Human nonverbal vocalizations

ANDREY ANIKIN
COGNITIVE SCIENCE | LUND UNIVERSITY

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 978-91-88899-86-6

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 978-91-88899-86-6
ISSN 1101-8453 Lund University Cognitive Sciences 178

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

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

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

1. Nonverbal vocalizations as a form of communication.................................13

1.1 Signal production...14

1.1.1 Somatic signals...15

1.1.2 Innate form, innate context..16

1.1.3 Innate form, flexible context..18

1.1.4 Learned form...19

1.1.5 Summary of signal production...21

1.2 Signal perception...21

1.2.1 Direct effects...22

1.2.2 Innate responses..22

1.2.3 Learned responses..24

1.2.4 Summary of signal perception..26

1.3 Research questions...27

2. Species-typical component...29

2.1 Sources of spontaneous vocalizations (Paper I)....................................30

2.2 Are spontaneous vocalizations different? (Paper II)..............................34

2.3 Human nonverbal repertoire (Paper III)...35

3. Cracking the code...37

3.1 Testing acoustic manipulations (Papers IV-VI)....................................37

3.2 The logic of the acoustic code (Paper VII)...41

4. Summary...43

4.1 Conclusions..43

4.2 Broader significance...44

5. References...45

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.

Links to original papers

Paper I: https://doi.org/10.3758/s13428-016-0736-y

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: https://doi.org/10.1080/17470218.2016.1270976 [Full Access]

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: https://doi.org/10.1007/s10919-017-0267-y [Open Access]

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: https://doi.org/10.3758/s13428-018-1095-7 [Open Access]

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

Paper V: https://doi.org/10.1080/09524622.2019.1581839 [Open Access]

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: https://www.karger.com/PHO [Link to Journal]

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.

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 978-91-88899-86-6
ISSN 1101-8453

9 789188 899811

LUND
UNIVERSITY