4

Emotion communication in animal vocalizations, music and language: An evolutionary perspective

Piera Filippi[1 2 3] and Bruno Gingras[4]
[1] Institut of Language, Communication and the Brain, Aix-en-Provence
[2] Laboratoire Parole et Langage LPL UMR 7309, CNRS
[3] Laboratoire de Psychologie Cognitive LPC UMR 7290, CNRS, Aix-Marseille University
[4] Department of Psychology, University of Innsbruck

ABSTRACT Emotions are a biological universal which allow each individual to react appropriately to the surrounding physical and social environment. Emotions can be expressed and perceived through several sensory modalities. Here, we focus on the communication of emotions in the auditory domain. Specifically, we begin by defining emotions from a broad comparative perspective before turning our attention to acoustic universals in emotion processing in animal vocalizations, human communication (including language), and finally, music. Building on recent findings on cognitive processes and mechanisms underpinning emotional processing shared by these three domains, and on the adaptive role of the ability for emotion communication, we hypothesize that this ability is an evolutionary precursor of the ability for language. Finally, our work offers insights for empirically testable questions within this research framework and proposes that future

investigations should incorporate a comparative lens and consider human non-semantic vocalizations, such as laughter, alongside animal vocalizations.

KEYWORDS arousal, comparative analysis, emotion, evolution, language, music, vocal communication

1 EMOTIONS: BIOLOGICAL UNIVERSALS

Consider a mother happily talking to her baby during playtime, a dog barking to get her owner to open the door and an instrumental melody played with increasing intensity. Although these three examples are quite different from each other in terms of communicative systems, they all share a key aspect: each of these signals expresses an emotional state and may trigger a physiological or behavioral reaction in the listener. In this chapter, we provide a comparative analysis of recent findings on the role of emotions in animal vocalizations, music and language. We do so with the overarching aim to shed light on the adaptive role of emotional communication in the auditory domain for the evolution of linguistic communication in humans.

As a first step into this comparative analysis, we propose an operative definition of emotional states (hereafter 'emotions'), which we will use throughout this chapter: *Emotions* are physiological states of the brain that increase an animal's ability to react appropriately in the surrounding environment or in the given social context, and are accompanied by preparatory tuning of the somatovisceral and motor systems (Scherer, 2005, 2009; see also Nesse, 1990).

Emotions have often been classified in multiple ways. For instance, a long-standing research tradition is centered around the study of emotions as discrete categories such as happiness or anger (Anderson & Adolphs, 2014; Cowen & Keltner, 2017; Ekman, 1992). These emotion categories are often referred to as *basic emotions* because they are thought to be culturally universal (Sauter, Eisner, Ekman, & Scott, 2010; Sauter, Eisner, Ekman, & Scott, 2015). In addition, *dimensional emotion models* (Anderson & Adolphs, 2014) argue that emotional states can be classified based on their *valence* (positive or negative) and their *arousal* level (i.e.,

activation or responsiveness levels, classified as high or low). Emotional arousal is a state of the brain or the body reflecting responsiveness to sensory stimulation, ranging from very low responsiveness to frenetic excitement (Russell, 1980). Accordingly, increases in arousal levels are correlated with increases in behavioral, hormonal and/or neurological activity. In addition, changes in emotional states may affect cognitive performances (Mendl, Burman, & Paul, 2010).

The identification of basic emotions is key for the investigation of physiological correlates and variance of human emotions. However, it may not be the best model to capture variance in non-human animal emotions. Indeed, while measuring degrees of valence and arousal is empirically possible (Briefer, 2012; Briefer et al., 2015; Briefer, Tettamanti, & Mcelligott, 2015), identifying the physiological correlates of emotional categories such as happiness or disgust in nonhuman animals is more challenging and may lead to a more anthropocentric interpretation of the signal (Briefer, 2012).

2 ANIMAL VOCALIZATIONS: ACOUSTIC UNIVERSALS IN EMOTION COMMUNICATION

Emotional communication is pervasive in animal vocalizations. A large number of studies have investigated the expression of affective states through voice modulation (see, e.g., McComb, Taylor, Wilson, & Charlton, 2009; Pongráz, Molnár, & Miklósi, 2006; Zimmermann, Leliveld, & Schehka, 2013). In particular, much research has focused on the expression of emotional arousal, which can be directly linked to the physiological state of the signaler. For instance, states of heightened emotional arousal correlate with quantifiable changes in physiological parameters and induce increased activation of muscles. Specifically, increased activation of muscles involved in vocal production, such as of the diaphragm, intercostals and vocalis muscles, alters subglottal pressure, that is the way air flows through the vocal tract. This may induce vocal folds to vibrate at their natural limit, generating sound waves at heightened amplitude. These sound waves may be perceived as harsh sounds (Taylor & Reby, 2010). In addition, states of heightened emotional arousal may induce heightened

muscular tension, which prepares the signaler for immediate action (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015).

In turn, the ability to recognize heightened levels of arousal in vocalizations may help avoiding threats or disturbances in the surrounding environment, as for instance the imminent approach of a predator. A vocalization produced by a signaler that is in a heightened arousal state, as for instance, an alarm call or a disturbance call, may induce fear or alertness in the listeners, thus prompting them to avoid dangers or disturbances in the surroundings. Hence, a high arousal vocalization may have been shaped by selection to affect others' behavior in an urgent manner (Fitch, Neubauer, & Herzel, 2002).

In a recent study, Filippi et al. (2017a) suggest that humans are able to recognize heightened states of arousal in vocalizations from nine animal species spanning across all classes of terrestrial vertebrates and varying in size, social structure and ecology. These vocalizations were emitted in negatively valenced contexts, namely threat, competition or general disturbance. Arousal level (high versus low) was the only emotional dimension that varied across calls. This research provides evidence showing that humans use acoustic parameters related to F0 (*fundamental frequency*) to recognize heightened states of arousal in calls across all these species. These results hold true across three different language groups (English, German and Mandarin native speakers), suggesting that this ability is biologically rooted in humans. Hence, this study points to a phylogenetic continuity of emotional communication across species, in terms of acoustic parameters involved in vocal production (Bowling, Gingras, Han, Sundararajan, & Opitz, 2013; Briefer, 2012; Linhart, Ratcliffe, Reby, & Špinka, 2015; Morton, 1977; Reichert, 2013; Stoeger, Baotic, Li, & Charlton, 2012; Stoeger, Charlton, Kratochvil, & Fitch, 2011; Templeton, Greene, & Davis, 2005) and in the perception of emotional content in these vocalizations (Belin et al., 2008; Faragó et al., 2014; McComb et al., 2009; Pongrácz et al., 2006; Sauter et al., 2010). In the context of emotional vocalizations, the ability to identify higher arousal states of callers expressed through voice modulation, which reflect higher levels of urgency or danger, is a necessary condition for adaptive reactions such as escape. Therefore, this perceptual ability is as important as the ability to produce calls that differ in the level of expressed arousal.

Furthermore, the findings of Filippi et al. (2017a) suggest that the ability to identify emotional arousal in animal calls, which constitutes a biological phenomenon key to survival, may have emerged in the early stages of the evolution of vocalizing animals and has been preserved across a broad range of animal species. Moreover, these results are in line with Darwin's hypothesis of a shared set of mechanisms grounding vocal emotional expression across terrestrial vertebrates (Darwin, 1871, 1872).

Within this research framework, Filippi et al. (2017b) further analyzed humans' sensitivity to high-arousal calls with negative valence and to high-arousal calls with positive valence in silver foxes. Interestingly, they found that humans are not able to identify high levels of arousal in vocalizations of positive valence. This may be due to a *negativity bias*, namely the fact that people are particularly sensitive to emotionally negative events (Kahneman & Tversky, 1984; Yuan et al., 2007). As argued by Filippi and colleagues, high-arousal calls with positive valence may not be as salient as high-arousal calls with negative valence to the human ear. In other words, these findings suggest that humans' ability to recognize arousal in calls with positive valence might have not been selected by evolution because they are not as crucial for survival as arousal calls produced in negatively valenced contexts. Certainly, in order to corroborate this finding, it is necessary to test humans' perception of vocalizations from more species. These vocalizations need to vary systematically in valence and arousal.

Taken together, these studies suggest two points: a) the modulation of acoustic features in the voice conveys information on the emotional arousal state of the vocalizing animal; b) the ability to identify emotional arousal in animal calls – which is linked to the ability to express emotional content by modulating specific acoustic parameters – may have been preserved across animal species, based on its key role in preparing for action in contexts that were crucial for survival. Building on these findings, we aim to provide a review on studies addressing whether – and if so, how – humans' ability to process animal vocalizations expressing heightened states of arousal is evolutionary linked to their ability for emotional processing in music and language.

3 EMOTION PROCESSING IN HUMAN COMMUNICATION: INSIGHTS FOR THE STUDY OF LANGUAGE EVOLUTION

Humans typically combine two sources of information to comprehend each other in communicative contexts: linguistic (e.g., segmental information, morphology or syntax) and nonlinguistic (e.g., body posture, facial expression, prosodic modulation of the voice and pragmatic context) information. For the purposes of this chapter, we will focus on the interaction of two auditory channels: *segmental information* and *prosodic modulation*, the so-called *musical* aspect of speech, which includes timing, frequency spectrum and amplitude (Lehiste, 1970).

Traditionally, studies have addressed language focusing on the segmental level of information, overlooking the analysis of prosodic modulation (Hockett, 1960; Martinet, 1980). These studies often fail to consider the role of prosody, although speakers across cultures and languages modulate segmental information *within* prosodic contours in the spoken signal. It is possible to distinguish two types of prosodic modulation of the voice: *affective prosody*, which affects perception of linguistic stimuli in a cross-culturally similar manner by expressing emotional content (Sauter, Eisner, Ekman, & Scott, 2010), and *linguistic prosody*, which affects the spoken signal in a way that is specific to natural languages.

Linguistic prosody strongly affects the segmental dimension of the signal, orienting the identification of segmental information (Bosker, 2017), lexical items (Van Donselaar, Koster, & Cutler, 2005) and syntactical connections (Soderstrom, Seidl, Nelson, & Jusczyk, 2003). Specifically, prosodic parameters affect perception of phrase boundaries, of a word (*lexical stress*), or of specific words within a sentence (*sentence focus*). Consider for instance, "MARY gave the book to John" vs. "Mary gave the book to JOHN". Here, the two sentences are identical from a segmental point of view. However, by accenting one word or the other through voice modulation, the speaker orients the listener's comprehension of the utterance. In addition, linguistic prosody may be used to distinguish word meanings (in tone languages) or statement types, for instance an assertion from a question or a command (Cutler, Oahan, & Van Donselaar, 1997).

Hence, in human communication, multiple channels may simultaneously determine utterance comprehension. But what is the relative sa-

lience of segmental and intonational channels in the process of meaning identification? Multiple brain and behavioral studies have addressed this question, mainly focusing on the communication of emotional valence (Pell et al., 2011; Schirmer & Kotz, 2003). Regarding emotional arousal, in line with Storbeck and Clore (2008), we suggest that the biologically anchored ability to express and identify heightened levels of arousal affects language processing and related cognitive processes in modern humans. Indeed, learning processes and long-term memory are enhanced when involving high arousal states, which are linked to urgency.

The analysis of emotional communication is crucial for the study of language evolution. Several studies suggest that the informational value of animal calls is given by their emotional content (Seyfarth & Cheney, 2003). Since the ability to both express and identify emotions may favor survival across animal species, it is plausible to assume that the ability for language in humans built on this pre-existing ability to express emotional content. In modern humans, emotional communication can take place by integrating multiple channels, which can interact with each other, for instance, through priming or simultaneous interaction. Research has shown that the verbal content and/or the prosodic modulation of spoken units prime the interpretation of a following target word in an emotion-congruent manner (Nygaard & Lunders, 2002). Evidence further shows that, when the two channels are congruent, emotional prosody biases memory of affective words (Schirmer, 2010; Schirmer, Kotz, & Friederici, 2002). Notably, in emotion communication, segmental information and prosodic modulation of the voice can simultaneously express different contents. For instance, this is the case when someone says "I'm sad!" with happy prosody. In this case, segmental information and voice modulation conflict in the meaning they convey. Recent research suggests that in cases like this, where the two communicative channels are incongruent, prosody dominates over segmental information in recruiting cognitive resources for emotion identification (Filippi et al., 2017c).

The analysis of the role of these two channels – segmental information and prosody – in emotion communication is relevant to the debate on the evolution of vocal communication. Indeed, as outlined in the previous section, humans use specific features in prosodic modulation of the voice to identify emotional content in non-verbal vocalizations, an ability that is universal and biologically rooted (Filippi et al., 2017a). The

combination of these data with evidence on the coexistence of prosodic modulation and segmental information in modern human's language suggests that the ability for emotional communication through prosodic modulation of the voice is evolutionary older than the ability to process segmental information (Brown, 2017; Darwin, 1871; Fitch, 2010; Filippi, 2016; Mithen, 2005; Panksepp & Trevarthen, 2009) and may have paved the emergence of the ability to articulate segmental information within prosodic contours. In line with this hypothesis, multiple studies suggest that prosody drives words' segmentation (Johnson & Jusczyk, 2001), the ability to map sounds to meanings (Filippi, Gingras, & Fitch, 2014) and syntactic disambiguation (Soderstrom et al., 2003). Accordingly, research shows that prosodic cues favor lexical access and syntactic analysis at an ontogenetic level, orienting language acquisition in preverbal children (de Carvalho, Dautriche, Lin, & Christophe, 2017; Gout, Christophe & Morgan, 2004).

4 INTERPERSONAL COORDINATION IN EMOTIONAL AUDITORY SIGNALS: MUSICAL ORIGINS OF LANGUAGE?

Human communication typically takes place in interactional contexts such as conversations. In these contexts, interlocutors take turn in exchanging utterances, following precise time patterns (Sacks, Schegloff, & Jefferson, 1978). Indeed, recent research provides cross-cultural evidence suggesting that, in conversations, listeners show the following abilities: a) predict the end of the speaker's utterance with the help of prosodic cues such as lengthening of final phrase, b) plan the production of an answer, and c) reply after an average time window of 250 ms (Levinson & Holler, 2014; Magyari, De Ruiter, & Levinson, 2017; Stivers et al., 2009).

A wide range of comparative studies investigating animal communication systems report on the ability for vocal coordination in a number of species spanning all classes of animals. At least three types of behaviors depend on this ability: *choruses*, *duets* and *antiphonal calling* (Yoshida & Okanoya, 2005). These temporally organized behaviors occur in contexts such as territorial defense, social bonding and sexual advertisement,

which evidently involve various degrees of emotional activation. Choruses, commonly found in insects, anurans and in some species of birds, are produced by males only, whereas duets, which are observed in insects, anurans, birds and mammals, are typically performed by male-female pairs. Antiphonal calls, encountered in several species of birds and mammals, occur between any species members, independently from their sex. In choruses, males simultaneously produce a signal for sexual advertisement or as an anti-predator defensive behavior. Duets occur when members of a pair (e.g., sexual mates, caregiver-juvenile) exchange calls within a precise time window. In duets, whose function is to strengthen and display pair bonding, two sexual mates vocally interact with one another, each responding to the preceding vocalization. An animal may respond to signals from one individual, while ignoring those from another. Antiphonal calling occurs when more than two members of a group exchange calls within an interactive context, favoring group cohesion and diverting outsiders.

In humans, prosodic modulation of the voice is key to coordinating these interactive behaviors (see Filippi, 2016, for a review). Darwin described these vocal interactions as *musical* (Gamba et al., 2016; Geissmann, 2000):

> Primeval man, or rather some early progenitor of man, probably first used his voice in *producing true musical cadences, that is in singing*, as do some of the gibbon-apes at the present day; and we may conclude, from a wide-spread analogy, that this power would have been especially exerted during the courtship between sexes, – would have expressed various emotions, such as love, jealousy, triumph, – and would have served as a challenge to rivals (Darwin, 1871, pp. 56–57; our emphasis).

Following Darwin, the distinctive musical cadence of these interactions has induced multiple researchers to describe these behaviors using musical terms such as ‘rhythmic’ and ‘singing’ among others. For instance, Bryant (2013; see also Hagen & Bryant, 2003) suggests that human abilities for music evolved from the ability to use time-coordinated behaviors as a *coalition system*, as we find in other animal species.

Building on evidence on this kind of musical abilities in nonhuman animals, multiple studies have identified the evolutionary origins of

language in the ability for music production (Brown, 2001; Fitch, 2010, 2012; Mithen, 2005; Patel, 2006). This hypothesis is supported by recent evidence on neural and cognitive underpinnings shared by music and language. For instance, common cognitive mechanisms involved in the production and perception of structural relations have been identified in instrumental music and propositional morpho-syntax (Fedorenko, Patel, Casasanto, & Winawer, 2009; Patel, 2010). Moreover, brain imaging research suggests that amusic participants (see also Marin in this volume) show deficits in fine-grained perception of pitch (Foxton, Dean, Gee, Peretz, & Griffiths, 2004) and are not able to distinguish a question from a statement relying on pitch changes (Liu, Patel, Fourcin, & Stewart, 2010). These data supports the hypothesis that music and speech intonation share specific neural resources for processing fine-grained pitch changes. Further brain imaging studies report a considerable overlap in the brain areas involved in the perception of pitch and rhythm in songs and lexical units within sentences (Merrill et al., 2012; Zatorre, Belin, & Penhune, 2002) as well as in melodies and linguistic phrases (Brown, Martinez, & Parsons, 2006). In line with these findings, several studies on adults and children suggest that musical training facilitates syllabic and pitch processing in language (Schön, Boyer, Moreno, Besson, & Peretz, 2008; Schön, Magne, & Besson, 2004).

In addition to studies reporting evidence on the activation of overlapping brain areas for music and language perception, a key field for the investigation of musical abilities as the evolutionary precursor of language is the perception of emotions in both domains.

5 EMOTION COMMUNICATION: COMMONALITIES BETWEEN MUSIC AND LANGUAGE

As early as 1781, Jean-Jacques Rousseau noted that melody, by imitating the inflections of the voice, could express "all the vocal signs of the passions" (1781/1998, p. 322). A few decades later, Spencer linked music with the vocal communication of emotions (Spencer, 1857). More recently, Scherer's (1986) theory predicted a correspondence between emotion-specific physiological changes and voice production (see also

section 2). On the basis of this theory, several authors have suggested that the communication of emotion in music may mimic the emotion-induced physiological changes leading to changes in voice timbre, pitch, loudness, or rate. For instance, Juslin & Laukka (2003, p. 799) suggested that "musicians communicate emotions to listeners on the basis of the principles of vocal expression of emotion".

Indeed, to a large extent, emotions are expressed through shared acoustic correlates in music and language (Coutinho & Dibben, 2012; Juslin & Laukka, 2003). For example, both music and speech express heightened arousal through an increase in either tempo or loudness (Ilie & Thompson, 2006). Moreover, researchers have observed that smaller melodic pitch intervals tend to be associated with sadness in both speech and music (Bowling, Sundararajan, Han, & Purves, 2012; Curtis & Bharucha, 2010). However, one notable difference between music and speech is that whereas high-pitched speech is associated with increased arousal, there is no clear association between pitch height and music-induced arousal (Ilie & Thompson, 2006).

The commonalities between music and language are perhaps best exemplified in infant-directed speech or *motherese*, which is characterized by a simplified grammar, an exaggerated prosody and a repetitive structure. Infants prefer motherese speech to adult-directed speech (Fernald, 1985) and seem to extract information from its melodic patterns in the absence of semantic content (Fernald, 1992). Moreover, infants display more interest in maternal singing than in maternal speech (Nakata & Trehub, 2004; Trehub & Nakata, 2001). Along these lines, Juslin and Laukka (2003) noted that the developmental curve regarding the identification of emotions in music by infants and young children parallels that of vocal expression. There is also mounting evidence that the explicit recognition of emotions in both musical stimuli and nonsensical speech activates similar brain regions (Escoffier, Zhong, Schirmer, & Qiu, 2013).

As already noted by Bolinger (1978), speakers from different cultures use similar intonation patterns to convey emotions. Researchers have empirically confirmed that people can decode the emotional meaning in an utterance spoken in an unfamiliar language (Thompson & Balkwill, 2006), although a meta-analysis shows that the recognition accuracy of the vocal expression of emotions is lower across cultures than within the same culture (Juslin & Laukka, 2003). Analogously, people can recog-

nize emotions conveyed by music from unfamiliar cultures (Balkwill & Thompson, 1999; Fritz et al., 2009), suggesting that some acoustic features are associated with specific emotions (Balkwill & Thompson, 1999).

However, music-induced emotions also rely on domain-specific mechanisms, primarily those based on musical expectations induced by pitch structures (harmonic and/or melodic, Juslin & Västjäll, 2008). Thus, while 3-year-old children can detect basic musical emotions above chance accuracy (Kastner & Crowder, 1990), children up to the age of five years rely on tempo but are unable to use *mode* (e.g., minor versus major scales associated with Western common-practice tonality) to decode emotions (Dalla Bella, Peretz, Rousseau, & Gosselin, 2001). Nevertheless, although pitch systems are more culturally determined, basic categories of musically-expressed emotions can be recognized across cultures. The universal conservation of arousal-inducing factors such as tempo, sound amplitude and pitch height across musical cultures was noted by Brown and Jordania (2013).

The link between emotion communication in music and speech can also be observed by examining the impact of musical expertise on the recognition of emotions in speech. On the one hand, Lima and Castro (2011) have shown that musical expertise enhances the recognition of emotion in speech prosody. On the other hand, individuals suffering from *congenital amusia*, a deficit related mainly to pitch perception, find it more difficult than matched controls to decode emotional prosody in semantically neutral sentences (Thompson, Marin, & Stewart, 2012).

Recent results suggest that the observation that the recognition of emotions in music and speech share similar mechanisms may be generalized to a wide range of environmental acoustic stimuli. For instance, Ma and Thompson (2015) showed that changes in acoustic attributes such as frequency, intensity, or rate (i.e., speed relative to the original version of the stimulus) affect the subjective emotional evaluation of environmental sounds including human and animal-produced sounds as well as machine and natural sounds in much the same way as with music and speech. These findings suggest that the subjective emotional evaluation of auditory stimuli may have an even more general basis than previously thought, insofar as basic attributes such as frequency and intensity are considered.

A clue regarding the nature of this general mechanism may be found in developmental research comparing emotion recognition in visual versus auditory displays. Siu and Cheung (2017) showed that the ability to recognize facial emotions was correlated with the ability to assess the emotional congruency of music-face displays in 20-month-old infants. Notably, these two abilities did not correlate with parental income, quality of parent-child interaction, or language skills, suggesting that these abilities may be subsumed by a common capacity to assess emotion from social cues. Along the same lines, recent findings indicate that the impairment in the detection of emotional prosody observed in congenital amusia also extends to nonverbal vocalizations and even to visual displays of facial emotions (Lima et al., 2016).

6 CONCLUSION

The theoretical and empirical studies reviewed in this chapter point to the ability for emotion communication in nonhuman vocalizations as the biological basis of the ability to process sound modulation in human music and language. Specifically, building on these studies, we hypothesized that the ability to process emotional states conveyed through the voice is a key aspect in the evolution of language. We propose that the systematic investigation of emotional arousal and valence lays out a fertile research venue that may favor comparative work across animal communication, music and language. This may, in turn, provide crucial insights for a fine-tuned analysis of the processes underlying language evolution.

In order to enhance our understanding of the role of vocal communication of emotions in the evolution of language, future research should investigate the ability for music through a *comparative adaptationist lens* across animal species. Furthermore, a fecund path of investigation can be envisioned by extending the analysis of emotional communication to domains other than the auditory and visual ones, focusing, for instance on the effect of chemo-signals (de Groot, Smeets, Kaldewaij, Duijndam, & Semin, 2012). Moreover, the emotional content of proto-musical behaviors in animals should be explored, with the aim to gather further evidence on the adaptive role of these behaviors in animals (Bryant, 2013;

Gingras, 2017). Finally, this research would benefit from the investigation of laughter, cries, and screams in humans. These are key nonverbal human behaviors that, similarly to nonhuman animal vocalizations, are produced in emotional contexts and as an automatic reaction, often within interactional dynamics. Hence, the study of these behaviors may provide a fruitful window for the investigation of the animal nature of human spoken utterances (Bryant & Aktipis, 2014).

To conclude, we suggest that the investigation of emotion processing in species spanning all classes of vocalizing animals, and across communicative channels and human cultures, will certainly help unravel the adaptive value of emotional communication. This line of research will accelerate our understanding of the biological roots of the human ability for language, thus fertilizing current debates on the evolution of language.

ACKNOWLEDGEMENTS

PF was supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). BG was supported by grant 241135 from the Hypo-Tirol Bank.

REFERENCES

Anderson, D. J., & Adolphs, R. (2014). A framework for studying emotions across species. *Cell, 157*, 187–200.

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*, 2051–2056.

Belin, P., Fecteau, S., Charest, I., Nicastro, N., Hauser, M. D., & Armony, J. L. (2008). Human cerebral response to animal affective vocalizations. *Proceedings of the Royal Society B: Biological Sciences, 275*, 473–481.

Bolinger, D. (1978). Intonation across languages. In J. Greenberg, C. A. Ferguson, & E. A. Moravcsik (Eds.), *Universals in human language: Vol. 2. Phonology* (pp. 472–524). Palo Alto: Stanford University Press.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics, 79*, 333–343.

Bowling, D. L., Sundararajan, J., Han, S. E., & Purves, D. (2012). Expression of emotion in Eastern and Western music mirrors vocalization. *PLoS ONE, 7*, e31942.

Bowling, D. L., Gingras, B., Han, S., Sundararajan, J., & Opitz, E. C. L. (2013). Tone of voice in emotional expression: Relevance for the affective character of musical mode. *Journal of Interdisciplinary Music Studies, 7*, 29–44.

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

Briefer, E. F., Maigrot, A.-L., Mandel, R., Freymond, S. B., Bachmann, I., & Hillmann, E. (2015). Segregation of information about emotional arousal and valence in horse whinnies. *Scientific Reports, 4*, 9989.

Briefer, E. F., Tettamanti, F., & Mcelligott, A. G. (2015). Animal studies repository emotions in goats: Mapping physiological, behavioural and vocal profiles. *Animal Behaviour, 99*, 131–143.

Brown, S. (2001). Are music and language homologues? *Annals of the New York Academy of Sciences, 930*, 372–374.

Brown, S. (2017). A joint prosodic origin of language and music. *Frontiers in Psychology, 8*, 1894.

Brown, S., & Jordania, J. (2013). Universals in the world's musics. *Psychology of Music, 41*, 229–248.

Brown, S., Martinez, M. J., & Parsons, L. M. (2006). Music and language side by side in the brain: A PET study of the generation of melodies and sentences. *European Journal of Neuroscience, 23*, 2791–2803.

Bryant, G. A. (2013). Animal signals and emotion in music: Coordinating affect across groups. *Frontiers in Psychology, 4*, 990.

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

Charlton, B. D., Filippi, P., & Fitch, W. T. (2012). Do women prefer more complex music around ovulation? *PloS ONE, 7*, e35626.

Coutinho, E., & Dibben, N. (2012). Psychoacoustic cues to emotion in speech prosody and music, *Cognition and Emotion*, *27*, 658–684.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences USA*, *114*, 7900–7909.

Cutler A., Oahan D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201.

Darwin, C. (1871). *The descent of man and selection in relation to sex*. London: Murray.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: John Murray.

de Carvalho, A., Dautriche, I., Lin, I. & Christophe, A. (2017). Phrasal prosody constrains syntactic analysis in toddlers. *Cognition, 163*, 67–79.

de Groot, J. H., Smeets, M. A., Kaldewaij, A., Duijndam, M. J., & Semin, G. R. (2012). Chemosignals communicate human emotions. *Psychological Science, 23*, 1417–1424.

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

Escoffier, N., Zhong, J., Schirmer, A., & Qiu, A. (2013). Emotional expressions in voice and music: Same code, same effect? *Human Brain Mapping, 34*, 1796–1810.

Faragó, T., Andics, A., Devecseri, V., Kis, A., Gácsi, M., & Miklósi, D. (2014). Humans rely on the same rules to assess emotional valence and intensity in conspecific and dog vocalizations. *Biology Letters, 10*, 20130926.

Fedorenko, E., Patel, A., Casasanto, D., & Winawer, J. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition, 37*, 1–9.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8*, 181–195.

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In H. Papousek, U. Jürgens, & M. Papousek (Eds.), *Nonverbal vocal communication: Comparative and developmental aspects* (pp. 262–282). Cambridge: Cambridge University Press.

Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology, 7*, 1393.

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., … Güntürkün, O. (2017a). 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*, 20170990.

Filippi, P., Gingras, B., & Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Frontiers in Psychology, 5*, 1468.

Filippi, P., Gogoleva, S. S., Volodina, E. V., Volodin, I. A., & de Boer, B. (2017b). Humans identify negative (but not positive) arousal in silver fox vocalizations: Implications for the adaptive value of interspecific eavesdropping. *Current Zoology, 63*, 445–456.

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & de Boer, B. (2017c). More than words (and faces): Evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion, 31*, 879–891.

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

Fitch, W. T. (2012). The biology and evolution of rhythm: Unravelling a paradox. In P. Rebushat, M. Rohrmeier, J. A. Hawkins, & I. Cross (Eds.), *Language and music as cognitive systems* (pp. 73–95). Oxford: Oxford University Press.

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

Foxton, J. M., Dean, J. L., Gee, R., Peretz, I., & Griffiths, T. D. (2004). Characterization of deficits in pitch perception underlying 'tone deafness'. *Brain, 127*, 801–810.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., & Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology, 19*, 573–576.

Gamba, M., Torti, V., Estienne, V., Randrianarison, R. M., Valente, D., Rovara, P., ... Giacoma, C. (2016). The indris have got rhythm! Timing and pitch variation of a primate song examined between sexes and age classes. *Frontiers in Neuroscience, 10*, 249.

Geissmann, T. (2000). Gibbon songs and human music from an evolutionary perspective. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 103–123). Cambridge: MIT Press.

Gingras, B. (2017). Music across the species. In R. Ashley & R. Timmers (Eds.), *Routledge companion to music cognition* (pp. 391–402). Oxford: Routledge.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language, 51*, 548–567.

Hagen, E. H., & Bryant, G. A. (2003). Music and dance as a coalition signaling system. *Human Nature, 14*, 21–51.

Hockett, C. F. (1960). Logical considerations in the study of animal communication. In W. E. Lanyon & W. N. Tavolga (Eds.), *Animal sounds and communication* (pp. 392–430). Washington: American Institute of Biological Sciences.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception, 23*, 319–330.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*, 770–814.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral & Brain Sciences, 31*, 559–575.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341–350.

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130302–20130302.

Lima, C. F., Brancatisano, O., Fancourt, A., Müllensiefen, D., Scott, S. K., Warren, J. D., & Stewart, L. (2016). Impaired socio-emotional processing in a developmental music disorder. *Scientific Reports, 6*, 34911.

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion, 11*, 1021–1031.

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

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain, 133*, 1682–1693.

Ma, W., & Thompson, W. F. (2015). Human emotions track changes in the acoustic environment. *Proceedings of the National Academy of Sciences USA, 112*, 14563–14568.

Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology, 8*, 211.

Martinet, A. (1980). *Eléments de Linguistique Générale*. Paris: Armand Collin.

McComb, K., Taylor, A. M., Wilson, C., & Charlton, B. D. (2009). The cry embedded within the purr. *Current Biology, 19*, R507-R508.

Mendl, M., Burman, O. H. P., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences, 277*, 2895–2904.

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology, 3*, 76.

Mithen, S. (2005). *The singing neanderthals: The origins of music, language, mind, and body*. Cambridge: Harvard University Press.

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

Nakata, T., & Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development, 27*, 455–464.

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature, 1*, 261–289.

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition, 30*, 583–593.

Panksepp, J., & Trevarthen, C. (2009). The neuroscience of emotion in music. In S. Malloch & C. Trevarthen (Eds.), *Communicative musicality: Exploring the basis of human companionship* (pp. 105–146). Oxford: Oxford University Press.

Patel, A. D. (2006). Musical rhythm, linguistic rhythm, and human evolution. *Music Perception, 24*, 99–104.

Patel, A. D. (2010). Music, biological evolution, and the brain. In M. Bailar (Ed.), *Emerging disciplines* (pp. 91–144). Houston: Rice University Press.

Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition and Emotion, 25*, 834–853.

Pongrácz, P., Molnár, C., & Miklósi, Á. (2006). Acoustic parameters of dog barks carry emotional information for humans. *Applied Animal Behaviour Science, 3*, 228–240.

Reichert, M. S. (2013). Sources of variability in advertisement and aggressive calling in competitive interactions in the grey treefrog, Hyla versicolor. *Bioacoustics, 22*, 195–214.

Rousseau, J. J. (1781/1998). Essay on the origin of languages. In J. Scott (Ed.), *Essay on the origin of languages and writings related to music*. Hanover: University Press of New England.

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

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 7–55). New York: Academic Press.

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 USA, 107*, 2408–2412.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological Science, 26*, 354–356.

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

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*, 695–729.

Scherer, K. R. (2009). Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society of London B, 364*, 3459–3474.

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research, 14*, 228–233.

Schirmer A., & Kotz, S. A. (2003). ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience, 15*, 1135–1148.

Schön, D., Boyer, M., Moreno, S., Besson, M., & Peretz, I. (2008). Songs as an aid for language acquisition. *Cognition, 106*, 975–983.

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*, 341–349.

Seyfarth, R. M., & Cheney, D. L. (2003). Meaning and emotion in animal vocalizations. *Annals of the New York Academy of Sciences, 1000*, 32–55.

Siu, T. S. C., & Cheung, H. (2017). Infants' sensitivity to emotion in music and emotion-action understanding. *PloS ONE, 12*, e0171023.

Soderstrom, M., Seidl, A., Nelson, D. G. K., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*, 249–267.

Spencer, H. (1857). The origin and function of music. *Fraser's Magazine, 56*, 396–408.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences USA, 106*, 10587–10592.

Stoeger, A. S., Baotic, A., Li, D., & Charlton, B. D. (2012). Acoustic features indicate arousal in infant giant panda vocalisations. *Ethology, 118*, 896–905.

Stoeger, A. S., Charlton, B. D., Kratochvil, H., & Fitch, W. T. (2011). Vocal cues indicate level of arousal in infant African elephant roars. *The Journal of the Acoustical Society of America, 130*, 1700–1710.

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass, 2*, 1824–1843.

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

Templeton, C. N., Greene, E., & Davis, K. (2005). Allometry of alarm calls: Black-capped chickadees encode information about predator size. *Science, 308*, 1934–1937.

Thompson, W. F., & Balkwill, L. L. (2006). Decoding speech prosody in five languages. *Semiotica, 158*, 407–424.

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences USA, 109*, 19027–19032.

Trehub, S. E., & Nakata, T. (2001). Emotion and music in infancy. *Musicae Scientiae, 5*, 37–61.

Van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *The Quarterly Journal of Experimental Psychology, 58*, 251–273.

Yoshida, S., & Okanoya, K. (2005). Evolution of turn-taking: A bio-cognitive perspective. *Cognitive Studies, 12*, 153–165.

Yuan, J., Zhang, Q., Chen, A., Li, H., Wang, Q., Zhuang, Z., & Jia, S. (2007). Are we sensitive to valence differences in emotionally negative stimuli? Electrophysiological evidence from an ERP study. *Neuropsychologia, 45*, 2764–2771.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6*, 37–46.

Zimmermann, E., Leliveld, L., & Schehka, S. (2013). Toward the evolutionary roots of affective prosody in human acoustic communication: A comparative approach to mammalian voices. In E. Altenmüller, S. Schmidt, & E. Zimmermann (Eds.), *Evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man* (pp. 116–132). Oxford: Oxford University Press.