off is that the synthetic nature of stimuli may affect the results. For example, it remains unclear whether the effects of acoustic manipulations on synthetic laughs were weak because the stimuli were not sufficiently realistic, even if authenticity is less of a concern for the other tested call types (Anikin 2018). It must also be emphasized that the reported findings concern only the effects of nonlinearities on the audience – other methods are needed to clarify what they actually reveal about the vocalizer's affective state and intentions.

Contrary to previous reports, neither nonlinear phenomena nor HNR increased the perceived level of arousal, with the exception of a small effect of subharmonics. This is surprising and needs to be verified, since the association between nonlinear vocal phenomena and high arousal is well documented in many non-human animals (Fitch et al. 2002; Schneider and Anderson 2011; Briefer 2012). A possible explanation is that in real life nonlinear phenomena correlate with other acoustic features indicative of high arousal, such as greater loudness, longer syllables, stronger harmonics, and higher and more variable f_0 (Briefer 2012). Perhaps listeners utilize these other features, rather than nonlinear phenomena per se, to detect high arousal. Another possibility is that nonlinearities signal the intensity of emotional experience rather than the general level of alertness. Following psychological research (Russell 1980; Belin et al. 2008; Lima et al. 2013), arousal in this study was defined as being energetic and alert (vs. relaxed and sleepy), whereas some previous reports of correlations between spectral noise and arousal refer to emotion intensity (Schneider and Anderson 2011; Briefer 2012) rather than alertness. It is thus possible that participants interpreted nonlinear phenomena as indicating a more intense emotional experience, and this was partly captured by the valence but not arousal scale.

In line with predictions, the addition of nonlinear vocal phenomena made valence ratings lower and dominance ratings higher. However, there was an unexpected 'division of labour' between different types of nonlinear phenomena: negative valence was primarily associated with chaos or pitch jumps, while high dominance, aggression, or physical effort were associated with subharmonics. Crucially, the perceptual effects of nonlinear vocal phenomena went beyond what would be expected from their impact on the amount of spectral noise alone, as measured by harmonics-to-noise ratio (HNR). This indicates that listeners distinguished between turbulent noise and nonlinear phenomena, on the one hand, and between pitch jumps, subharmonics, and chaos as different kinds of nonlinear phenomena, on the other.

This finding has immediate practical implications. Although the need for a more fine-grained analysis of spectral noise is increasingly recognized in bioacoustics (Fitch et al. 2002; Riede et al. 2007; Cazau et al. 2016), the acoustic measures reported in most perceptual studies (HNR, jitter, and shimmer) cannot reliably distinguish between turbulent noise, subharmonics, and chaos. Unless spectrograms are inspected manually, pitch jumps slip entirely under the radar of a conventional acoustic analysis, since there are no tools for their automatic detection. At the same time, the present results indicate that different types of nonlinear vocal phenomena convey highly salient and specific information to listeners. The field can thus benefit from more widespread use of specialized tools for analyzing nonlinear dynamics (Tokuda et al. 2002; Cazau et al. 2016) and from experimental manipulation of nonlinear vocal phenomena, as in the present study. A particularly urgent task is to elucidate the role of nonlinearities in

human vocal communication, where they have been largely neglected outside infant studies.

Naturally, the reported perceptual effects of nonlinear phenomena will need to be confirmed in future experiments, and many pieces of the puzzle are still missing. For example, the dominance scale in Experiment 1 had low reliability, and the corresponding effect size was small. The findings for screams and roars in Experiment 2 were consistent with the notion that nonlinearities project a dominant or aggressive attitude, but this effect was attributable to any spectral noise and relatively weak, possibly because all scream-like sounds tended to be interpreted as manifestations of fear. In addition, subharmonics were not more effective than chaos in turning a laugh into a jeer, and the effect of both was again non-specific and fully explained by HNR. The reason may be that individual voiced fragments in laughs are too short for listeners to distinguish between different sources of spectral noise. In any case, the role of nonlinear phenomena, and particularly subharmonics, in asserting social dominance or expressing an aggressive, self-confident attitude is plausible, but uncertain.

In future studies it will also be interesting to investigate how the perceptual effects of nonlinear vocal phenomena depend on their strength, duration, and temporal position within a call. Exploratory analyses of the available data produced inconclusive results, and more extensive testing would be needed to determine the communicative significance of the subharmonic frequency g_0 and of the strength and duration of subharmonics and chaos. In particular, although in most cases g_0 was between 100–200 Hz, it ranged from sidebands corresponding to amplitude modulation at less than 100 Hz to period doubling in screams with g_0 of up to 700–800 Hz. It is an open question whether the perceptual effects of such widely varying subharmonic regimes are comparable enough to treat them as the same manipulation, as in this study.

Perhaps the most important pending question is whether nonlinear phenomena are necessarily interpreted as aversive. Since the most positive call type in this study, laughter, produced relatively weak results, it is still possible that nonlinearities might increase the perceived intensity of any expressed emotion regardless of its valence. The explanation most consistent with the present findings, however, is that listeners generally associate nonlinear phenomena with more negative valence, particularly in call types that can be either aversive or hedonistic. It is even less certain to what extent the production, as opposed to perception, of nonlinearities is restricted to aversive contexts. In the corpus of human nonverbal vocalizations by Anikin and Persson (2017), nonlinear effects are particularly common in cries, screams, and roars, mostly implicating negative emotional states (sadness, anger, pain, etc.). On the other hand, triumph and extreme pleasure also appear to be expressed through intense calls rich in nonlinear phenomena, such as roars of jubilant football fans or orgasmic moans. It remains to be seen whether the listeners' interpretation is correct – that is, whether nonlinear phenomena reliably indicate an aversive context.

In conclusion, using parametric synthesis in order to study the perceptual consequences of changing voice quality in human nonverbal vocalizations has shed new light on the role of nonlinear vocal phenomena and source spectrum. The experimental results presented here can also guide further research on the role of voice quality in acoustic communication using non-synthetic vocalizations and a more traditional, non-interventional approach.

Acknowledgements

I am grateful to Stephan Reber, Tomas Persson, Christian Balkenius, and two anonymous reviewers for their comments on the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Andrey Anikin http://orcid.org/0000-0002-1250-8261

References

Alonso JB, Ferrer MA, Henríquez P, López-de-Ipina K, Cabrera J, Travieso CM. 2015. A study of glottal excitation synthesizers for different voice qualities. Neurocomputing. 150:367–376.
Anikin A. 2018. Soundgen: an open-source tool for synthesizing nonverbal vocalizations. Behav Res Methods. 1–15. doi:10.3758/s13428-018-1095-7
Anikin A, Bååth R, Persson T. 2018. Human non-linguistic vocal repertoire: call types and their meaning. J Nonverbal Behav. 42(1):53–80.
Anikin A, Persson T. 2017. Nonlinguistic vocalizations from online amateur videos for emotion research: a validated corpus. Behav Res Methods. 49(2):758–771.
Arnal LH, Flinker A, Kleinschmidt A, Giraud AL, Poeppel D. 2015. Human screams occupy a privileged niche in the communication soundscape. Curr Biol. 25(15):2051–2056.
August PV, Anderson JG. 1987. Mammal sounds and motivation-structural rules: a test of the hypothesis. J Mammal. 68(1):1–9.
Bachorowski JA, Smoski MJ, Owren MJ. 2001. The acoustic features of human laughter. J Acoust Soc Am. 110(3):1581–1597.
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. Behav Res Methods. 40 (2):531–539.
Blumstein DT, Recapet C. 2009. The sound of arousal: the addition of novel nonlinearities increases responsiveness in marmot alarm calls. Ethol. 115(11):1074–1081.
Boersma P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc Inst Phon Sci. 17(1193):97–110.
Briefer EF. 2012. Vocal expression of emotions in mammals: mechanisms of production and evidence. J Zool. 288(1):1–20.
Bürkner PC. 2017. brms: an R package for Bayesian multilevel models using Stan. J Stat Softw. 80 (1):1–28.
Carvalho CM, Polson NG, Scott JG. 2009. Handling sparsity via the horseshoe. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS); Apr 15; Florida, USA: Clearwater Beach. p. 73–80.
Cazau D, Adam O, Aubin T, Laitman JT, Reidenberg JS. 2016. A study of vocal nonlinearities in humpback whale songs: from production mechanisms to acoustic analysis. Sci Rep. 6:31660.
Facchini A, Bellieni CV, Marchettini N, Pulselli FM, Tiezzi EB. 2005. Relating pain intensity of newborns to onset of nonlinear phenomena in cry recordings. Phys Lett A. 338(3–5):332–337.
Fastl H, Zwicker E. 2006. Psychoacoustics: facts and models. Berlin: Springer.
Fee MS, Shraiman B, Pesaran B, Mitra PP. 1998. The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. Nature. 395(6697):67–71.

Fichtel C, Hammerschmidt K, Jürgens U. 2001. On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. Behav. 138 (1):97–116.
Fitch WT, Neubauer J, Herzel H. 2002. Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. Anim Behav. 63(3):407–418.
Fontaine JR, Scherer KR, Roesch EB, Ellsworth PC. 2007. The world of emotions is not two-dimensional. Psychol Sci. 18(12):1050–1057.
Gelman A, Bois F, Jiang J. 1996. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. J Am Stat Assoc. 91(436):1400–1412.
Goudbeek M, Scherer K. 2010. Beyond arousal: valence and potency/control cues in the vocal expression of emotion. J Acoust Soc Am. 128(3):1322–1336.
Herbst C, Herzel H, Švec JG, Wyman MT, Fitch WT. 2013. Visualization of system dynamics using phasegrams. J R Soc Interface. 10(85):20130288.
Herzel H, Berry D, Titze I, Steinecke I. 1995. Nonlinear dynamics of the voice: signal analysis and biomechanical modeling. Chaos. 5(1):30–34.
Karp D, Manser MB, Wiley EM, Townsend SW. 2014. Nonlinearities in meerkat alarm calls prevent receivers from habituating. Ethology. 120(2):189–196.
Koutseff A, Reby D, Martin O, Levrero F, Patural H, Mathevon N. 2018. The acoustic space of pain: cries as indicators of distress recovering dynamics in pre-verbal infants. Bioacoustics. 27 (4):313–325.
Lavan N, Scott SK, McGettigan C. 2016. Laugh like you mean it: authenticity modulates acoustic, physiological and perceptual properties of laughter. J Nonverbal Behav. 40(2):133–149.
Lima CF, Castro SL, Scott SK. 2013. When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing. Behav Res Methods. 45(4):1234–1245.
Mann DA, O'Shea TJ, Nowacek DP. 2006. Nonlinear dynamics in manatee vocalizations. Mar Mamm Sci. 22(3):548–555.
McAleer P, Todorov A, Belin P. 2014. How do you say 'Hello'? Personality impressions from brief novel voices. PLoS One. 9(3):e90779.
Mende W, Herzel H, Wermke K. 1990. Bifurcations and chaos in newborn infant cries. Phys Lett A. 145(8–9):418–424.
Mergell P, Herzel H. 1997. Modelling biphonation – the role of the vocal tract. Speech Commun. 22(2–3):141–154.
Morton ES. 1977. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. Am Nat. 111(981):855–869.
Neubauer J, Edgerton M, Herzel H. 2004. Nonlinear phenomena in contemporary vocal music. J Voice. 18(1):1–12.
Ohala JJ. 1984. An ethological perspective on common cross-language utilization of F_0 of voice. Phonetica. 41(1):1–16.
Puts DA, Gaulin SJ, Verdolini K. 2006. Dominance and the evolution of sexual dimorphism in human voice pitch. Evol Hum Behav. 27(4):283–296.
Puts DA, Hodges CR, Cárdenas RA, Gaulin SJ. 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. Evol Hum Behav. 28(5):340–344.
Raine J, Pisanski K, Simner J, Reby D. 2018. Vocal communication of simulated pain. Bioacoustics. 1–23. doi:10.1080/09524622.2018.1463295
Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton-Brock T. 2005. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. Proc R Soc Lond B Biol Sci. 272(1566):941–947.
Reby D, Wyman MT, Frey R, Passilongo D, Gilbert J, Locatelli Y, Charlton BD. 2016. Evidence of biphonation and source–filter interactions in the bugles of male North American wapiti (Cervus canadensis). J Exp Biol. 219(8):1224–1236.
Riede T, Arcadi AC, Owren MJ. 2007. Nonlinear acoustics in the pant hoots of common chimpanzees (Pan troglodytes): vocalizing at the edge. J Acoust Soc Am. 121(3):1758–1767.

Riede T, Herzel H, Mehwald D, Seidner W, Trumler E, Böhme G, Tembrock G. 2000. Nonlinear phenomena in the natural howling of a dog–wolf mix. J Acoust Soc Am. 108 (4):1435–1442.
Riede T, Mitchell BR, Tokuda I, Owren MJ. 2005. Characterizing noise in nonhuman vocalizations: acoustic analysis and human perception of barks by coyotes and dogs. J Acoust Soc Am. 118(1):514–522.
Robb MP, Saxman JH. 1988. Acoustic observations in young children's non-cry vocalizations. J Acoust Soc Am. 83(5):1876–1882.
Russell JA. 1980. A circumplex model of affect. J Pers Soc Psychol. 39(6):1161–1178.
Schneider JN, Anderson RE. 2011. Tonal vocalizations in the red wolf (Canis rufus): potential functions of nonlinear sound production. J Acoust Soc Am. 130(4):2275–2284.
Siebert ER, Parr LA. 2003. A structural and contextual analysis of chimpanzee screams. Ann N Y Acad Sci. 1000(1):104–109.
Sueur J. 2018. Sound analysis and synthesis with R. Cham: Springer. doi:10.1007/978-3-319-77647-7
Szameitat DP, Alter K, Szameitat AJ, Darwin CJ, Wildgruber D, Dietrich S, Sterr A. 2009. Differentiation of emotions in laughter at the behavioral level. Emotion. 9(3):397–405.
Titze IR. 2008. Nonlinear source–filter coupling in phonation: theory. J Acoust Soc Am. 123 (4):1902–1915.
Tokuda I, Riede T, Neubauer J, Owren MJ, Herzel H. 2002. Nonlinear analysis of irregular animal vocalizations. J Acoust Soc Am. 111(6):2908–2919.
Tokuda IT. 2018. Non-linear dynamics in mammalian voice production. Anthropol Sci. 171130:1–7.
Townsend SW, Manser MB. 2011. The function of nonlinear phenomena in meerkat alarm calls. Biol Lett. 23. 7(1):47–49.
Tyson RB, Nowacek DP, Miller PJ. 2007. Nonlinear phenomena in the vocalizations of North Atlantic right whales (Eubalaena glacialis) and killer whales (Orcinus orca). J Acoust Soc Am. 122(3):1365–1373.
Wilden I, Herzel H, Peters G, Tembrock G. 1998. Subharmonics, biphonation, and deterministic chaos in mammal vocalization. Bioacoustics. 9(3):171–196.
Wood A, Martin J, Niedenthal P. 2017. Towards a social functional account of laughter: acoustic features convey reward, affiliation, and dominance. PLoS One. 12(8):e0183811.

Paper VI

A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations

Andrey Anikin

Division of Cognitive Science, Lund University, Sweden

Abstract

Prosodic features, such as intonation and voice intensity, have a well-documented role in communicating emotion, but less is known about the role of laryngeal voice quality in speech and particularly in nonverbal vocalizations such as laughs and moans. Potentially, however, variations in voice quality between tense and breathy may convey rich information about the speaker's physiological and affective state. In this study breathiness was manipulated in synthetic human nonverbal vocalizations by adjusting the relative strength of upper harmonics and aspiration noise. In Experiment 1 (28 prototypes x 3 manipulations = 84 sounds), otherwise identical vocalizations with tense vs. breathy voice quality were associated with higher arousal (general alertness), higher dominance, and lower valence (unpleasant states). Ratings on discrete emotions in Experiment 2 (56 x 3 = 168 sounds) confirmed that breathiness was reliably associated with positive emotions, particularly in ambiguous vocalizations (gasps and moans). Spectral centroid did not fully account for the effect of manipulation, confirming that the perceived change in voice quality was more specific than a general shift in timbral brightness. Breathiness is thus involved in communicating emotion with nonverbal vocalizations, possibly due to changes in low-level auditory salience and perceived vocal effort.

Keywords: voice quality, nonverbal vocalizations, emotion, voice synthesis, glottal source

Introduction

Nonverbal vocalizations, such as moans or laughs, are ubiquitous in everyday interaction, express a wide range of easily recognizable emotions and attitudes (Anikin & Persson, 2017; Lima, Castro, & Scott, 2013; Sauter, Eisner, Calder, & Scott, 2010; Wood, Martin, & Niedenthal, 2017), and display significant cross-cultural similarities (Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016). Their relatively simple acoustic structure, intuitiveness, and flexible meaning make nonverbal vocalizations attractive options for enriching speech synthesis (Campbell, 2006) and human-machine interaction (Haddad, Çakmak, Sulír, Dupont, & Dutoit, 2016). Furthermore, the acoustic structure and production context of some vocalizations, such as laughs (Ross, Owren, & Zimmermann, 2009) and screams (Högstedt, 1983), show marked similarities across species, suggesting that these sounds predate language and have deep biological roots. Insights from ethology can therefore prove instrumental for research on human nonverbal communication; in turn, learning more about human nonverbal repertoire can shed new light on acoustic communication in non-human animals.

In order to elucidate how humans communicate with nonverbal vocal cues, a crucial task is to understand the underlying acoustic code. This is an active area of research; several studies have reported detailed acoustic analyses of human nonverbal vocalizations, looking for acoustic correlates of emotion either across all types of vocalizations (Anikin & Persson, 2017; Lima et al., 2013; Sauter et al., 2010) or in particular vocalizations or call types such as laughter (Szameitat et al., 2009; Wood et al., 2017). In many ways this research parallels the more extensive search for acoustic markers of emotion in speech (e.g., Banse & Scherer, 1996; Murray & Arnott, 1993; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). In both cases, the main focus is on easily measured prosodic characteristics such as intonation, intensity, and temporal features. More subtle acoustic features, such as the spectrum of laryngeal vocal source, the presence and spectrum of turbulent noise, the variability in the period (jitter) and amplitude (shimmer) of glottal pulses, or nonlinear vocal phenomena (e.g., pitch jumps and subharmonics), are more challenging to measure accurately and consequently less well-understood (Gobl & Ní Chasaide, 2003). At the same time, these aspects of vocal production, which give the voice a particular coloring or “voice quality”, provide important information about the speaker's age, sex, and emotion in speech (Airas & Alku, 2006; Cummings & Clements, 1995; Grichkovtsova, Morel, & Lacheret, 2012; He, Lech, & Allen, 2010; Johnstone & Scherer, 1999; Laukkanen, Vilkman, Alku, & Oksanen, 1996; Murray & Arnott, 1993; Patel, Scherer, Björkner, & Sundberg,

2011; Waaramaa, Laukkanen, Airas, & Alku, 2010) and probably also in nonverbal vocalizations (Lima et al., 2013; Mittal & Yegnanarayana, 2014; Wood et al., 2017). In fact, non-speech sounds may be particularly suitable for studying the intrinsic link between voice quality and emotion because they are free both from semantic contents (Lavan, Scott, & McGettigan, 2016) and from the constraints imposed by language-specific phonemic structure or socio-cultural rules (Patel et al., 2011). For example, adding different types of nonlinear phenomena to human nonverbal vocalizations revealed that these acoustic features are perceptually salient and associated with distinct affective states: abrupt pitch jumps in screams enhance the impression of fear, episodes of unstable phonation with subharmonics or chaos in gasps and moans make the speaker sound hurt rather than pleased, and so on (Anikin, 2019b). The focus in the present paper is on the perceptual consequences of manipulating another aspect of voice quality, namely breathiness.

What makes a voice tense or breathy?

Vibrating vocal folds produce changes in air pressure that are periodic, but not perfectly sinusoidal. As a result, the glottal source of excitation contains the lowest frequency component determined by the rate at which the vocal folds oscillate (known as the fundamental frequency or f_0) and a number of other frequency components that are multiples of f_0 (harmonics). The energy of harmonics above f_0 dissipates, or rolls off - hence, "rolloff" - approximately exponentially at a rate of about 6-12 dB per octave, depending on the mode of phonation, alveolar pressure, f_0, and other factors (Stevens, 2000, Ch. 2). The buzz-like glottal pulses, as well as other excitation sources such as aspiration noise, are then modified by the resonances of the vocal tract before being perceived by the listeners, as described by the source-filter model (Fant, 1960). For a particular speaker, the exact shape of glottal pulses, and therefore the spectrum of glottal source, is primarily determined by the tension of the vocal folds and subglottal pressure, which are controlled by laryngeal and respiratory muscles (Gobl & Ní Chasaide, 2010). In addition, source spectrum can be significantly affected by nonlinear interactions between glottal source and filter, which are particularly relevant when f_0 is high enough to cross formant frequencies, as in many nonverbal vocalizations (Titze, 2008).

The spectrum of glottal pulses prior to their filtering by the vocal tract (source spectrum) has a major impact on the perceived voice quality (Gobl & Ní Chasaide, 2010; Kreiman, Gerratt, Garellek, Samlan, & Zhang, 2014). In fact, used in the narrow sense, the term "voice quality" may refer specifically to laryngeal source (Gobl & Ní Chasaide, 2003), although other aspects of vocal production are often also included, and the exact terminology varies across disciplines. One of the most

widely recognized and perceptually important dimensions is breathiness, which describes variations in voice quality from tense to breathy, with modal phonation as the neutral type. A tense, or pressed, voice is characterized by complete and abrupt closure of the vocal folds, strong harmonics, and little or no aspiration noise. In contrast, a breathy voice is characterized by loosely closed glottis, a strong f_0, weak harmonics, and audible aspiration noise caused by air leaking through the partially closed glottis (Gobl & Ní Chasaide, 2003, 2010; Stevens, 2000). The strength of upper harmonics and the amount of aspiration noise are thus the main acoustic correlates of voice quality changes along the tense-breathy continuum, which are referred to as "breathiness" in the remainder of the text.

Evidence linking breathiness and emotion

Because the prominence of upper harmonics depends on subglottal pressure and the activity of laryngeal muscles (Gobl & Ní Chasaide, 2010), breathiness has the potential to convey rich information about the speaker's physiological and affective state. To test whether listeners do utilize this information, it is necessary to manipulate, or at least to measure accurately, source spectrum. The most reliable way to estimate source spectrum is to monitor the oscillations of glottal folds directly using electroglottography (Laukkanen et al., 1996), but a more common indirect approach involves inverse filtering, which removes the contribution of the vocal tract by deconvolution of the signal with an estimated vocal tract transfer function (Drugman, Alku, Alwan, & Yegnanarayana, 2014). Several studies performed inverse filtering to demonstrate that source spectrum varied depending on the speaker's emotion in vowel sounds extracted from speech (Cummings & Clements, 1995; He et al., 2010; Johnstone & Scherer, 1999; Laukkanen et al., 1996; Patel et al., 2011), and there have been attempts to apply inverse filtering to laughs (Mittal & Yegnanarayana, 2014). Despite many methodological challenges and inconsistent definitions of voice quality in different studies (Gobl & Ní Chasaide, 2010), one of the most robust findings appears to be the association of pressed phonation with anger or other intense emotions, and of breathy phonation with more relaxed or subdued affective states. There are also reports that variations along the pressed-breathy continuum in synthesized speech are associated with perceived speaker's arousal or general alertness (Brady, 2005; Gobl & Ní Chasaide, 2003).

Researchers who do not have access to highly specialized recording facilities, or who wish to analyze large collections of recordings, are usually unable to perform electroglottography or inverse filtering and are confined to describing the observed spectrum (Gobl & Ní Chasaide, 2010). Conventional measures of the general shape of spectral envelope that have been reported in relation to emotion in speech

or nonverbal vocalizations include spectral center of gravity or centroid (Lavan et al., 2016; Lima et al., 2013; Sauter et al., 2010), peak frequency with the highest amplitude within the spectrum (Scheiner, Hammerschmidt, Jürgens, & Zwirner, 2002), spectral slope or tilt (Goudbeek & Scherer, 2010; Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001), ratios of energy above and below a certain frequency (Patel et al., 2011; Wood et al., 2017), dominant frequency bands and quantiles of spectral energy distribution (Fichtel, Hammerschmidt, & Jürgens, 2001; Hammerschmidt and Jürgens, 2007), or principal components combining several of these features. All these measures say something about the balance of low- and high-frequency energy in the spectrum, but their informativeness about source spectrum is limited by two factors. First, they do not distinguish between the contribution of source and filter. Increasing the frequency of one or more formants has the effect of raising the spectral centroid and making the voice "brighter" regardless of the glottal source (Fastl & Zwicker, 2006; Stevens, 2000). As a result, a vowel like [a] (high F1, average F2) will have noticeably stronger harmonics and sound brighter than [u] (low F1 and F2), even when both are pronounced or synthesized with the same glottal source. Second, glottal pulses are not the only source of excitation: the presence of turbulent noise can have a major effect on the shape of the resulting spectrum. For example, the spectrum flattens and its center of gravity rises as harmonics increase in strength in a tonal sound (Fig. 1, left panel), but this effect is much less noticeable in the presence of aspiration noise (right panel). In other words, harmonics in a breathy voice are weaker than might be expected based on the overall spectral slope (Gobl & Ní Chasaide, 2010). As a result, summary measures of spectral envelope may fail to capture variation in voice quality that is potentially informative to listeners.

Keeping in mind these limitations of non-specific spectral measures, such as peak frequency or spectral centroid, there are several reports that listeners interpret high-frequency spectral energy as a sign of high arousal in speech (Johnstone & Scherer, 1999; Schröder et al., 2001) and in nonverbal vocalizations (Lavan et al., 2016; Lima et al., 2013). Raine, Pisanski, Simner, and Reby (2018) also report that listeners associate breathy voices with low pain intensity. In addition, there is extensive evidence from ethological literature that the spectrum contains more high-frequency energy when the animal is highly aroused (Briefer, 2012; Fichtel & Hammerschmidt, 2002) or distressed (Lingle, Wyman, Kotrba, Teichroeb, & Romanow, 2012). These studies support the generalization that stronger harmonics in source spectrum may be associated with higher general alertness (arousal) or emotion intensity not only in speech, but also in human nonverbal vocalizations and animal calls.

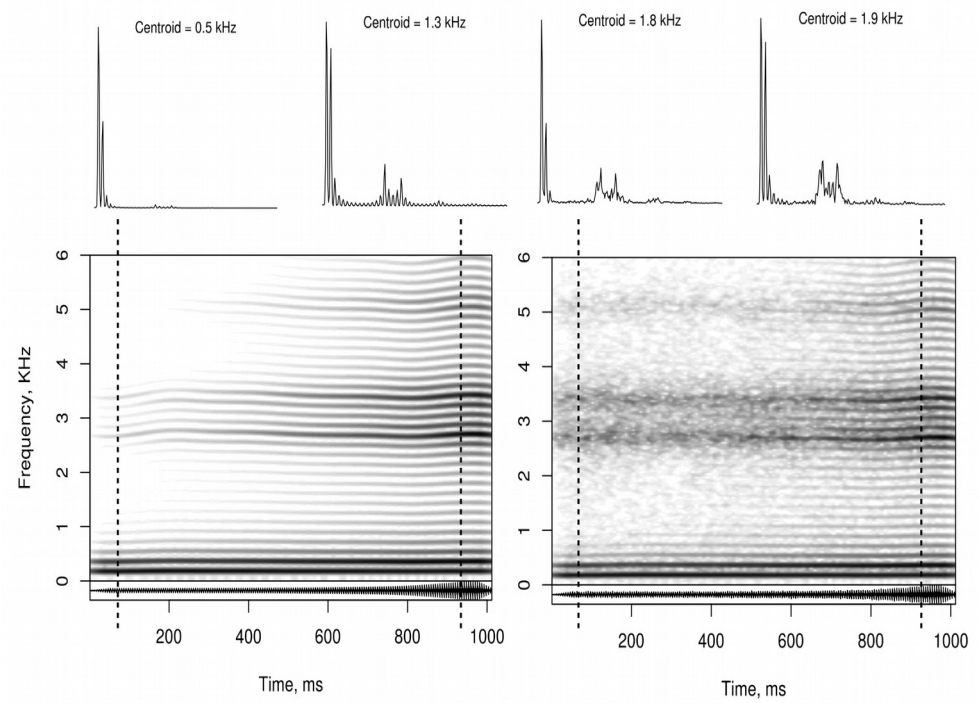

Fig. 1. Manipulation of rolloff from -16 to -6 dB/octave in a synthetic vowel that is purely harmonic (left panel) or contains turbulent noise at a constant level of -14 dB relative to the harmonic component (right panel). Spectrograms and spectral slices at ~50 ms and 950 ms. Observe that the spectral centroid is less dependent on the strength of harmonics in the presence of turbulent noise. AUDIO #1 in Supplements.

Besides the clearly motivated connection of a tense voice quality with arousal, various measures of high-frequency energy have been associated with unpleasant emotional experiences (negative valence) in human nonverbal vocalizations (Sauter et al., 2010; Scheiner et al., 2002) and speech (Goudbeek & Scherer, 2010; Hammerschmidt & Jurgens, 2007). On the other hand, spectral centroid was positively correlated with reward ratings of laughs in the study by Wood et al. (2017). The evidence is also mixed for animal vocalizations (Briefer, 2012), and the issue is further complicated by the fact that strongly negative affective states are usually associated with high arousal, making it difficult to know whether it is the intensity of emotion or its unpleasantness that is responsible for the observed acoustic characteristics. In one of the most methodologically rigorous studies, peak frequency was the best predictor of negative valence in squirrel monkeys, but only for relatively ambiguous vocalizations (Fichtel et al., 2001). Earlier maximum peak frequency has also been associated with positive valence,

suggesting that a downward trajectory of peak frequency may mark positive valence (Briefer, 2012; Hammerschmidt & Jurgens, 2007). The relationship between spectral envelope and perceived aversive or hedonistic nature of vocalizations may thus depend on temporal dynamics, the type of vocalization, and/or it may be mediated by arousal. Finally, several reports have linked high-frequency spectral energy to speaker's dominance or aggression in speech (Banse & Scherer, 1996; Gobl & Ní Chasaide, 2010; Hammerschmidt & Jurgens, 2007; McAleer, Todorov, & Belin, 2014) and in human nonverbal vocalizations (Sauter et al., 2010; Wood et al., 2017). However, many of these effects may be mediated by arousal or the intensity of affect in general, since high-frequency spectral energy has also been reported to correlate with the perceived intensity of several emotions (e.g., Banse & Scherer, 1996).

To summarize, the available evidence suggests that listeners interpret tense voices with strong harmonics as indicators of high alertness (arousal) or intense emotional states. Shifts from tense to breathy phonation may also be interpreted as a sign of unpleasant affective states (negative valence) or an assertive attitude (high dominance), but the evidence in this respect appears to be more mixed. In addition, there seems to be no experimental data showing that listeners attend to the strength of harmonics independently of the overall distribution of energy in the spectrum. Crucially, the most acoustically informative evidence of the role of laryngeal source – obtained with inverse filtering or direct manipulations of voice quality in synthetic stimuli – comes from studies of isolated vowels or short verbal utterances. The role of laryngeal voice quality in naturalistic nonverbal vocalizations or animal calls remains largely uncharted.

The present study

Affective speech synthesis has been slow in coming because of many technical challenges (Schröder, 2009), but it is an attractive complementary approach to inverse filtering that offers an ability to modify the laryngeal source according to stringent definitions and without acoustic confounds common in correlational studies (Gobl & Ní Chasaide, 2003). The aim of the present study was to capitalize on this underutilized methodological opportunity and to shed new light on the role of tense-breathy voice quality in emotional nonverbal vocalizations. *Soundgen*, an open-source formant synthesizer developed and validated specifically for parametric synthesis of nonverbal vocalizations (Anikin, 2019a), was used to synthesize a number of laughs, screams, and other non-speech vocalizations; each sound was created in three versions that differed only in voice quality and had identical duration, intonation and other acoustic characteristics. Most stimuli included both a harmonic component and some turbulent noise, and the

manipulation had the effect of simultaneously modifying (1) the strength of higher harmonics relative to f_0 and (2) the strength of the harmonic component as a whole relative to the noise component. Qualitatively, this approximately corresponds to changing the perceived voice quality along the tense-breathy continuum, although this terminology was developed for speech and may not adequately describe acoustically "extreme" sounds like high-pitched screams.

Importantly, because the rolloff of harmonics was not the sole determinant of the observed spectrum in the presence of turbulent noise, it was possible to tease apart the effects of excitation source and overall spectral balance of energy. Statistically, this was achieved by analyzing the effect of manipulation after controlling for spectral centroid – the most common summary measure of spectral envelope and an excellent predictor of perceived timbral brightness of human voice (Fastl & Zwicker, 2006) and musical tones (Schubert, Wolfe, & Tarnopolsky, 2004). Based on the evidence reviewed above, it was hypothesized that shifts from breathy to tense phonation would produce an impression of intense or unpleasant emotional states.

To test this hypothesis, listeners rated nonverbal vocalizations with manipulated breathiness on valence, arousal, and dominance scales (Experiment 1) and then on discrete emotions (Experiment 2). Valence and arousal are among the most commonly used dimensions of emotional experience in both human (Belin, Fillion-Bilodeau, & Gosselin, 2008; Lima et al., 2013) and animal (Briefer, 2012) research. Dominance is less well established as an emotional dimension, and the literature mentions a variety of conceptually related measures such as control, power, or potency (Goudbeek & Scherer, 2010). Categorization into discrete emotions is sometimes obtained alongside ratings on continuous dimensions in perceptual studies (e.g., Lima et al., 2013); in this study this was a natural choice given that (a) sounds in the original corpus were obtained from contexts related to several well-defined affective states, and (b) two previous studies established which emotions are most commonly perceived by people who hear these sounds (Anikin Bååth, & Persson, 2018; Anikin, 2019a).

Experiment 1

Methods

Stimuli.

The stimuli were synthetic versions of human nonverbal vocalizations from the corpus collected by Anikin & Persson (2017). The original vocalizations were mostly non-staged, spontaneous emotional bursts that had been captured on video in real-life situations and then uploaded to social media. These sounds included little or no phonemic structure and were associated with a powerful emotional experience such as incurring a physical injury (pain), being the victim of a scare prank (fear), witnessing a funny accident (amusement), and so on for a total of nine emotions, whose recognition was tested in a validation study (Anikin & Persson, 2017). The sounds have also been sorted into call types based on linguistic labeling by English-, Swedish-, and Russian-speaking participants as well as acoustic analysis (Anikin et al., 2018).

For the present study, 28 prototypes were chosen from the larger corpus to represent the following seven call types: cry, gasp, grunt, laugh, moan, roar, and scream. These particular sounds were chosen primarily based on the relatively high authenticity ratings of their synthetic versions in the *soundgen* validation study (Anikin, 2019a), and together they represent a broad range of vocalizations from the human nonverbal repertoire. The 28 prototype vocalizations were all from different individuals (17 women and 11 men) and varied between 0.22 and 2.7 s in duration (Table 1).

Each of these 28 sounds was manually analyzed, and then a similar synthetic version was created with original / strengthened / weakened harmonics in source spectrum using *soundgen* 1.1.2 (Anikin, 2019a), for a total of 28 x 3 = 84 stimuli. *Soundgen* is an implementation of the source-filter model written in the *R* language that creates a separate sine wave for each harmonic, adds a noise component, and filters this excitation source with a simulated transfer function. All control parameters are set manually by specifying a few anchors that are interpolated to produce smooth curves for an entire syllable or bout; for example, the intonation contour is given by a few pitch anchors at different time points. The quality of this parametric synthesis had previously been validated, in the sense that synthetic vocalizations were shown to be similar to the original recordings in

terms of their ratings on valence and arousal as well as perceived emotion, and in most cases the authenticity of synthetic stimuli was on a par with the originals (Anikin, 2019a). At the same time, the synthesized sounds in this study were not exact replicas of the 28 prototypes, and their fully synthetic nature made it possible to manipulate any desired acoustic characteristic, without any limitations associated with audio editing or resynthesis.

Table 1. Acoustic characteristics of the synthetic stimuli in Experiment 1.

Call type	Number of stimuli		Acoustic characteristics: mean [range]					
	Total (with noise)	Female / male	Rolloff manipu-lation (dB/oct)	Duration (s)	Median voiced syllable (s)	Median f_0 (Hz)	Median HNR* (dB)	Median spectral centroid (kHz)
Cry	4 (2)	3/1	±4 [4, 4]	2.2 [1.6, 2.6]	0.3 [0.1, 0.5]	468 [255, 790]	15.5 [13.1, 19]	1.5 [0.8, 2.1]
Gasp	4 (4)	3/1	±7.5 [6, 8]	1.3 [1.2, 1.5]	1.0 [0.6, 1.2]	362 [255, 498]	11.2 [5.2, 16.6]	1.9 [1.0, 2.8]
Grunt	4 (4)	2/2	±5.5 [4, 10]	0.4 [0.2, 0.5]	0.3 [0.1, 0.4]	289 [190, 378]	4.1 [0.7, 9.5]	1.3 [0.8, 1.6]
Laugh	4 (4)	1/3	±5 [4, 8]	1.8 [1.4, 2.6]	0.1 [0.1, 0.2]	459 [330, 617]	7.9 [3.2, 12.7]	2.1 [1.4, 2.6]
Moan	4 (4)	2/2	±5.5 [4, 10]	1.3 [0.7, 1.8]	0.5 [0.4, 0.6]	293 [161, 490]	8.6 [1.7, 14.8]	1.5 [1.0, 2.2]
Roar	4 (2)	2/2	±4 [4, 4]	0.8 [0.5, 1.1]	0.8 [0.5, 1.0]	454 [288, 701]	1.2 [-3.0, 5.2]	1.8 [0.9, 2.9]
Scream	4 (0)	4/0	±7 [4, 10]	0.9 [0.6, 1.2]	0.8 [0.4, 1.2]	1800 [1420, 2102]	9.1 [3.4, 16.7]	2.6 [1.9, 4]

*HNR = harmonics-to-noise ratio, measured based on autocorrelation. HNR and spectral centroid were measured in the synthesized audio files using *soundgen*, and the rest of acoustic characteristics were derived directly from the settings used for synthesizing the stimuli.

The harmonic structure of spectral source is controlled by several parameters in *soundgen*, but its overall slope can be manipulated with a single parameter called "rolloff", which corresponds to the rate of exponential decay of the energy in harmonics above f_0, in dB/octave (Fig. 2). Depending on the sound, rolloff was changed by ±4-10 dB/octave, aiming to make the magnitude of manipulation comparable across all stimuli in terms of the perceptual salience of voice quality changes. All sounds were normalized for peak amplitude, so changes in rolloff did not entail major changes in the overall sound pressure level, although subjective loudness may have changed. Increasing or decreasing the strength of harmonics changed the spectral centroid on average by +316 Hz and -183 Hz, respectively. However, 20 out of 28 prototype sounds included some turbulent noise, whose spectrum was not affected by the rolloff setting. Because the amplitude of the noise component was tied to the amplitude of the first harmonic, changing the amplitude of harmonics relative to f_0 also affected the relative amplitudes of the voiced and unvoiced (noise) components. In particular, noise partly replaced higher harmonics in manipulated sounds with steeper (more negative) rolloff,

making the voice sound breathy, while in manipulated sounds with shallower (less negative) rolloff harmonics tended to replace the noise, creating the impression of pressed phonation (Fig. 2). High-pitched screams and roars were synthesized without a noise component because in real life these vocalizations are normally too loud for aspiration noise to be audible. In these sounds, the manipulation affected only the strength of harmonics relative to f_0.

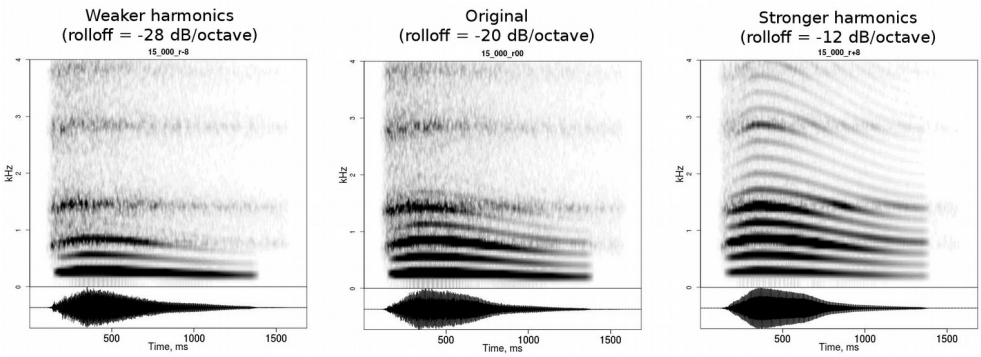

Fig. 2. Spectrograms illustrating the manipulation of rolloff in synthetic gasp #15 corresponding to changes in voice quality along the breathy-tense continuum. AUDIO #2 in Supplements.

Procedure.

The rating experiment was performed online. To avoid presenting very similar sounds repeatedly to the same participants and to optimize the effectiveness of data collection, the rolloff manipulation was tested partly as a separate study, and partly in conjunction with another experiment on nonlinear vocal phenomena, which employed different manipulations, but the same prototype sounds and design (Anikin, 2019b). The stimuli were divided in such a manner that each batch contained at most two manipulated versions of the same prototype stimulus, and each batch was rated by a different sample of participants. As a result, each participant heard only a subsample of experimental stimuli and rated each stimulus on three scales (valence, arousal, and dominance), which were explained to participants as follows:

Valence *is high if the experience is pleasant, so the speaker is happy, pleased, relieved, etc. Valence is low if the experience is unpleasant, so the speaker is sad, afraid, in pain, etc.*

Arousal *is high if the person is very energetic, alert, wide-awake. Arousal is low if the person is sleepy, relaxed, calm.*

Dominance *is high if the speaker sounds assertive, self-confident, superior, perhaps aggressively so. Dominance is low if the person sounds submissive, uncertain, perhaps as someone who seeks reassurance or needs a hug.*

The rating was performed on a continuous horizontal visual analog scale. To minimize the correlation between valence, arousal, and dominance ratings, the experiment was divided into three blocks, in random order. For example, one participant might first rate the stimuli on valence, then on arousal, and finally on dominance; another participant would begin with dominance, etc. The order of sounds within each block was also randomized for each participant. Prior to each block the upcoming scale was illustrated with two contrasting examples: non-synthetic recordings of a person crying (low valence) or laughing (high valence), sighing (low arousal) or screaming (high arousal), and whimpering (low dominance) or roaring (high dominance; adapted from Puts, Gaulin, & Verdolini, 2006).

Participants.

Data quality was ensured by carefully checking all submissions and reimbursing only participants with minimum 40 trials and responses that were not obviously faked (extremely fast and stereotypical). Out of 136 submissions that passed this minimal quality control, the responses of four participants were clear outliers in terms of their low correlation with the global median ratings across all three scales ($r < 0.2$), presumably indicating that they had not attended to the task. The responses of these four participants (2.2% of all data) were therefore excluded from the analysis. In addition, individual trials with a response time under 500 ms presumably represented accidental clicks and were removed from the dataset (<0.5% of all data). The final sample consisted of 132 participants, of whom 18 were unpaid volunteers contacted via online advertisements and 114 were recruited from https://www.prolific.ac. Each sound was rated on average 43 times (range 40 to 47) on each scale. No demographic characteristics were collected; according to the statistics on https://www.prolific.ac/demographics, over 80% of participants on this platform are native English speakers, and about 75% are between 20 and 40 years of age.

Statistical analysis.

The response variable was the rating of a sound on a continuous scale (valence, arousal, or dominance) provided in a single trial by a particular participant. These ratings were modeled using the beta distribution in Bayesian mixed models with random intercepts per participant, per stimulus, and per prototype (shared by all sounds that were modifications of the same original vocalization). Rolloff manipulations were treated as a continuous variable with three values: 0 for more

negative rolloff (breathy voice), 0.5 for original, and 1 for less negative (tense voice). To compare the relative contributions of the relative strength of harmonics and the overall spectral shape, rolloff manipulation and the logarithm of spectral centroid were entered simultaneously in multiple regression models. Mixed models were fit in Stan computational framework (http://mc-stan.org/) accessed with R package *brms* (Bürkner, 2017). To improve convergence and guard against overfitting, regularizing priors were used for all regression coefficients. The effects were summarized as the median of posterior distribution and 95% credible interval. The stimuli, R code for their generation, experimental datasets, and scripts for statistical analysis can be downloaded from http://cogsci.se/publications.html.

Results

Inter-rater reliability was moderate for valence (ICC = .48, 95% CI [.41, .57]) and arousal (ICC = .53 [.46, .61]). These levels agreement of agreement among raters were similar to those in previous studies of both real (Anikin, 2019a) and synthetic (Anikin, 2019a, 2019b) versions of similar sounds. In contrast, dominance was rated less consistently by different participants: ICC = .22 [.17, .29]. The results for the dominance scale should therefore be treated with caution.

Increasing the rolloff parameter in *soundgen* (making it less negative) made the voice tenser (strong harmonics, less aspiration noise), while decreasing the rolloff made it more breathy (weak harmonics, more aspiration noise). The valence ratings were 3.4% (95% CI [0.4, 6.3]) lower in otherwise identical sounds with tense vs. breathy phonation. This small negative effect remained essentially unchanged after controlling for spectral centroid (-3.0% [-6.6, 0.5]), while the independent effect of spectral centroid after controlling for rolloff was highly uncertain: -2.6% [-17.8, 11.7] over the observed range of 0.8 to 4 kHz.

As predicted, making the voice tenser increased arousal ratings by 6.9% [3.6, 10.1]. The effect of rolloff on arousal became slightly weaker after controlling for spectral centroid (5.3% [1.0, 9.3]), whereas the effect of spectral centroid itself on perceived arousal was statistically uncertain after controlling for rolloff: 11.3% [-6.0, 29.0]. In other words, the effect of manipulations on arousal may be partly mediated by the general shape of the spectral envelope, but the strength of harmonics as such clearly makes an independent contribution. Dominance ratings were slightly higher for tense vs. breathy voice quality (2.8% [0.4, 5.1]). It is not clear to what extent this change may be mediated by spectral centroid: the effect of rolloff manipulation controlling for spectral centroid is predicted to be 2.1% [-0.8,

5.1], and that of spectral centroid 4.9% [-6.6, 16.2], which is too uncertain to draw any firm conclusions.

Although the three resynthesized versions of each prototype vocalization – with original / increased / decreased rolloff – differed only in the synthetic equivalent of laryngeal voice quality, there was a large variation among the 28 prototype sounds in terms of their duration, temporal structure, average pitch, and other prosodic characteristics. To test whether the effect of rolloff manipulation depended on some of these acoustic characteristics, likelihood ratio tests in non-Bayesian mixed models were employed to test the significance of interaction terms between rolloff manipulation and the following acoustic characteristics: duration, median f_0, median and maximum length of voiced syllables (taken directly from the control parameters supplied to the synthesizer), and the original speaker's sex. None of these interaction terms were significant after Bonferroni correction in models predicting valence, arousal, or dominance ratings, suggesting that the effects of rolloff manipulations were broadly consistent across acoustically diverse stimuli. On the other hand, the number of stimuli and effect sizes were not large enough to reveal relatively subtle interactions; the resolution of this analysis could be improved by creating and testing a larger number of stimuli.

Discussion

The aim of Experiment 1 was to perform an initial exploration of the perceptual consequences of manipulating laryngeal voice quality along the tense-breathy continuum in several types of human nonverbal vocalizations. The tested manipulations affected the ratings of synthetic vocalizations in a manner broadly consistent with theoretical predictions. Making the voice tenser enhanced the perceived level of speaker's general alertness or arousal; it also and made the vocalizations slightly more aversive and enhanced the speaker's perceived dominance, although the latter effect was uncertain. A commonly reported measure of timbral brightness or overall balance of low- and high-frequency energy in the spectrum – spectral centroid – appeared to contribute to the observed effects of voice quality, but did not fully account for them. From a listener's perspective, the experimental manipulations of voice quality thus appeared to be both salient and more specific than a general shift of energy towards higher frequencies.

Experiment 2

Some perceptual effects of source spectrum may be specific to particular types of vocalizations, but these differences could not be estimated in Experiment 1 with only four prototype sounds per acoustic class. Accordingly, in the follow-up study the number of sounds of the same type, such as laugh or scream, was increased. To keep the overall number of stimuli manageable, only four acoustic classes were investigated: one that was predominantly positive in valence (laughs), one negative (screams or high-pitched roars), and two more ambivalent (gasps and moans). Laughs, screams, and moans are among the four most universally recognized human nonverbal vocalizations (Anikin et al., 2018). Gasps have not been extensively studied, but their ingressive nature makes them very distinct acoustically.

Only a few studies have looked at voice quality in specific vocalization types. Lavan et al. (2016) showed that ratings of breathiness provided by trained phoneticians were associated with higher valence ratings in all laughs and with higher arousal ratings in spontaneous, but not in volitional laughs. In contrast, laughs with a higher spectral centroid were rated as slightly higher on both dominance and reward scales in the study by Wood et al. (2017). Apart from laughs, some measures of voice quality have been reported in human screams. Mostly these relate to the presence of nonlinear phenomena (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015; Raine et al., 2018), but Hansen et al. (2017) also report more high-frequency energy and flatter spectral slopes in screams compared to neutral speech. This evidence is not sufficiently detailed to make specific predictions regarding the perceptual effects of breathiness in different call types. The most parsimonious assumption is that the same general acoustic code applies to both speech and all nonverbal vocalizations. On the other hand, there are significant differences between such vocalizations as gasps and screams in their production mechanism (e.g., ingressive or egressive) and general acoustic characteristics (the degree of voicing, pitch, syllable structure), which may affect the perceptual consequences of changes in voice quality.

Instead of the dimensions of valence, arousal, and dominance, participants in Experiment 2 rated the intensity of discrete emotions, aiming to provide a complementary outcome measure that was arguably more intuitive for participants. Three emotional labels were provided for each call type: *Pleased* / *Hurt* / *Surprised* for gasps, *Amused* / *Evil (jeering)* / *Polite* for laughs (adapted from Szameitat et al., 2009 and Wood et al., 2017), *Pleased* / *Hurt* / *Effortful* for

moans, and *Pleased* / *Afraid* / *Aggressive* for screams. These emotions correspond to the most common classifications of acoustically similar vocalizations in earlier studies (Anikin et al., 2018; Anikin, 2019a). They may not cover every possible interpretation, but the objective in Experiment 2 was primarily to contrast responses to the same stimulus after a particular acoustic manipulation. The measure of interest was thus the difference in the weights of particular emotions caused by strengthening or weakening harmonics in the source spectrum of the same prototype sound.

Methods

Stimuli.

The experimental stimuli were 168 modifications of 56 prototype vocalizations, selected from the same source as in Experiment 1 (Anikin & Persson, 2017) and re-synthesized with original, weakened or strengthened harmonics with *soundgen* 1.2.0 (Anikin, 2019a). The prototypes included 10 gasps, 14 laughs, 15 moans or grunts, and 17 screams or high-pitched roars with duration ranging from 0.4 to 3.4 s (Table 2). Out of 56 prototypes, 24 were by men, 29 by women, and 3 by preadolescence children; 15 were also used in Experiment 1. The range of rolloff manipulation was ±4-15 dB/octave.

Table 2. Acoustic characteristics of the synthetic stimuli in Experiment 2.

Call type	Number of stimuli		Acoustic characteristics: mean [range]					
	Total (with noise)	Female / male / child	Rolloff manipu-lation (dB/oct)	Duration (s)	Median voiced syllable (s)	Median f_0 (Hz)	Median HNR* (dB)	Median spectral centroid (kHz)
Gasp	10 (10)	6/4/0	±7.5 [4, 15]	0.8 [0.5, 1.4]	0.6 [0.2, 1.2]	345 [111, 634]	8.4 [0.4, 16.8]	1.8 [1.0, 2.7]
Laugh	14 (14)	6/6/2	±4.4 [4, 6]	1.7 [1.1, 3.1]	0.2 [0.1, 0.7]	517 [167, 846]	10.2 [1.5, 19.1]	1.8 [1.2, 2.9]
Moan	15 (15)	7/8/0	±5.9 [4, 8]	0.8 [0.4, 2.0]	0.6 [0.2, 1.8]	288 [135, 423]	13.3 [3.7, 20.1]	1.3 [0.5, 2.8]
Scream / roar	17 (4)	10/6/1	±6.1 [4, 8]	0.9 [0.3, 1.9]	0.7 [0.2, 1.6]	1205 [295, 3063]	18.4 [11.3, 22.4]	2.4 [0.9, 4.3]

*HNR = harmonics-to-noise ratio, measured based on autocorrelation. HNR and spectral centroid were measured in the synthesized audio files using *soundgen*, and the rest of acoustic characteristics were derived directly from the settings used for synthesizing the stimuli.

Procedure.

The experiment was performed in a web browser and began with training, in which participants rated eight human nonverbal vocalizations in order to become familiar with the rating tool and the nature of stimuli. Following training,

participants rated 81 or 84 synthetic vocalizations each – a subsample of a larger corpus prepared for this study and a companion study on nonlinear phenomena (Anikin, 2019b). The rating tool was a triadic scale designed for rating the proportional weight of three response categories with a single click. Three labels were placed in the corners of an equilateral triangle (Fig. 3), and the weights of these three categories were related to the position of the marker within the triangle via a nonlinear transformation under the constraint that the three weights should sum to 100%. Participants were instructed to set these weights, which were displayed as bars under the triangle.

Participants.

Participants were recruited via https://www.prolific.ac. As in Experiment 1, only submissions that were complete and not obviously faked were accepted and reimbursed. Beyond this initial quality check, no participants were excluded from the analysis because there were no clear outliers among participants in terms of how well their responses agreed with the typical response pattern (see Results). Individual trials with an unusually rapid response time suggestive of a technical problem or accidental clicking were excluded; they represented ~1.4% of data. The response time threshold was higher than in Experiment 1 (2000 vs. 500 ms) because of a different design involving at least two clicks per trial in Experiment 2. The responses of 151 participants provided on average 49 (range 46 to 51) ratings per sound.

Statistical analysis.

The triadic rating scale returns a vector of three weights that sum to one – a so-called simplex – which was modeled with a redundantly-parameterized normal distribution that forces the means for the three categories to sum to one with a softmax transform (Gelman, Bois, & Jiang, 1996):

$$\mu_i = exp(\varphi_i) / (exp(\varphi_1) + exp(\varphi_2) + exp(\varphi_3))$$

for i in $\{1, 2, 3\}$, where μ_i is the mean of the normal distribution for the weight of each category and φ_i is normally distributed with a mean of zero. The φ parameters are not uniquely identifiable, but they do provide valid inference on μ. A corresponding Bayesian model was defined in Stan and extended to include main effects (condition, spectral centroid) and two random intercepts (per participant and per prototype sound) with regularizing priors. A separate model was fit for each of four call types.

Results

The emotion category with the greatest weight was identified for each of 151 participants and for the entire sample. The proportion of sounds for which a given participant assigned the greatest weight to the same category as the majority was then used as a measure of inter-rater agreement. This proportion was approximately normally distributed with no clear outliers and ranged from .26 to .71 (mean = .45), suggesting that most participants understood the experimental procedure and were reasonably consistent in their responses.

The effect of voice quality manipulations can be visualized as the change in the average coordinates within the triangle per prototype sound, as shown in Figure 3. Since participants were instructed to set the relative weights of three emotions, the main analysis focused on how acoustic manipulations affected these weights, not coordinates as such. Similarly to Experiment 1, the outcome was modeled before and after controlling for spectral centroid (Table 3).

Making the voice quality tenser in gasps (less negative rolloff, condition "R+" vs. "R-") made the speaker sound more *Hurt* (+13.4%, 95% CI [9.1, 17.7]) and less *Pleased* (-9.9% [-14.4, -5.5]). For moans, a tense voice was likewise associated with being *Hurt* (+12.8% [9.5, 16.1]) rather than *Pleased* (-12.5% [-17.2, -7.8]). Interestingly, voice quality in moans had no effect on perceived effort (-0.3% [-3.6, 3.0]). Accounting for the spectral centroid did not substantively alter the observed effects on the perceived emotion in either gasps or moans, and spectral centroid had no independent effect on the weight of any emotion after controlling for rolloff, except for making the speaker sound 13.9% [1.4, 25.5] more *Hurt* in gasps (Table 3).

Table 3. Contrasts between the effect of different acoustic manipulations on the weight of emotion categories for each call type: median of posterior distribution (%) and 95% CI.

Model	Contrast	Gasps			Laughs		
		Hurt	Surprised	Pleased	Evil	Polite	Amused
Without SC*	R+ vs. R-	13.4 [9.1, 17.7]	-3.5 [-9.7, 2.7]	-9.9 [-14.4, -5.5]	0.7 [-4.0, 5.6]	-5.1 [-8.5, -1.7]	4.4 [1.1, 7.9]
With SC	R+ vs. R-	13.7 [9.4, 17.9]	-3.7 [-9.9, 2.4]	-9.9 [-14.3, -5.5]	-8.1 [-14.1, -1.2]	3.4 [-2.0, 8.5]	4.6 [-0.8, 9.8]
	SC***	13.9 [1.4, 25.5]	-11.6 [-29.6, 6.3]	-2.0 [-15.5, 11.0]	26.7 [10.3, 39.0]	-25.4 [-37.5, -12.8]	-1.0 [-13.0, 12.0]

Model	Contrast	Moans			Screams		
		Hurt	Effortful	Pleased	Afraid	Aggressive	Pleased
Without SC	R+ vs. R-	12.8 [9.5, 16.1]	-0.3 [-3.6, 3.0]	-12.5 [-17.2, -7.8]	5.1 [2.4, 7.9]	2.4 [-0.3, 5.1]	-7.6 [-11.4, -3.7]
With SC	R+ vs. R-	12.7 [8.3, 17.3]	0.4 [-4.0, 4.8]	-13.1 [-19.2, -7.0]	-3.6 [-10.0, 2.9]	1.9 [-3.1, 7.3]	1.7 [-6.8, 9.9]
	SC	0.6 [-13.8, 14.9]	-3.5 [-17.8, 10.8]	2.8 [-17.1, 22.4]	4.6 [1.5, 7.6]	0.3 [-2.2, 2.5]	-4.9 [-8.6, -1.0]

* SC = spectral centroid, R- = weaker harmonics, R+ = stronger harmonics.
** Cells in **bold** contain 95% CIs that exclude or nearly exclude zero. This is merely a visualization aid, not significance testing.
*** Effect over the observed range of 530 to 4260 Hz, controlling for rolloff manipulation.

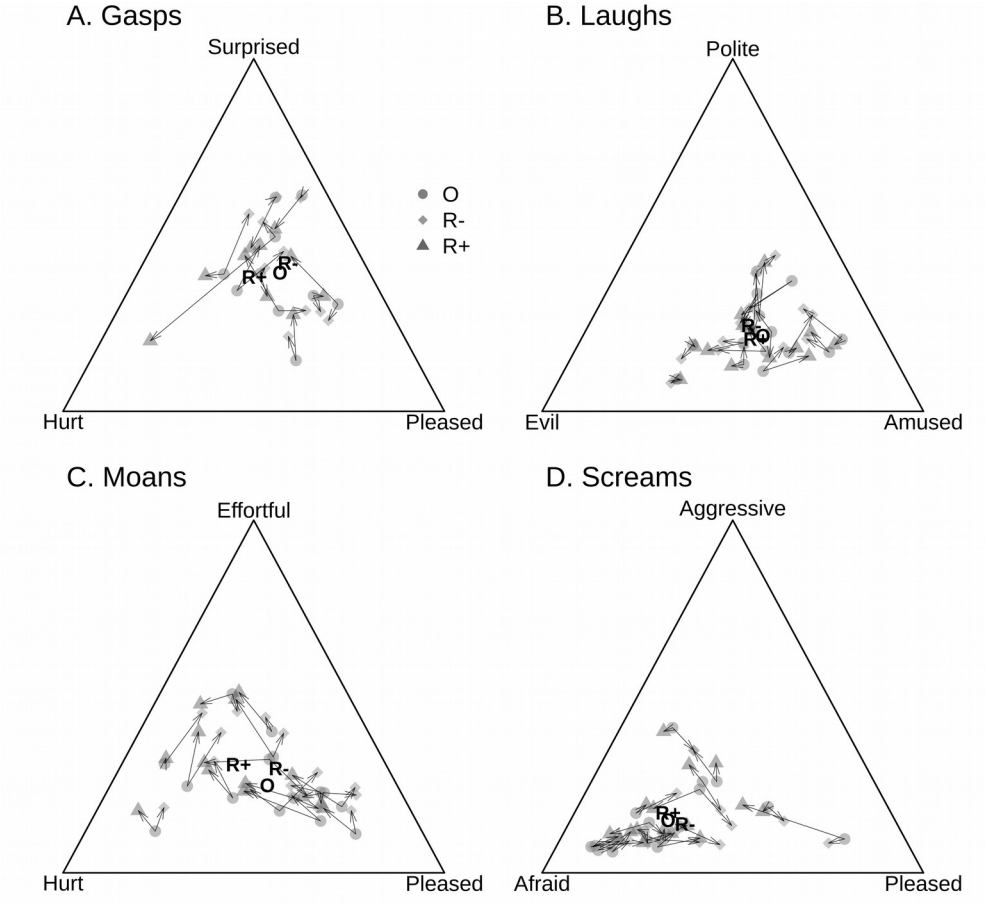

Fig. 3. Mean coordinates representing the perceived emotion of different call types and acoustic manipulations. Labels in bold show the average for all sounds, while individual symbols and arrows show the effect of manipulations for each prototype sound. O = original, R- = less energy in harmonics (breathy), R+ = more energy in harmonics (tense).

Tenser voice quality in laughs enhanced the impression that the speaker was genuinely *Amused* (+4.4% [1.1, 7.9]) rather than merely *Polite* (-5.1% [-8.5, -1.7]). However, this effect was relatively small and disappeared after controlling for spectral centroid (Table 3). Interestingly, a higher spectral centroid was strongly associated with sounding *Evil* (+26.7% [10.3, 39.0]) rather than *Polite* (-25.4% [-37.5, -12.8]), presumably because strongly aspirated giggles with little voicing are considered impolite or malicious.

Screams with stronger harmonics were associated with being *Afraid* (+5.1% [2.4, 7.9]) rather than *Pleased* (-7.6% [-11.4, -3.7]). Since screams were synthesized

with little or no noise component, rolloff of harmonics was the sole determinant of the spectral shape, and therefore the effects of rolloff manipulation and spectral centroid in screams could not be separated in multiple regression, leading to uncertain results when both were entered simultaneously (Table 3).

Discussion

Experiment 2 aimed to verify and nuance the findings from Experiment 1 by increasing the number of stimuli and using discrete emotions instead of the dimensions of valence, arousal, and dominance. The results confirmed that a shift in voice quality from breathy to tense was associated with more aversive emotions, but with interesting differences between call types.

The clearest picture emerged for the most ambiguous of the tested call types – gasps and moans. Tense voices with strong harmonics and little aspiration noise were perceived as considerably more aversive than breathy voices with weak harmonics. In effect, manipulating only the voice quality, without changing the intonation or any other acoustic characteristic, was enough to turn a gasp or moan of pleasure into pain. Since this effect of voice quality persisted after controlling for spectral centroid, listeners appeared to attend specifically to breathiness, and not simply to the amount of high-frequency energy in the spectrum.

In screams, the manipulation did not create a breathy voice as such, since no aspiration noise was added. Furthermore, screams were predominantly interpreted as an expression of fear, and this limited variation in responses partly masked the effect of manipulations. Nevertheless, strengthening upper harmonics relative to the fundamental frequency noticeably shifted the interpretation of screams from pleasure to fear, as predicted.

As for laughs, making the voice quality tense within one particular sound - without changing any prosodic characteristics - made the speaker sound slightly more amused rather than merely polite. When comparing different laughs, on the other hand, a higher spectral centroid (a measure of timbral brightness) was strongly associated with sounding malicious. Laughs come in a great variety of acoustic forms and contain a large amount of aspiration noise, complicating the relationship between summary measures of spectral shape, such as spectral centroid, and glottal source spectrum. This is possibly the reason for seemingly contradictory reports in previous correlational studies, which found that breathy laughs scored higher on both arousal and valence (Lavan et al., 2016), but also that laughs with a higher spectral centroid were rated as more rewarding (Wood et al., 2017). Taking the present results at face value, tensing the voice quality in laughs does not make them more negative - unlike other analyzed vocalizations - but it appears to

enhance the perception of genuine amusement. Because high-intensity emotional expressions are often perceived as more authentic (Anikin & Lima, 2018; Lavan et al., 2016), this is in line with the association of tense voice quality with higher perceived arousal in Experiment 1. Considering that the manipulation effect for laughs was relatively weak and uncertain, however, it should be treated with caution. In addition, laughs present a formidable challenge to manual parametric synthesis because of their complex and dynamic spectrotemporal characteristics, and the perceived authenticity of synthetic laughs was lower than for other vocalizations in a previous validation study (Anikin, 2019a). It is therefore possible that the synthesis of laughs was not successful enough to analyze the effects of voice quality. The relatively short and noisy syllables that laughs consist of also make changes in laryngeal source harder to detect because breathiness is more salient in longer sustained syllables.

General Discussion

Two experiments were carried out to test the perceptual consequences of modifying the laryngeal voice quality in otherwise identical synthetic nonverbal vocalizations. The manipulation consisted in changing the rolloff of harmonics in source spectrum, which had two effects: it made upper harmonics more or less pronounced relative to the fundamental frequency and simultaneously changed the amplitude of aspiration noise relative to the voiced component. Perceptually, this manipulation approximately corresponds to shifting the voice quality along the tense-breathy continuum. As predicted, breathiness was associated with less intense and more pleasant emotions, particularly for those vocalizations that can be either positive or negative in valence. Some implications of these findings are discussed below.

In line with previous reports based on conventional measures of spectral shape (Briefer, 2012; Lima et al., 2013; Lingle et al., 2012; Schröder et al., 2001), stronger harmonics were associated with higher arousal ratings. This is not surprising, since physiologically this change in voice quality is caused by greater pharyngeal constriction (Briefer, 2012) and higher subglottal pressure (Stevens, 2000), both of which are associated with an active, aroused state. In addition, increasing the strength of harmonics had a negative effect on valence ratings in Experiment 1. This finding was strongly confirmed in Experiment 2, particularly for intrinsically ambiguous vocalizations such as gasps and moans. For example, a breathy gasp or moan with a strong f_0 and weak harmonics (breathy voice) was likely to be interpreted as a sign of pleasant surprise, whereas an otherwise identical sound with stronger harmonics and less aspiration noise (tense voice)

was interpreted as an expression of pain. Voice quality had a smaller effect on vocalizations that were interpreted as predominantly hedonistic (laughs) or predominantly aversive (screams). This tallies with the earlier observation that the distribution of energy in the spectrum correlates with valence only in the more ambiguous primate vocalizations (Fichtel et al., 2001). More generally, it calls for caution when generalizing acoustic observations across call types, since perceptual effects of acoustic features may be vocalization-specific (Linhart et al., 2015).

In addition to testing the relevance of laryngeal voice quality to the communication of emotion specifically in nonverbal vocalizations, an important novelty of the chosen experimental approach lies in the ability to distinguish between the contribution of "glottal source" (or rather, its synthetic counterpart) and the general balance of low- and high-frequency energy in the spectrum. A non-specific measure of timbral brightness (spectral centroid) appeared to contribute to some of the observed effects, but did not fully account for them. In line with speech research (Garellek et al., 2016), this suggests that listeners are sensitive to the relative strength of individual harmonics in nonverbal vocalizations when they decide what emotion the speaker is experiencing. As a result, relatively subtle modifications of voice quality – changes that would be nearly invisible to a conventional acoustic analysis – are sufficient to cause a major effect on perceived emotion, sometimes "flipping" the valence of a vocalization.

The reported manipulation of voice quality was somewhat simplified: in reality breathiness also includes other acoustic characteristics that were not modeled in this study, such as increased formant bandwidth and additional zero-pole pairs in the source spectrum due to coupling with supralaryngeal resonators (Gobl & Ní Chasaide, 2010). In addition, the manipulation affected all harmonics in source spectrum, whereas speech research suggests that humans are sensitive to the difference between specific harmonics, the overall spectral slope, and high-frequency noise excitation (Garellek, Samlan, Gerratt, & Kreiman, 2016; Kreiman, Gerratt, & Antoñanzas-Barroso, 2007; Kreiman et al., 2014). On the other hand, the narrow range of f_0 and moderate vocal effort typical of speech are different from the conditions of voice production in nonverbal vocalizations such as screams or roars, making it problematic to apply linguistic terminology or speech-specific measures of source spectrum. Furthermore, the role of laryngeal voice quality in nonverbal vocalizations in largely terra incognita, and the results reported here are only a preliminary investigation that will need to be confirmed and elaborated in future studies.

An interesting question for follow-up research is whether a tense voice with strong harmonics is intrinsically associated with intense and unpleasant affective states, or whether this effect is due to changes in perceived loudness and pitch. Although

pitch is usually considered to be the perceptual equivalent of the fundamental frequency, it may in fact depend on other spectral characteristics (McPherson & McDermott, 2018), including voice quality. In particular, a tense voice literally sounds higher than a breathy voice (Kuang, Guo, & Liberman, 2016). Likewise, sounds with more high-frequency energy are subjectively experienced as louder because human hearing is more sensitive to high frequencies (Fastl & Zwicker, 2006), and loudness can enhance the impression of high activation states (Yanushevskaya, Gobl, & Ní Chasaide, 2013). It is therefore possible that strengthening the harmonics rather mechanistically makes the stimulus appear louder, brighter, and more high-pitched, thus enhancing its low-level perceptual salience to the auditory system.

In addition, listeners may interpret vocalizations with strong harmonics as originally produced with high vocal effort - loudly and close to the upper limit of the speaker's pitch range (Kuang, Guo, & Liberman, 2016; but see Bishop & Keeting, 2012). Because louder vocalizations tend to have stronger harmonics (Traunmüller & Eriksson, 2000) and therefore a higher peak frequency (Stout, 1938; Gustison & Townsend, 2015), some information about the loudness of the original utterance is still available to listeners even from recordings with normalized amplitude. This estimate of the original speaker's vocal effort may then be taken into account when interpreting the vocalization. For example, listeners may expect that a moan of pleasure will be produced in a quieter and breathier voice than a moan of pain, that a person in pain will scream with greater vocal effort and thus a tenser voice than a person who is delighted, and so on. Indeed, hedonistic vocalizations, such as moans of pleasure, are more likely to occur in intimate, close-range contexts, whereas aversive vocalizations, such as screams of fear, are meant to broadcast the signal to distant observers, call for help, or warn others about the presence of predators, necessitating high vocal effort (Gustison & Townsend, 2015). Similarly, listeners in perceptual experiments may assume, often mistakenly, that intense emotional expressions are more likely to be aversive rather than hedonistic (Anikin & Persson, 2017). On the other hand, the association of tense voice quality with feeling more genuinely amused in laughs, although relatively uncertain, could indicate that higher perceived vocal effort is simply interpreted as a sign of greater emotion intensity, making purely hedonistic vocalizations such as laughs more positive, and aversive or ambiguous vocalizations more negative. It will be a productive avenue for future research to look more closely at the differences between vocalization types when investigating the link between voice acoustics and emotion, not least because it could elucidate the cognitive mechanisms involved.

To end on a methodological note, the present results further underline the importance of estimating glottal source when analyzing field recordings of non-speech vocalizations. In speech research it is common to compare the amplitudes

of the first few harmonics with each other and with the harmonic nearest to the first formant or a specific frequency as an indirect measure of spectral source (Garellek et al., 2016; Gobl & Ní Chasaide, 2010; Kreiman et al., 2007; Kreiman et al., 2014), but this may be inappropriate for nonverbal vocalizations and animal calls with an extreme range of f_0 that routinely crosses formant frequencies. An interesting option is to estimate how high detectable harmonics reach in the spectrum (cf. "frequency range" in Fichtel et al., 2001; Fichtel & Hammerschmidt, 2002; Hammerschmidt & Jurgens, 2007), although high levels of jitter and noise will affect this measure. Above all, it is advisable to extract multiple measures of spectral shape instead of a single descriptive, such as mean or peak frequency, and to provide access to the original recordings.

References

Airas, M., & Alku, P. (2006). Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, *63*(1), 26-46.

Anikin, A. (2019a). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods, 51*(2), 778-792.

Anikin, A. (2019b). The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics*, 1-22. doi: 10.1080/09524622.2019.1581839

Anikin, A., & Lima, C. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology*, *71*(3), 622-641.

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

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior*, *42*(1), 53-80.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051-2056.

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

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.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100-1112.

Brady, M. C. (2005). Synthesizing affect with an analog vocal tract: glottal source. In *Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop* (pp. 25-26).

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1-20.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1171-1178.

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.

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, *98*(1), 88-98.

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. (2014). Glottal source processing: From analysis to applications. *Computer Speech & Language*, *28*(5), 1117-1138.

Fant, G. (1960). *Acoustic theory of speech perception*. Mouton, The Hague.

Fastl, H., and Zwicker, E. (2006). *Psychoacoustics: facts and models, 2nd ed.* (Vol. 22). Springer Science & Business Media.

Fichtel, C., & Hammerschmidt, K. (2002). Responses of redfronted lemurs to experimentally modified alarm calls: Evidence for urgency based changes in call structure. *Ethology*, *108*(9), 763-778.

Fichtel, C., Hammerschmidt, K., & Jürgens, U. (2001). On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. *Behaviour*, *138*(1), 97-116.

Garellek, M., Samlan, R., Gerratt, B. R., & Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, *139*(3), 1404-1410.

Gelman, A., Bois, F., & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, *91*(436), 1400-1412.

Gobl, C., & Ní Chasaide, A. (2010). "Voice source variation and its communicative functions". In Hardcastle, W. J., Laver, J., & Gibbon, F. E. (Eds.). *The handbook of phonetic sciences* (2nd ed.) (pp. 378-423). Singapore: Wiley-Blackwell.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*(1-2), 189-212.

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322-1336.

Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, *54*(3), 414-429.

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.

El Haddad, K., Çakmak, H., Sulír, M., Dupont, S., & Dutoit, T. (2016). Audio affect burst synthesis: A multilevel synthesis system for emotional expressions. In *2016 24th European Signal Processing Conference (EUSIPCO)* (pp. 1158-1162).

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, *21*(5), 531-540.

Hansen, J. H., Nandwana, M. K., & Shokouhi, N. (2017). Analysis of human scream and its impact on text-independent speaker verification. *The Journal of the Acoustical Society of America*, *141*(4), 2957-2967.

He, L., Lech, M., & Allen, N. (2010). On the importance of glottal flow spectral energy for the recognition of emotions in speech. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 2346-2349).

Hogstedt, G. (1983). Adaptation unto death: function of fear screams. *The American Naturalist*, *121*(4), 562-570.

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the XIVth international congress of phonetic sciences* (pp. 2029-2032). San Francisco: University of California, Berkeley.

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research*, *50*, 595-610.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1). doi:10.3989/loquens.2014.009.

Kuang, J., Guo, Y., & Liberman, M. (2016). Voice quality as a pitch-range indicator. In *Proceeding of Speech Prosody* (pp. 1061-1065).

Laukkanen, A. M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics*, *24*(3), 313-335.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, *40*(2), 133-149.

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.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology*, *58*(5), 698-726.

Linhart, P., Ratcliffe, V. F., Reby, D., & Špinka, M. (2015). Expression of emotional arousal in two different piglet call types. *PloS one*, *10*(8), e0135414.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one*, *9*(3), e90779.

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour*, *2*(1), 52-66.

Mittal, V. K., & Yegnanarayana, B. (2014). Study of changes in glottal vibration characteristics during laughter. In *Fifteenth Annual Conference of the International Speech Communication Association* (pp. 1777-1781).

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097-1108.

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, *87*(1), 93-98.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417-435.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283-296.

Raine, J., Pisanski, K., Simner, J., & Reby, D. (2018). Vocal communication of simulated pain. *Bioacoustics*, 1-23. doi: 10.1080/09524622.2018.1463295

Ross, M. D., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology*, *19*(13), 1106-1111.

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.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2002). Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice*, *16*(4), 509-529.

Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In J. Tao and T. Tan (Eds.) *Affective information processing* (pp. 111-126). London: Springer.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology* (pp. 1-4). Sep 3-7; Aalborg, Denmark.

Schubert, E., Wolfe, J., & Tarnopolsky, A. (2004). Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois* (pp. 112-116).

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Stout, B. (1938). The harmonic structure of vowels in singing in relation to pitch and intensity. *The Journal of the Acoustical Society of America*, *10*(2), 137-146.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., & Sterr, A. (2009). Differentiation of emotions in laughter at the behavioral level. *Emotion*, *9*(3), 397-405.

Titze, I. R. (2008). Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*(4), 1902-1915.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438-3451.

Waaramaa, T., Laukkanen, A. M., Airas, M., & Alku, P. (2010). Perception of emotional valences and activity levels from vowel segments of continuous speech. *Journal of Voice*, *24*(1), 30-38.

Wood, A., Martin, J., & Niedenthal, P. (2017). Towards a social functional account of laughter: Acoustic features convey reward, affiliation, and dominance. *PloS one 12*(8), e0183811.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2013). Voice quality in affect cueing: does loudness matter? *Frontiers in psychology*, *4*, 335, 1-14.

Paper VII

The link between auditory salience and emotion intensity

Andrey Anikin

Division of Cognitive Science, Lund University, Sweden

Abstract

To ensure that listeners pay attention and do not habituate, emotionally intense vocalizations may be under evolutionary pressure to exploit processing biases in the auditory system by maximizing their bottom-up salience. This "salience code" hypothesis was tested using 128 human nonverbal vocalizations representing eight emotions: amusement, anger, disgust, effort, fear, pain, pleasure, and sadness. As expected, within each emotion category salience ratings derived from pairwise comparisons strongly correlated with perceived emotion intensity. For example, while laughs as a class were less salient than screams of fear, salience scores almost perfectly explained the perceived intensity of both amusement and fear considered separately. Validating self-rated salience evaluations, high- vs. low-salience sounds caused 25% more recall errors in a short-term memory task, whereas emotion intensity had no independent effect on recall errors. Furthermore, the acoustic characteristics of salient vocalizations were similar to those previously described for non-emotional sounds (greater duration and intensity, high pitch, bright timbre, rapid modulations, and variable spectral characteristics), confirming that vocalizations were not salient merely because of their emotional content. The acoustic code in nonverbal communication is thus aligned with sensory biases, offering a general explanation for some non-arbitrary properties of human and animal high-arousal vocalizations.

Keywords: emotion, arousal, salience, sensory bias, nonverbal vocalizations

Introduction

Like other animals, humans communicate with a variety of nonverbal signals: we smile and frown at each other, giggle when tickled, scream when frightened, and so on. One way to describe these behaviors is to focus on the form of these signals. For instance, identifying the acoustic properties of nonverbal vocalizations that are associated with particular emotions (Anikin & Persson, 2017; Lima, Castro, & Scott, 2013; Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001) helps to clarify how meaning is encoded acoustically in these signals. In addition to such proximate or *how* questions, we can ask *why* signals assume this particular form, and not another (Tinbergen, 1963). Could we just as easily have evolved to laugh in sadness and to smile ferociously at our opponents? In this article I argue that high-intensity nonverbal vocalizations of humans, and presumably also of other animals, have non-arbitrary acoustic properties that make them maximally salient to the audience. In the rest of the text, I develop this "salience code" hypothesis of emotion expression, discuss the available evidence, and report the results of three perceptual experiments designed to test it.

Sensory biases

Researchers have long known that there are many similarities between the vocal signals of distantly related animal species. For example, in many mammals and birds aggressive calls are low-pitched and harsh, while friendly or appeasing calls are high-pitched and tonal (Morton, 1977; Ohala, 1984). Although not universal, this frequency code explains a large variety of vocal behaviors (August & Anderson, 1987; Briefer, 2012) and has a plausible cognitive mechanism, namely a reliably developing cross-modal association between auditory frequency and size (Hamilton-Fletcher et al., 2018; Spence, 2011).

Another dimension of affect (Russell, 1980), and a very common way to characterize both animal (Briefer, 2012; Filippi et al., 2017) and human (Lassalle et al., 2019; Lima et al., 2013; Lingle et al., 2012; Schröder et al., 2001) vocalizations, is the degree of physiological activation (arousal) or emotion intensity that they express. Acoustic changes associated with high arousal include increased duration with shorter pauses between syllables, high fundamental frequency (pitch), more high-frequency energy in the spectrum (bright timbre), variable pitch and sound intensity, and low harmonicity (a harsh voice quality;

Briefer, 2012; Fitch, Neubauer, & Herzel, 2002; Lima et al., 2013; Lingle et al., 2012; Schröder et al., 2001). Confirming the predictable nature of high-arousal calls, participants recognize high emotion intensity not only in foreign languages (Lassalle et al., 2019), but even in calls of unfamiliar animal species (Filippi et al., 2017).

Many of the acoustic properties associated with high-intensity calls are direct consequences of the physiological changes triggered in the vocal apparatus by a powerful emotion. For example, activating laryngeal muscles and elevating subglottal pressure raises the pitch and makes the voice louder and brighter due to boosted harmonics (Gobl & Ní Chasaide, 2010). As vocal effort increases further, the vibration of vocal folds becomes irregular, causing unpredictable pitch jumps and other nonlinear phenomena that make the voice harsh (Fitch et al., 2002). The exact mechanisms linking emotion with vocal production are not yet sufficiently well understood (Gobl & Ní Chasaide, 2010), but they probably involve direct motor control of laryngeal and respiratory musculature as well as slower effects mediated by the autonomous nervous system and hormonal levels (LeDoux, 2012; Scherer, 1986).

Pushing the question one step further, however, what evolutionary pressures might have caused intense emotional states to have these specific effects on the voice? Because it is usually in the signaler's interest to make high-intensity calls easy for conspecifics to detect and difficult to ignore, such calls may be under selective pressure to optimally activate the auditory system of receivers. This line of reasoning is known in biology as the *sensory bias* or *sensory exploitation* hypothesis (Ryan, 2013). It was initially applied to the relatively narrow problem of sexual selection of male calls and visual ornaments to match the sensitivity of female peripheral receptors. The hypothesis remains controversial in this original context (Fuller, Houle, & Travis, 2005; Ron, 2008), but the logic behind it is more general. Perception has evolved to promote survival and reproduction, rather than to represent the external world accurately (Neuhoff, 2018), and the resulting perceptual distortions or biases – such as enhanced sensitivity to particular sensory features – can be exploited by signalers.

Naturally, in the long run signal production and perception must coevolve. In the case of vocalizations, this is particularly manifest for specialized adaptations such as echolocation that simultaneously place new demands on both vocal production and auditory processing. In general, the auditory system is tuned for optimal discrimination of biologically relevant sounds (Woolley, Fremouw, Hsu, & Theunissen, 2005), and adaptations such as improved hearing in the most relevant frequency ranges have been described in mammals at large (Stebbins, 1980) as well as in individual species (Theunissen & Elie, 2014). However, the key assumption behind the sensory bias hypothesis is that production often outpaces

perception in terms of the speed and flexibility of its evolution. For example, reptiles and birds neither hear nor produce sounds above approximately 10 kHz because their middle ear is anatomically incapable of handling high frequencies; in contrast, the independently evolved mammalian middle ear is well-equipped to exploit this frequency range, and many mammals have evolved ultrasonic calls for communication or echolocation (Köppl, 2009; Stebbins, 1980).

Apart from evolvability constraints, the auditory system must perceive and localize (Köppl, 2009) not only vocalizations, but also environmental sounds, which have very different acoustic properties (Singh & Theunissen, 2003; Smith & Lewicki, 2006). This relative inertia of auditory perception is illustrated by the fact that hominins rapidly evolved very specific neurological adaptations for speech, while the auditory system remained similar to that of other apes (Fitch, 2018). In fact, human speech combines the acoustic features of environmental sounds and vocalizations, possibly to match the preferred input to the auditory system (Smith & Lewicki, 2006; Theunissen & Elie, 2014).

In sum, with regard to meaningful acoustic variation within a species' repertoire and on shorter time scales, vocal production is more likely to adapt to the auditory perception than the other way round. Accordingly, vocalizations associated with intense emotional states may have been under evolutionary pressure to "tickle the brain" – that is, to involuntarily attract and hold the attention of conspecifics.

Salience

The capacity of sensory stimuli for commanding attention in a bottom-up, involuntary fashion is known as their *salience* and distinguished from the voluntary, top-down component of attention. The last decade has seen active development of experimental paradigms and computer algorithms for the evaluation of visual and, more recently, auditory salience. The most general finding is that salient sounds tend to be highly variable and unpredictable. Simply maintaining high intensity or periodically varying the sound is not enough to make it salient: even loud bursts become less surprising after a few repetitions (Kaya & Elhilali, 2017; Huang & Elhilali, 2017). In contrast, attention is drawn to unpredictable acoustic changes: a tone presented against a noisy background (Kayser, Petkov, Lippert, & Logothetis, 2005), a random tone sequence (Southwell et al., 2017), a sudden change in overall amplitude (Kim, Lin, Walther, Hasegawa-Johnson, & Huang, 2014) or in a particular frequency range (Kayser et al., 2005; Kaya & Elhilali, 2014), rapid amplitude modulation or "roughness" (Zhao et al., 2018), etc. There is also evidence that acoustic changes in a particular direction are experienced as more salient. For example, looming stimuli with

increasing intensity are more salient than receding ones, perhaps as a defense mechanism for rapid detection of approaching hazards (Neuhoff, 2018; Tajadura-Jiménez, Väljamäe, Asutay, & Västfjäll, 2009). More generally, salient acoustic events are associated with sudden increases – rather than decreases – in loudness, pitch, spectral centroid (timbral brightness), harmonicity, and spectral bandwidth (Huang & Elhilali, 2017). In addition, long sounds are generally more salient than short ones (Kayser et al., 2005).

Although the research on auditory salience is still at an early stage, these characteristics of salient acoustic events appear to have many parallels with the most common characteristics of high-intensity vocalizations. As described above, greater duration, loudness, pitch, and a bright timber are all associated with high general activation or arousal in both speech and animal vocalizations. High-intensity vocalizations also tend to be more variable, particularly due to the prevalence of nonlinear vocal phenomena, which elicit a stronger response from the audience (Reby & Charlton, 2012) and prevent habituation (Karp, Manser, Wiley, & Townsend, 2014). In fact, crying babies and barking dogs are self-reported to be harder to ignore than specially designed sirens and fire alarms (Ball & Bruck, 2004), and it has been suggested that the acoustics of dog barks is functionally optimal for attracting the attention of humans (Jégh-Czinege, Faragó, & Pongrácz, 2019). There is thus indirect evidence that sounds produced in highly aroused states possess many acoustic characteristics associated with auditory salience. To the best of my knowledge, this study is the first to test this salience code hypothesis empirically.

The present study

The chosen emotional sounds were human nonverbal vocalizations. Unlike speech, they are free from semantic confounds and language-specific phonological constraints. At the same time, nonverbal vocalizations express a wide range of emotions and related states such as pain (Anikin & Persson, 2017; Lima et al., 2013; Maurage, Joassin, Philippot, & Campanella, 2007). Moreover, similar sounds are produced by modern-day apes (Davila-Ross, Owren, & Zimmermann, 2009; Lingle, Wyman, Kotrba, Teichroeb, & Romanow, 2012), and at least some were presumably present in our prelinguistic hominin ancestors. To some extent, the results may therefore generalize to both speech and vocalizations of non-human mammals.

To avoid circularity, bottom-up salience of these vocalizations had to be measured independently of their emotional content. In contrast to vision research, where eye tracking gives a ground-truth measure of bottom-up salience (Kaya & Elhilali,

2016), there is no generally accepted method for measuring the salience of sounds. Various approaches have been explored: annotation of soundscapes (Kim et al., 2014), detection of targets against noisy backgrounds (Kayser et al., 2005, Kaya & Elhilali, 2014), distraction from a working memory task (Vachon, Labonté, & Marsh, 2017), pairwise comparisons (Kayser et al., 2005), pupil dilation (Huang & Elhilali 2017), microsaccade inhibition (Zhao et al., 2018), etc.

In this study two independent samples of participants provided two measures of salience: (1) *self-reported salience* based on explicit pairwise comparisons and (2) *objective salience* operationalized as the drop in performance in a short-term memory task. These two salience measures were then compared with the ratings of emotion intensity and the acoustic characteristics of experimental stimuli. The two key predictions were (1) that there would be a close match between perceived emotion intensity and auditory salience, and (2) that the acoustic characteristics of salient nonverbal vocalizations would be similar to the previously described predictors of salience in mixed synthetic and environmental sounds – that is, that the responsible acoustic features are a property of human auditory perception in general, rather than specific to emotional vocalizations.

Methods

Stimuli

The experimental stimuli (N = 128) were human nonverbal vocalizations of amusement, anger, disgust, effort, fear, pain, pleasure, and sadness: 8 male + 8 female = 16 sounds in each category (Table 1). Of these 128 sounds, 118 were selected from a corpus of spontaneous, non-staged vocalizations obtained from social media (Anikin & Persson, 2017), some of which were shortened to keep the duration of all stimuli under 3 s. The selected vocalizations were relatively free from background noise, and although audio quality was more variable than in studio recordings, a machine learning algorithm classified these sounds based on acoustic measurements with an accuracy on a par with human raters in the validation study (Anikin & Persson, 2017).

Spontaneous sounds of anger, fear, and sadness consisted of high-intensity vocalizations such as high-pitched screams. To maintain a variety of intensity levels within each emotion, ten milder, intentionally produced vocalizations were added from another corpus (Maurage et al., 2007). To control for the tendency of high-pitched sounds to appear louder, the subjective loudness was estimated in

sone (assuming sound pressure level of 70 dB above a reference value of $2x10^{-5}$ Pa) within each 200-ms frame throughout the duration of each sound. As a compromise between normalizing for peak and average loudness, the stimuli were normalized for the 75% quantile of per-frame loudness values. Loudness estimation and normalization, as well as the acoustic analysis of experimental stimuli, were performed using the R package *soundgen* (Anikin, 2019).

Table 1. Experimental stimuli.

Emotion	*N* (male/female)		Description	Acoustics: mean [range]	
	Spont.*	Acted**		Duration, ms	Pitch, Hz
Amusement	8/8	-	Laughs	1780 [745, 2818]	465 [164, 902]
Anger	6/6	2/2	Roars, growls, screams, vowels	795 [419, 1737]	417 [141, 1170]
Disgust	8/8	-	Grunts, groans	745 [443, 1535]	242 [112, 560]
Effort	8/8	-	Grunts, growls, roars	943 [448, 1945]	361 [188, 647]
Fear	6/6	2/2	Screams, vowels	1207 [432, 2468]	902 [201, 1870]
Pain	8/8	-	Screams, groans	1257 [474, 2678]	607 [259, 1580]
Pleasure	8/8	-	Moans, groans, sighs	1201 [542, 2758]	290 [100, 614]
Sadness	7/7	1/1	Cries, vowels	1858 [698, 2707]	374 [137, 702]

*Anikin & Persson (2017) **Maurage et al. (2007)

Acoustic analysis

Each sound was described with 22 acoustic parameters (italicized below) chosen to facilitate comparisons with previous research (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015; Huang & Elhilali, 2017):

(1) General descriptives: *duration*, *RMS amplitude* (mean and SD across all frames), *loudness* in sone (mean and SD across all frames).

(2) Intonation: to ensure optimal accuracy, all intonation contours were extracted manually using *soundgen*'s interactive pitch editor *pitch_app* and summarized as mean *pitch* (on a musical scale) and its standard deviation (in semitones). Harmonics-to-noise ratio (HNR or *harmonicity* – a measure of tonality, the ratio of energy in the harmonic component and aperiodic noise) was estimated in each frame as the peak of autocorrelation function and summarized by its mean and SD across frames.

(3) Auditory spectrum (Fig. 1A): the input for this analysis was a mel-transformed auditory spectrogram produced with *melfcc* function from *tuneR* package (Ligges, Krey, Mersmann, & Schnackenberg, 2018) with 20 ms Hamming windows, a step of 5 ms, and 128 frequency bins on the mel scale. Following Huang & Elkilali (2017), the spectrum of each frame was summarized by its *centroid* (center of gravity, a measure of timbral brightness), *bandwidth* (weighted distance from the centroid), *flatness* (a ratio of geometric and arithmetic means), and *irregularity* (average difference in strength between adjacent frequency bins). Each of these four measures was then averaged across all frames (mean and SD).

(4) Modulation spectrum (Fig. 1B): the result of a two-dimensional Fourier transform of a spectrogram, modulation spectrum represents the sound as a combination of amplitude and frequency modulation (Singh & Theunissen, 2003). It was extracted for an entire vocalization, rather than per frame, using the *modulationSpectrum* function in *soundgen*. The result was summarized by the centroid of *amplitude modulation* (2-32 Hz) and *frequency modulation* (0-10 cycles/kHz; see Huang & Elkilali, 2017) as well as by the proportion of high-frequency amplitude modulation in the "*roughness*" range of 30-150 Hz (Arnal et al., 2015; Zhao et al., 2018).

(5) Self-similarity matrix (Fig. 1C): this transformation represents the internal structure of a sound by comparing its parts to each other. The auditory spectrogram was divided into 40-ms windows, and cosine similarity was calculated for all window combinations, resulting in a square self-similarity matrix (SSM). Acoustic *novelty* was then calculated by sliding a 200-ms Gaussian checkerboard matrix along SSM's diagonal (Foote, 2000). Peaks in the novelty contour correspond to sudden changes in spectral structure on the time scale determined by the length of the checkerboard filter. The novelty contour of each sound was extracted using the *ssm* function in *soundgen* and summarized by its mean and SD.

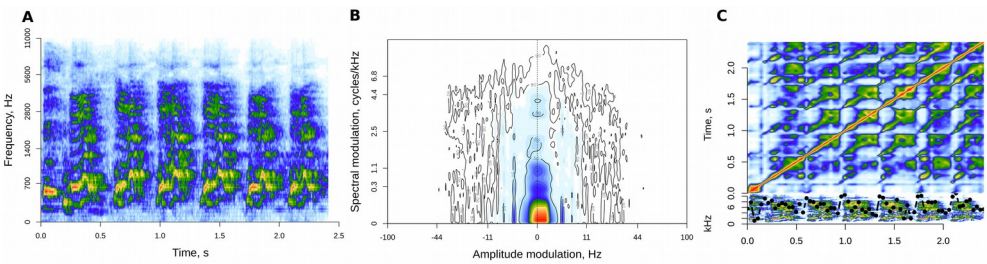

Fig. 1. Three representations of the same laugh: mel-scaled auditory spectrogram (A), modulation spectrum (B), and self-similarity matrix (C). SSM's lower panel shows the novelty contour (dotted black line) over the input spectrogram.

Procedure

Three experiments were performed to measure:

(1) Emotion intensity. Participants rated the intensity of 128 sounds arranged into eight blocks, one block for each emotion. For example, for the 16 laughs in the “amusement” block the question was: *"How intense is the amusement"?* The order of both blocks and sounds within each block was randomized for each participant. The intensity was rated on a continuous horizontal scale ranging from *mild* to *extreme*. The sound could be repeated as many times as needed, and there was no time limit for responding.

(2) Self-reported salience. A new sample of participants were presented with two sounds and asked: “*Which sound is more attention-grabbing and harder to ignore?*” They could then select one of the sounds or answer *similar* to indicate a tie. These comparisons were performed among all 128 sounds, rather than within one emotion category. Emotion was not mentioned in the instructions to avoid priming the subjects for thinking in terms of emotion intensity or arousal. The sounds could be replayed as many times as needed, and there was no time limit for responding. The order of sound pairs was randomized for each participant under the constraint that the same stimulus should never appear in two consecutive pairs. Each participant rated 100 sound pairs.

(3) Objective salience. A new sample of participants had to memorize and repeat a spoken sequence of six numbers presented through earphones in one channel, while ignoring a distractor vocalization in the other channel. The degree to which different distractor vocalizations disrupted accurate recall in this dichotic-listening short-term memory task was regarded as an implicit measure of their "objective" salience. A target sequence consisted of six non-repetitive single-digit numbers, each spoken by a random computer-synthesized voice and played with unpredictable timing from a randomly chosen stereo channel (left or right). About 10% of sequences did not contain a distractor, and the rest contained one vocalization timed to coincide with the onset of one of the six spoken digits. Each distractor vocalization could overlap with one to four targets, but never in the same channel (Fig. 2). To reduce the pop-out effect of any sound against silence, both targets and distractors were embedded in a background of steady street noise at 13 dB below the level of targets and distractors. High accuracy (4.8 out of 6 digits on average, all six correct in over half the trials) proves that targets could be clearly heard over both background and distractors. Fifty unique sequences were generated with each distractor

vocalization to cancel out possible effects of the timing and channel, incidental qualities of target sequences, and variations in the background noise. Pilot pre-testing revealed that estimating implicit salience required much more data compared to emotion intensity and self-reported salience. Accordingly, only half of the 128 vocalizations were tested (8 out of 16 in each emotion category, selected to preserve the full range of emotion intensity). Each trial began with the screen showing a fixation cross, while the participant listened to the sequence of digits followed by a five-second retention period. Five seconds after the last target a number pad appeared, and the participant had to enter target digits in the original order. Digits flashed green for correct responses and red for errors. The average accuracy was updated after each block and continuously displayed on the screen.

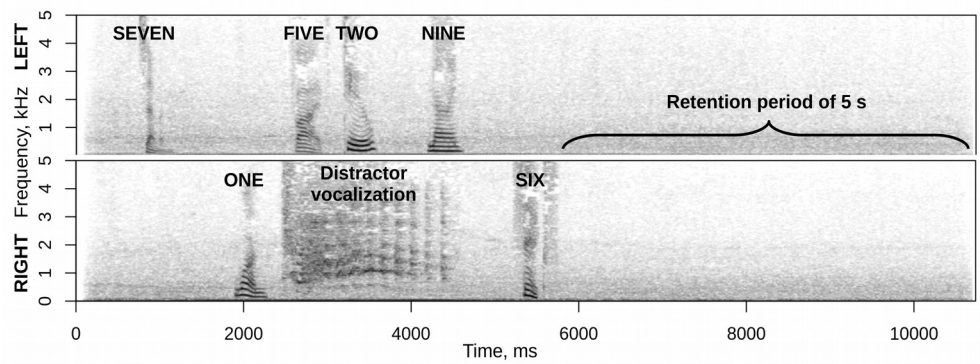

Fig. 2. An example sequence with six target digits (715296) and a distractor (laugh) from the short-term memory experiment designed to obtain an implicit measure of salience. Spectrograms of left and right stereo channels with 25 ms Gaussian windows and 50% overlap.

Participants

Separate samples of participants, fluent in English and with no self-reported hearing problems, were recruited for the three experiments (self-reported salience, objective salience, and emotion intensity) on the online platform https://www.prolific.co/. No demographic characteristics were collected; according to the statistics on the website, about 75% of participants on this platform are native English speakers, 57% are female, and 72% are between 20 and 40 years of age. To ensure the quality of collected data, clearly invalid submissions were not considered or reimbursed (e.g., very rapid identical responses to many consecutive sounds). Additional verification steps were

experiment-specific, such as catch trials with a pair of identical sounds (see Appendix for details). Since there were no expectations as to the effect size, power analysis was not performed, and data collection continued until the desired precision of a-posteriori estimates was achieved (Kelley, Maxwell, & Rausch, 2003), namely the width of 95% credible intervals of ±5% for each sound for both emotion intensity and self-reported salience. Unfortunately, it proved unfeasible to achieve a comparable precision level for objective salience, even after limiting this analysis to 64 out of 128 vocalizations. The final sample sizes after quality control were as follows: emotion intensity N = 45 participants (5727 trials, 43 to 45 trials per sound), self-reported salience N = 90 (10146 trials, 128 to 200 trials per sound), objective salience N = 238 (11695 trials, 147 to 177 trials per sound).

Data analysis

All analyses were performed on unaggregated, trial-level data using Bayesian multilevel models, which are more flexible than the corresponding frequentist models and provide a natural framework for combining the results of multiple experiments. In Bayesian modeling, the output of the analysis is a posterior distribution of credible parameter values given the data, model structure, and prior assumptions (McElreath, 2018). These posterior distributions were summarized by their medians and 95% credible intervals, reported in the text as 95% CIs. Linear models were fit using the R package *brms* (Bürkner, 2017) with default conservative priors. To estimate salience scores for each sound based on pairwise comparisons in Experiment 2 ("self-reported salience"), a custom model was written directly in *Stan* (https://mc-stan.org/). The outcome distributions were as follows: beta for intensity scores, ordered logistic for pairwise comparisons, and zero-inflated binomial for recall errors, with appropriate random effects. Please see the Appendix for more details on data analysis. All materials for running the experiments (HTML, audio stimuli, etc.), datasets of responses, and scripts for statistical analysis can be downloaded from http://cogsci.se/publications.html.

Results

Salience scores extracted from pairwise comparisons of 128 vocalizations were excellent predictors of emotion intensity scores produced by a different sample of participants (R^2 = .65 assuming a linear relationship), particularly within each emotion category (R^2 = .81 after adding an interaction with emotion). The likely

reason is that different emotions are expressed with distinct acoustic call types, whose auditory salience can be generally high (e.g., screams) or low (e.g., grunts or laughs). Thus, the most intense laughs had a salience score of only ~50%, far below the top salience of screams of fear or pain. At the same time, salience scores almost perfectly explained the perceived intensity of both amusement and fear considered separately (Fig. 3).

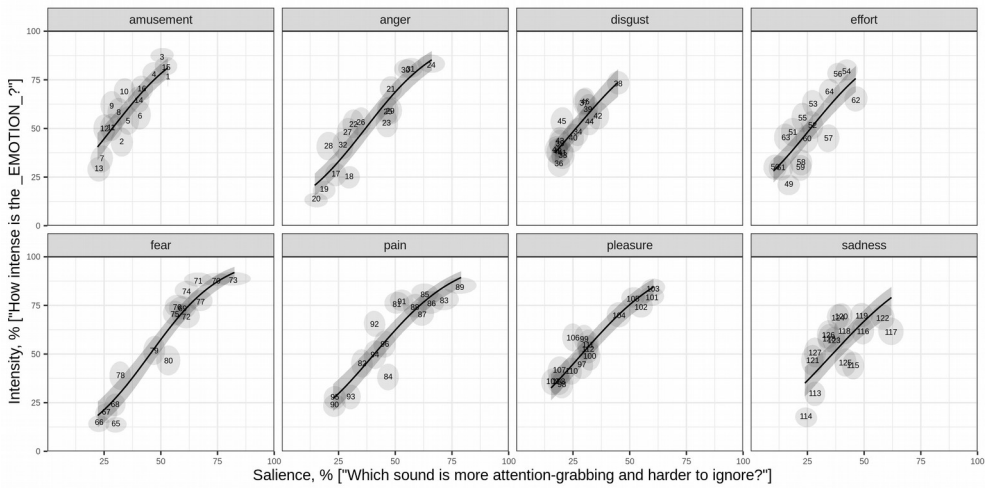

Fig. 3. Perceived intensity of emotion and self-reported salience of experimental stimuli. Light-gray ellipses mark the position of individual sounds (labeled 1 through 128) with two-dimensional 95% CIs. Solid lines show the relation between intensity and salience within each emotion class predicted with multilevel beta-regression, with shaded 95% CIs.

The average number of recall errors in the short-term memory task was about 22% higher with vs. without a distractor vocalization: odds ratio (OR) = 1.22, 95% CI [1.14, 1.31]. The effect of different distractors varied greatly: 23 out of 64 vocalizations seemed to have little or no effect on recall accuracy (95% CI of the OR overlapped with one), while the remaining 41 vocalizations increased the odds of recall errors by up to 70% (Fig. 4). None of the analyzed individual acoustic characteristics were significant predictors of recall errors, probably because of the relatively noisy nature of data from the memory task. However, both self-reported salience and intensity ratings were positive predictors of recall errors. The odds of a recall error were ~25% higher for sounds with the highest vs. lowest salience score: OR = 1.25, 95% CI [1.08, 1.45]. Crucially, the positive effect of self-reported salience on recall errors remained essentially unchanged after controlling for the duration of distractor vocalization (OR = 1.27 [1.06, 1.53]), while duration had no detectable independent effect on recall errors (OR = 0.98 [0.85, 1.13]). In other words, the elevated proportion of recall errors was not simply a matter of

longer vocalizations obscuring more targets, but a result of greater distraction caused by salient vocalizations.

Likewise, vocalizations that were rated higher on emotion intensity were responsible for ~16% more recall errors: OR = 1.16 [1.02, 1.32] for maximum vs. minimum intensity. However, multiple regression suggests that perceived emotion intensity had no independent effect on recall errors (OR = 0.99 [0.77, 1.25]) after controlling for self-reported salience (OR = 1.26 [0.96, 1.70]). In other words, participants were easily distracted by vocalizations suggestive of high emotion intensity, but this effect was mediated by the auditory salience of these sounds.

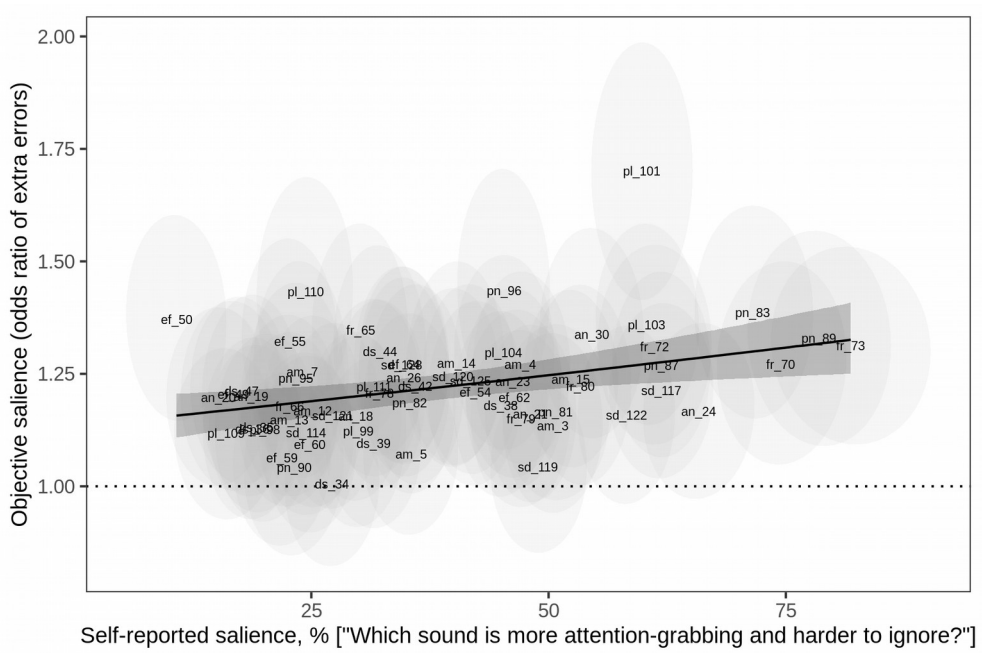

Fig. 4. Extra errors attributable to the presence of a distractor vocalization as a function of its self-reported salience. Light-gray ellipses mark the position of individual sounds (N = 64) with two-dimensional 95% CIs. The solid regression line shows the best fit with shaded 95% CI. The dotted line marks OR = 1 (as many errors as without any distractors). am = amusement, an = anger, ds = disgust, ef = effort, fr = fear, pn = pain, pl = pleasure, sd = sadness.

To see how well relative salience judgments could be predicted from low-level acoustic characteristics, the choice between two sounds in each trial was modeled as a function of their difference on 22 acoustic characteristics. A multilevel ordered logistic regression model correctly predicted ~66% of participants' responses in a three-choice task (chance level = 33%) and identified the acoustic

characteristics that made a vocalization highly salient: greater RMS amplitude and duration, higher pitch and spectral centroid (brightness), more variable bandwidth, more high-frequency modulation (roughness, amplitude and frequency modulation), higher novelty derived from the self-similarity matrix, and a few weaker predictors such as low spectral flatness and more variable harmonicity (Fig. 5, left panel). Likewise, intensity scores were predicted from the acoustic characteristics of each sound scaled relative to other sounds expressing the same emotion (Fig. 5, right panel). Since the stimuli were normalized for subjective loudness, its average level did not vary much, but the lack of effect of the variability in loudness is surprising, and so is the weak effect of pitch variability. Otherwise, these findings are in line with theoretical predictions: emotional vocalizations attract more attention if they produce a greater excitation in the auditory system due to the intensity and unpredictability of sensory stimulation.

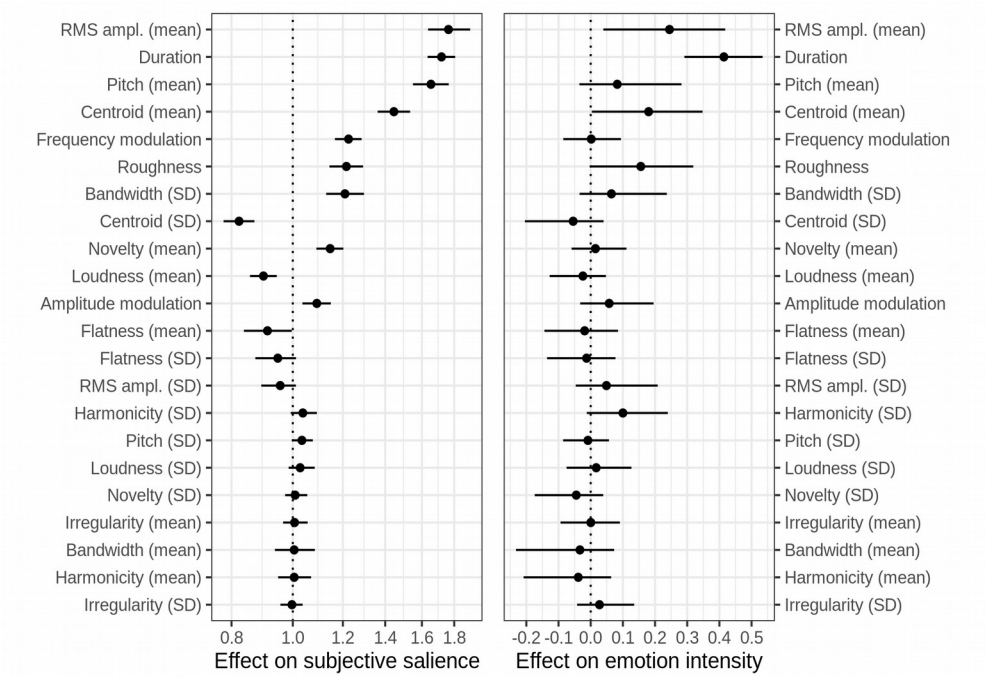

Fig. 5. Acoustic predictors of self-reported salience (left panel; odds scale, i.e. no effect = 1) and emotion intensity (right panel; linear scale, i.e. no effect = 0). Beta-coefficients from multiple regression showing the effect of a 1 SD difference in each acoustic characteristic, with 95% CI. Mirroring the design of the two experiments, acoustic predictors were normalized across all sounds for salience and within emotion category for intensity; the magnitude of effects is directly comparable within each panel.

Discussion

In line with the salience code hypothesis, high-intensity nonverbal vocalizations displayed acoustic characteristics previously reported to be associated with high bottom-up auditory salience. These acoustic properties made emotionally intense vocalizations more effective at involuntarily attracting the listeners' attention based on both self-reported distraction and an objectively measured drop in accuracy on a short-term memory task. In fact, the relationship between self-rated salience and emotion intensity ratings was so tight as to raise the suspicion that participants interpreted the two questions as synonymous, or that emotion intensity mediated both salience ratings and recall errors.

Several pieces of evidence argue against this interpretation. Multiple regression analysis suggests that salience ratings mediated the effect of emotion intensity on recall errors in a short-term memory task, not the other way round. Furthermore, the predictors of self-reported salience in emotional vocalizations were in line with psychoacoustic research on auditory salience, which is conducted using mixed vocal and environmental sounds, rather than emotional stimuli (Huang & Elhilali, 2017; Kayser et al., 2005; Zhao et al., 2018). In other words, high-intensity emotional vocalizations competed with goal-directed attention and were rated as highly distracting mostly because of their low-level acoustic characteristics, and not because listeners interpreted them as conveying a strong emotion.

It is also noteworthy that the correlation between salience and intensity ratings (which were provided by independent samples of participants) was considerably higher within each emotion category than across all 128 sounds. Each emotion was expressed with one or several, partly distinct types of sounds (laughs of amusement, moans and sighs of pleasure, etc.; see Table 1). These call types differ in their eliciting context and overall salience in addition to displaying meaningful within-call variation (Briefer, 2012; Fischer et al., 2017). For example, grunts and groans of disgust convey a wide range of intensity levels, although even the most intense ones are acoustically less salient than screams of relatively mild pain or fear (Fig. 3). At the same time, grunts with higher bottom-up salience express stronger disgust, and likewise, more salient screams are associated with more intense fear. The salience code reported here thus informs relatively subtle within-call acoustic variation, and not only qualitative differences between generally high-arousal (screams, roars) and low-arousal (sighs, moans) call types.

Vocalizations with high bottom-up salience possessed three acoustic characteristics. First, they were longer and louder, presumably causing more excitation in afferent sensory pathways. Second, salient sounds were highly variable and unpredictable, as reflected in such acoustic measures as amplitude

and frequency modulation, roughness, variability in spectral bandwidth and harmonicity, and novelty extracted from self-similarity matrices. From a predictive coding perspective, unpredictability in sensory input causes a mismatch with top-down expectations, attracting attention and making it difficult to ignore the elusive stimulus (Kaya & Elhilali, 2014; Southwell et al., 2017). Third, salient vocalizations tended to have relatively high pitch and a bright timbre, as also reported by Huang and Elhilali (2017). The mechanisms responsible for this effect are less clear. High-frequency sounds appear to be louder than low-frequency sounds of the same amplitude, but subjective loudness was controlled in this study, and the effect of pitch on salience was too large to be attributable to any remaining variations in loudness. Possibly, high-pitched vocalizations with strong harmonics were interpreted as evidence of high vocal effort (Gobl & Ní Chasaide, 2010), capturing attention in a top-down manner. Likewise, certain sounds, such as sexual moans #101 and 110 (Fig. 4), caused unexpectedly large drops in task performance, again suggesting a possible role of top-down attention.

This ambiguity highlights the difficulty of isolating purely bottom-up attentional mechanisms or measuring salience as a property fully encapsulated from top-down control (Huang & Elhilali, 2017). In real-life encounters, listeners take into account the speaker's identity, sex, age, facial expression, and other available contextual information, which may considerably affect the allocation of attention predicted by a purely bottom-up model. But even in a controlled experiment, it is challenging to separate bottom-up from top-down influences.

Another limitation of the present study is that it relied on the perceived emotion intensity rather than the actual (unknown) affective state of the speaker. As a result, while the current results demonstrate a close correlation between how emotionally intense and how distracting a vocalization is perceived to be, it remains an assumption that this perceived intensity is an accurate reflection of the speaker's arousal level. Listeners are good at identifying the emotion portrayed with nonverbal vocalizations (Lima et al., 2013; Maurage et al., 2007) as well as the production context of unstaged, naturalistic vocalizations of the kind used in this study (Anikin & Persson, 2017). Even so, in future studies it would be important to map neurological and somatic markers of the intensity of particular emotions or overall arousal, such as changes in skin conductance, onto the accompanying changes in vocal production instead of relying on subjective ratings of listeners. Another productive avenue for follow-up research would be to systematically manipulate salience-relevant acoustic characteristics of emotional speech or vocalizations. As our understanding of bottom-up auditory salience keeps improving, very specific hypotheses can be formulated and tested to either support or falsify the claim that more salient vocalizations convey – or are perceived to convey – more intense emotions. If the salience code is confirmed and shown to be generalizable beyond human nonverbal vocalizations, it can

provide a powerful framework for guiding research on acoustic communication and integrating it with the neuroscience of auditory perception.

Returning to the ultimate question of evolutionary causation, I would argue that the salience code of emotion expression described here is best seen as an example of sensory exploitation. The initial selective pressures would presumably have affected the vocalizations emitted in high-stake, survival-relevant contexts: infant separation cries (Lingle et al., 2012), anti-predator alarm calls of varying urgency (Manser, 2001), aggressive and mating calls (Reby & Charlton, 2012), etc. Natural selection would have favored changes in the physiological links between arousal and voice production that made high-intensity vocalizations of this kind easier for conspecifics to notice and harder to ignore. Once established, these physiological links would then automatically generalize to all vocal behaviors, explaining why even close-range vocalizations, such as grunts of disgust and moans of pleasure, obey the same salience code.

Other explanations are also possible. If vocal production and auditory perception have been co-evolving closely (Ron, 2008), the match between emotion intensity and bottom-up salience could be the result of the auditory system being exquisitely tuned to the acoustic properties of high-intensity calls, either innately or as a result of developmental plasticity and increasing sensitivity to the typical or most relevant sensory input (Bao, 2015; Woolley et al., 2005). The properties of high-intensity vocalizations, in turn, may be shaped by factors unrelated to salience, such as the need for the caller to demonstrate his fitness and stamina by producing long, loud, and high-pitched vocalizations (Fischer, Kitchen, Seyfarth, & Cheney, 2004). For instance, Ma and Thompson (2015) hypothesized that acoustic correlates of emotion (such as intensity, production rate, and spectral characteristics) have evolved to be affected by emotion because they reveal important biological information about the speaker, such as their size, proximity, and speed.

Whatever the exact evolutionary story, efficient communication requires some coordination between the form of signals and sensory biases. The close correspondence between the effects of strong emotion on the voice and the sensitivity of the auditory system demonstrated in this study is an example of this communicative principle, and possibly also a general explanation for some non-arbitrary properties of human and animal high-arousal vocalizations.

Acknowledgments

I would like to thank Tomas Persson and Peter Gärdenfors for their comments on the manuscript.

References

Anikin, A. (2019). Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behavoir Research Methods*, *51*(2), 778-792.

Anikin, A. & Persson, T. (2017). Non-linguistic vocalizations from online amateur videos for emotion research: a validated corpus. *Behavior Research Methods, 49*(2), 758-771.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051-2056.

August, P. V., & Anderson, J. G. (1987). Mammal sounds and motivation-structural rules: a test of the hypothesis. *Journal of Mammalogy*, *68*(1), 1-9.

Ball, M., & Bruck, D. (2004). The salience of fire alarm signals for sleeping individuals: a novel approach to signal design. In: Shields J. (Ed.). *Proceedings of the Third Human Behaviour in Fire Conference* (pp. 303-314), Belfast, 1-3 October.

Bao, S. (2015). Perceptual learning in the developing auditory cortex. *European Journal of Neuroscience*, *41*(5), 718-724.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1-20.

Bürkner, P. C. (2017). brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Davila-Ross, M. D., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology, 19*(13), 1106–1111.

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., ... & Newen, A. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1859), 20170990.

Fischer, J., Kitchen, D. M., Seyfarth, R. M., & Cheney, D. L. (2004). Baboon loud calls advertise male quality: acoustic features and their relation to rank, age, and exhaustion. *Behavioral Ecology and Sociobiology*, *56*(2), 140-148.

Fischer, J., Wadewitz, P., & Hammerschmidt, K. (2017). Structural variability and communicative complexity in acoustic communication. *Animal Behaviour*, *134*, 229-237.

Fitch, W. T. (2018). The biology and evolution of speech: a comparative analysis. *Annual Review of Linguistics*, *4*, 255-279.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, *63*(3), 407-418.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000*. IEEE (Vol. 1, pp. 452-455).

Fuller, R. C., Houle, D., & Travis, J. (2005). Sensory bias as an explanation for the evolution of mate preferences. *The American Naturalist*, *166*(4), 437-446.

Gobl, C., & Ní Chasaide, A. (2010). "Voice source variation and its communicative functions". In Hardcastle, W. J., Laver, J., & Gibbon, F. E. (Eds.). *The handbook of phonetic sciences* (2nd ed.) (pp. 378-423). Singapore: Wiley-Blackwell.

Hamilton-Fletcher, G., Pisanski, K., Reby, D., Stefańczyk, M., Ward, J., & Sorokowska, A. (2018). The role of visual experience in the emergence of cross-modal correspondences. *Cognition*, *175*, 114-121.

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, *141*(3), 2163-2176.

Jégh-Czinege, N., Faragó, T., & Pongrácz, P. (2019). A bark of its own kind – the acoustics of 'annoying' dog barks suggests a specific attention-evoking effect for humans. *Bioacoustics*, 1-16. doi:10.1080/09524622.2019.1576147

Karp, D., Manser, M. B., Wiley, E. M., & Townsend, S. W. (2014). Nonlinearities in meerkat alarm calls prevent receivers from habituating. *Ethology*, *120*(2), 189-196.

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*, 327.

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, *15*(21), 1943-1947.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions*, *26*(3), 258-287.

Kim, K., Lin, K. H., Walther, D. B., Hasegawa-Johnson, M. A., & Huang, T. S. (2014). Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, *38*, 78-85.

Köppl, C. (2009). Evolution of sound localisation in land vertebrates. *Current Biology*, *19*(15), R635-R639.

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., ... & Baron-Cohen, S. (2019). The EU-Emotion Voice Database. *Behavior Research Methods*, *51*(2), 493-506.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron, 73*(4), 653-676.

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2018). tuneR: Analysis of Music and Speech. https://CRAN.R-project.org/package=tuneR

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.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology, 58*(5), 698–726.

Ma, W., & Thompson, W. F. (2015). Human emotions track changes in the acoustic environment. *Proceedings of the National Academy of Sciences, 112*(47), 14563-14568.

Manser, M. B. (2001). The acoustic structure of suricates' alarm calls varies with predator type and the level of response urgency. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1483), 2315-2324.

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

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: Chapman and Hall/CRC.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855-869.

Neuhoff, J. G. (2018). Adaptive Biases in Visual and Auditory Looming Perception. In T. L. Hubbard (Ed.). *Spatial biases in perception and cognition*. Cambridge, UK: Cambridge University Press.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1-16.

Reby, D., & Charlton, B. D. (2012). Attention grabbing in red deer sexual calls. *Animal Cognition, 15*(2), 265-270.

Ron, S. R. (2008). The evolution of female mate choice for complex calls in túngara frogs. *Animal Behaviour*, *76*(6), 1783-1794.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178.

Ryan, M. J., & Cummings, M. E. (2013). Perceptual biases and mate choice. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 437-459.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of Eurospeech 2001*, Aalborg, Denmark, vol. 1, pp. 561–564.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394-3411.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978-982.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160105.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971-995.

Stebbins, W. C. (1980). The evolution of hearing in the mammals. In *Comparative studies of hearing in vertebrates* (pp. 421-436). New York: Springer.

Tajadura-Jiménez, A., Väljamäe, A., Asutay, E., & Västfjäll, D. (2010). Embodied auditory perception: The emotional impact of approaching and receding sound sources. *Emotion*, *10*(2), 216-229.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410-433.

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355-366.

Vachon, F., Labonté, K., & Marsh, J. E. (2017). Attentional capture by deviant sounds: a noncontingent form of auditory distraction? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 622-634.

Woolley, S. M., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371-1379.

Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Furukawa, S., Dick, F., ... & Chait, M. (2018). Rapid ocular responses are a robust marker for bottom-up driven auditory salience. *BioRxiv*, 498485.

Appendix: Supplemental Methods

Experiment 1: emotion intensity

Participants

Four out of 50 participants were excluded for obvious cheating (very rapid identical responses to many consecutive sounds), and another was excluded based on very low ($r < .05$) correlation of the provided scores with the median scores of all other participants, suggesting random responses. The remaining sample size was thus 45 – that is, each sound was rated 45 times. This sample size ensured that the average width of 95% CIs on emotion intensity estimates per sound was approximately ±5%.

Data analysis

Intensity ratings were modeled as beta-distributed, with random intercepts per sound and per participant. Modeling the differences between sounds as a random effect made the model more sparse compared to estimating 128 independent coefficients, thereby decreasing the risk of overfitting. To analyze the relationship between salience and emotion intensity, fitted self-reported salience values from Experiment 2 were added as a predictor, with a random slope per emotion. For plotting purposes, uncertainty from both experiments was incorporated by making repeated random draws from both posteriors, resulting in two-dimensional credible intervals (Fig. 3 in the main text). Acoustic predictors of emotion intensity were determined by normalizing (*z*-transforming) acoustic characteristics of sounds within each emotion category and using them as predictors of trial-level intensity scores. Modeling was performed with default conservative priors in the *brms* package, but models with acoustic predictors imposed additional shrinkage on regression coefficients with a horseshoe prior to control for multiple comparisons.

Experiment 2: self-reported salience

Participants

Participants were recruited until the precision of salience estimates was similar to that in Experiment 1 (95% CI ±5% for each sound); each participant rated 100 or 128 sound pairs. Two verification steps were employed to ensure data quality. First, there were three catch trials with a pair of identical sounds, which were expected to be ranked as similar. Second, the average salience was estimated based on all responses, and the proportion of responses violating this average ranking was calculated for each participant. On average, participants made choices that contradicted the population consensus in 12% of trials; if this proportion exceeded 20% (cutoff chosen based on an analysis of random response patterns), the participant was excluded from the analysis. Based on these two criteria, 12 out of 102 submissions were excluded, resulting in a final sample size of 90 participants. In a sensitivity analysis, salience estimates based on the full (N = 102) and high-quality (n = 90) samples were almost perfectly correlated (r = .998), indicating that the exclusion of 12 participants had only a negligible effect on results.

Data analysis

It was assumed that the choice between two sounds reflected the distance between them on a unidimensional salience scale: participants would respond sound 1 if the difference in salience between sounds 1 and 2 was large and positive, sound 2 if it was large and negative, and similar (tie) if it was close to zero. The threshold for calling a tie was assumed to vary across individuals, and the latent salience variable was mapped onto responses via ordered logistic regression. The corresponding Bayesian model was written in Stan (https://mc-stan.org/) and accessed from R (https://cran.r-project.org/). The latent salience variable is underspecified; to ensure convergence, the position of the first sound was therefore fixed at zero, and the overall scale was set by the normal prior on salience scores with an arbitrarily chosen standard deviation. To facilitate comparisons with intensity scores in Experiment 1, the posterior distribution of salience estimates was normalized to range from 0 to 100%.

To check the consistency of responses, the dataset was repeatedly split into two halves (45 participants each), salience ratings were calculated using a simple heuristic (+1 to the chosen sound and -1 to the other sound in each trial, ignoring ties) in each half-dataset, and then the resulting salience scores were correlated.

Averaging over 1000 iterations, this correlation was r = .95, 95% CI [.94, .96], suggesting that different groups of participants largely agreed about which sounds were more and which less salient. Acoustic predictors of salience were determined by modeling trial-level choices in pairwise comparisons with ordered logistic regression based on the difference between the two sounds on acoustic characteristics normalized (z-transformed) across all 128 stimuli.

Experiment 3: objective salience

Participants

The required sample size was estimated in a pilot study with 24 vocalizations, aiming to achieve reasonable precision on the estimates of effect size for each vocalization (extra errors attributable to its presence compared to no distractor). Another 40 vocalizations were then tested with a new sample of participants. Accuracy was high (36% to 100% per participant, 95% of submissions over 50% correct), so all submissions with minimum 40 out of 50 trials were included.

Data analysis

The number of errors per sequence was modeled as a draw from a zero-inflated binomial distribution with six trials (one for each digit), with a random intercept per participant to account for individual variability in general performance and another random intercept per distractor, with 65 levels (64 for different vocalizations and one for "none"). First, a model with no fixed effects was constructed to evaluate the expected number of errors for each vocalization (including "none"). The objective salience of each vocalization was operationalized as the ratio of the odds of making an error with a particular distractor vocalization vs. without any distractors. This odds ratio (OR) was calculated for each step in the MCMC chain and summarized as the median and 95% coverage interval.

Second, a similar model was constructed only for those trials that included a distractor vocalization to test whether acoustic characteristics of distractor vocalizations and their intensity and salience ratings from Experiments 1 and 2 predicted the number of errors. The ratings from Experiments 1 and 3 were treated as point estimates with measurement errors, calculated as the median and standard deviation, respectively, of their posterior distributions for each vocalization.