Emotional Voice Intonation: A Communication Code at the Origins of Speech Processing and Word-Meaning Associations?

Piera Filippi

Journal of Nonverbal Behavior (2020) 44:395–417
https://doi.org/10.1007/s10919-020-00337-z

REVIEW PAPER

Emotional Voice Intonation: A Communication Code at the Origins of Speech Processing and Word-Meaning Associations?

Piera Filippi[1,2,3,4,5]

Published online: 21 July 2020

Abstract
The aim of the present work is to investigate the facilitating effect of vocal emotional intonation on the evolution of the following processes involved in language: (a) identifying and producing phonemes, (b) processing compositional rules underlying vocal utterances, and (c) associating vocal utterances with meanings. To this end, firstly, I examine research on the presence of these abilities in animals, and the biologically ancient nature of emotional vocalizations. Secondly, I review research attesting to the facilitating effect of emotional voice intonation on these abilities in humans. Thirdly, building on these studies in animals and humans, and through taking an evolutionary perspective, I provide insights for future empirical work on the facilitating effect of emotional intonation on these three processes in animals and preverbal humans. In this work, I highlight the importance of a comparative approach to investigate language evolution empirically. This review supports Darwin's hypothesis, according to which the ability to express emotions through voice modulation was a key step in the evolution of spoken language.

Keywords Language evolution · Emotional intonation · Animal communication · Speech sound processing · Word-meaning association

✉ Piera Filippi
pie.fiippi@gmail.com

[1] Present Address: Department of Comparative Language Science, University of Zurich, 8032 Zurich, Switzerland

[2] Present Address: Center for the Interdisciplinary Study of Language Evolution, University of Zurich, 8032 Zurich, Switzerland

[3] Institute of Language, Communication and the Brain, Centre National de la Recherche Scientifique, Aix-Marseille Université, Aix-en-Provence, France

[4] Laboratoire Parole et Langage, LPL UMR 7309, Centre National de la Recherche Scientifique, Aix-Marseille Université, Aix-en-Provence, France

[5] Laboratoire de Psychologie Cognitive LPC UMR 7290, Centre National de la Recherche Scientifique, Aix-Marseille Université, Marseille, France

Introduction

A Comparative Approach to the Evolution of Human Language: Insights from Empirical Studies

In the last few decades, a surge of studies has emerged with the aim of applying an empirical and comparative approach to animal vocal communication as a way to enhance our understanding of the evolution of the mechanisms enabling human language (hereafter, language). The present work adopts this approach, and includes the following methodological steps: (a) identifying and defining core abilities involved in language; (b) tracing the presence of these abilities in multiple closely-related or phylogenetically independent species; (c) analyzing the factors that may have boosted the evolution of these abilities into their current form in language.

This method, which assumes that core mechanisms underpinning language are broadly shared across nonhuman animals (hereafter, animals), sheds light on the evolutionary link between language and other animal communication systems and thus on the biological foundations of language. Specifically, two main classes of shared traits provide information on the evolution of mechanisms involved in language: "homologies" and "analogies" (see Fitch 2017). Homologies are traits that are shared between different species and that were present in their last common ancestor. This type of shared trait informs the direct genetic inheritance underpinning the studied traits, and its phylogenetic path. For instance, humans and chimpanzees have the ability to mentally represent the goals, intentions, perceptions, and knowledge of other individuals (Fitch et al. 2010; Schmelz et al. 2011; Tomasello et al. 2003). It is, therefore, most parsimonious to suggest that, rather than evolving convergently in a relatively short period of evolutionary time, this trait was present in their last common ancestor, which lived between 6 and 7 million years ago (Kumar et al. 2005). Analogies, on the other hand, are traits that were not present in the last common ancestor between the focus species, but evolved convergently, as a result of similar selective pressures. Hence, the study of this type of shared mechanism provides insights into their adaptive function. An example of an analogous trait is the ability to arrange notes within songs, which humans share with songbirds (Berwick et al. 2011). Much research suggests that this ability, which serves as sexual advertisement in both groups, has evolved under comparable sexual selection pressures (Charlton 2014; Darwin 1871; Miller 2000).

By comparing data across species, this approach sheds light on anatomical, cognitive, and neural commonalities between humans and other animals that can be identified as factors enabling the emergence of linguistic communication in humans. In this regard, it is important to emphasize that the uniqueness of language might not be endowed by one or more core mechanisms that are specific to humans. On the contrary, language might have evolved from the integration of multiple mechanisms, each of which can be individually traced (sometimes in a simpler form) in at least another animal species (Fitch 2010, 2017).

Here, I will take a cross-species, comparative approach to studying language evolution by examining three core abilities underpinning language which are, to some extent, shared with nonhuman species (Fitch 2017; Fitch and Zuberbühler 2013; Rendall et al. 2009; Townsend et al. 2018): (a) the ability to identify and produce phonemes; (b) the ability to process compositional rules underlying vocal utterances; (c) the ability to associate vocal sounds with meanings. Importantly, I will highlight the importance of a key communicative factor, namely emotional intonation of the voice, with the aim to shed light on its facilitating effect on the evolution of these three cognitive abilities underpinning language.

Previous research has suggested that the expression of emotions through voice modulation or musical communication, which has been attested across multiple animal species, might have paved the way for the emergence of language in the first hominids (Altenmüller et al. 2013; Brown 2017; Darwin 1871; Filippi 2016; Filippi and Gingras 2018; Filippi et al. 2019; Fitch 2010; Panksepp 2009; Thompson et al. 2012). However, empirical studies addressing the facilitating effect of emotional intonation on each of these three core abilities within a cross-species and evolutionary perspective have never been conducted. In fact, finding this facilitating effect in animal species would provide support for the hypothesis suggested here, namely that emotional intonation might have boosted the evolution of the ability to process phonemes and combinatorial structures, and to associate words with meanings out of comparable abilities reported in animal species. Specifically, in the present work, I suggest that emotional intonation might have boosted the evolution of these abilities, facilitating cognitive processes such as selective attention, perception, memorization, and learning.

In support of this hypothesis, I will firstly review research attesting the presence of the ability to identify and produce phonemes, process compositional rules, and associate vocal sounds with meanings in animals. Secondly, I will review studies indicating that emotional vocalizations are used as a communication code across a wide variety of animal species (cf. Darwin 1872). Thirdly, I will link this research to recent work on the facilitating effect of emotional intonation of the voice on the human ability to perceive speech sounds within compositional structures and associate words with meanings. Finally, I will integrate these studies within a unified framework on the facilitating effect of emotional intonation on language evolution, suggesting specific research questions that can be addressed empirically within a cross-species perspective.

Language-Related Abilities and Emotional Intonation in Animals

The Animal Ability to Identify and Produce Phonemes

Much research has addressed phoneme identification—i.e., fine-tuned perceptual discrimination of vowel- and consonant-like sounds in animals. In this regard, a study reported that one chimpanzee, who had long been exposed to spoken English before being tested, was able to recognize spoken words, even when spectrally degraded (Heimbauer et al. 2011). Further work shows that animals can learn categorical discrimination of distinct phonemes along an acoustic continuum. For instance, macaques (*Macaca mulatta*) (Kuhl and Padden 1983) and budgerigars (*Melopsittacus undulatus*) (Dooling and Brown 1990) can learn to discriminate between voiced and voiceless consonants in the pairs /ba/-/pa/, /da/-/ta/, /ga/-/ka/. Similarly, chinchillas (*Chinchilla laniger*), a mammalian species with auditory abilities similar to humans, can be trained to discriminate a voiced plosive consonant, /d/, from a voiceless one, /t/ in the initial position of a syllable (Kuhl and Miller 1975).

In addition to research on animals' ability to learn to identify fine-grained differences between phonemes, extensive research has addressed their ability to produce phoneme-like sounds. A fundamental theory in the study of animal vocal production is the so-called source-filter theory, which identifies two main factors affecting the vocal output: the "source" and the "filter" (Fitch 2000; Titze 1994). The source of vocal sound production is the larynx in mammals, amphibians, and reptiles, and the syrinx in birds. Specifically, vocal sound is generated by tissue vibrations stimulated by the

passage of air through the vocal folds, in the source. The lowest frequency of the vocal folds' opening-closing cycles determines the fundamental frequency of the vocal sound (F0), which corresponds to the tonal sensation of the voice's pitch. Subsequently, the sound reaches the supralaryngeal vocal tract, i.e., the filter, where certain frequencies are enhanced while others are attenuated by articulation of various parts of the filter, e.g., lips, or tongue. This results in concentrations of acoustic energy in particular frequency bands (called 'formants'), which are perceivable in vowels and consonants (Fant 1960). For instance, if you produce a sequence of different vowels, equal in duration, F0, and amplitude, the perceived acoustic variation is resultant of the difference in formant frequencies.

Following Lieberman et al. (1969), until the last two decades it was commonly assumed that mammals (including primates) are not able to articulate sounds included in human speech due to an anatomical limitation in the filter, namely a heightened larynx. This has been argued to impact the range of articulatory movements in the vocal tract, and hence the formants that could be produced. However, a recent growing body of converging data from empirical studies and computer models of animal vocal production has been undermining Lieberman's hypothesis. For instance, research shows that, when resting, the larynx of red and fallow deer (*Cervus elaphus* and *Dama dama*, respectively) is in a position comparable to that of humans, and retracts even lower during vocalization (Fitch and Reby 2001). Furthermore, Boë et al. (2017) reported that vocalizations of baboons (*Papio papio*) have the formant structure of human [ɨ æ ɑ ɔ u] vowels. This finding suggests that, unless the ability to produce these vowels emerged independently in humans and baboons, the ability to articulate vowel-like sounds may be traced to the last common ancestor from which humans and Cercopithecoidea diverged, about 25 million years ago (Stevens et al. 2013). Consistent with this work, a study adopting a computer model based on vocal tract configurations of living rhesus macaques (*Macaca mulatta*) confirmed that the primate vocal apparatus is potentially capable of producing human-like vowel sounds, as well as a variety of consonants, including stop consonants that are widely shared across languages (e.g., /h/, /m/, /w/, /p/, /b/, /k/, and /g/) (Fitch et al. 2016). This study implies that the human ability for speech required the evolution of specific neural connections between forebrain and laryngeal muscles, rather than anatomical changes in the vocal apparatus. Importantly, this research supports findings from previous studies hypothesizing that, in humans, direct neural connections between the laryngeal motor cortex (LMC) and the brainstem laryngeal motoneurons (which are, in turn connected to the laryngeal muscles), as well as the location of the LMC in the primary motor cortex (as opposed to its location in the premotor cortex in nonhuman primates), might have been key evolutionary steps enabling the ability to control complex laryngeal movements involved in producing learned vocal utterances (Jürgens 2002, 2009; Simonyan 2014; Simonyan and Horwitz 2011). In monkeys, and presumably, other nonhuman primates, the LMC is linked only indirectly—namely, through the reticular formation—to the laryngeal motoneurons in the brainstem (Simonyan 2014). Critically, their innate vocal production seems to be enabled by a specific voice control system in the brain, involving the brain stem and spinal cord sensorimotor phonatory nuclei only (Simonyan and Horwitz 2011). This might explain why the destruction of the LMC region in monkeys does not affect their innate vocal production (Simonyan 2014), which can take place without involving voluntary coordination and control of laryngeal muscles. Comparative studies on primate vocal production are a clear example of how research can help shed light on which trait was present (the anatomy of the vocal tract) and which was still missing (e.g., direct neural connections from motor cortical regions onto laryngeal motoneurons) before full-blown speech evolved.

Strikingly, this picture could be placed into a broader evolutionary scale to gain a wider perspective on the selective pressures enabling the emergence of the neural connections necessary for articulating human speech sounds. Indeed, although much research reports on animals' ability to produce novel sounds and sound combinations by imitation (see section below), only a few species of mammals and birds seem to be able to learn to modulate their vocal tract to imitate words and sentences in existing human languages, e.g., Asian elephant, *Elephas maximusi*, (Stoeger et al. 2012); captive harbor seals, *Phoca vitulina*, (Ralls et al. 1985); gray seals, *Halichoerus grypus* (Stansbury and Janik 2019); and birds (Grey parrot, *Psittacus erithacus*, Pepperberg 2010; mynah bird, *Acridotheres tristis*, Stefanski and Klatt 1974). Social bonding can be identified as a potential selective factor boosting the ability to learn to produce novel sounds that are not included in the given species' vocal repertoire (Stoeger et al. 2012). Hence, this research provides crucial insights on the key role of social pressures in language evolution and is consistent with work suggesting that social bonding (which is likely highly connected to the use of vocal emotional intonation in inter-individual communications) might have promoted the evolution of neural connections enabling the production of human spoken language (Dunbar 2003).

Generally, although it is plausible that species that are able to produce speech sounds can equally discriminate them at a perceptual level (cf. Pulvermller 2005), more research is needed on this topic within a cross-species perspective. This will favor a broader understanding of the evolutionary pressures behind neural and anatomical predispositions for identifying and producing phonemes.

The Animal Ability to Process Compositional Rules in Vocal Utterances

A number of cross-species studies revolve around the ability to process vocal sequences according to compositional rules. This line of research aims at understanding the evolutionary precursors and selective pressures that led from the ability to parse simple forms of compositionality, which has been demonstrated in multiple species, to the human ability to parse fully-fledged syntactic systems of languages (Russell and Townsend 2017). Although much research on this topic is still ongoing, our understanding of the evolution of the human ability for syntax has significantly advanced in the last two decades (Collier et al. 2014; Engesser and Townsend 2019; Townsend et al. 2018; Zuberbühler 2020). For instance, it has been shown that Campbell's monkeys (*Cercopithecus campbelli*) add an acoustic modifier (i.e., a sort of affix) to predator-specific alarm calls (Ouattara et al. 2009). In this way, the meaning of the alarm call is no longer perceived as linked to a predator, but to the presence of a general disturbance. Intriguingly, the ability to process compositional structures was also found in birds, providing insights for comparative studies on its evolutionary origins. For example, Engesser et al. (2016) showed that southern pied babblers (*Turdoides bicolor*) respond to combinations of alert and recruitment calls with mobbing-like behavior, while no obvious reaction is elicited by control combinations of foraging and recruitment calls. In a similar study, Suzuki et al. (2016) report that, in Japanese tits (*Parus minor*), the combination of two calls—namely of a call typically eliciting scanning behaviors in the listeners, and a call typically eliciting approach behavior to the caller—results in the combination of these two behaviors, i.e., scanning and approach. In a control experiment, the inversion of these two calls did not elicit any behavior, suggesting that these birds are processing the call combination according to a specific order. Therefore, these studies suggest that the southern pied babblers and the Japanese tits are sensitive to compositional properties of call sequences, and that structural changes impair signal perception. These

systems can be fruitfully compared with compositional structures in language, where variation of words within a sequence (e.g., changing "gimme a break" into "apple a break" or "break a gimme") can turn a well-formed and meaningful spoken utterance into an ill-formed and meaningless sequence of words.

Further cross-species studies on the ability to discriminate syntactical structures have typically adopted artificial grammars that are created following specific formal rules. For instance, Spierings and ten Cate (2016) found that zebra finches (*Taeniopygia guttata*) are able to discriminate units of their own vocal repertoire, arranged in a XYX or XXY structure, and that budgerigars (*Melopsittacus undulatus*) can discriminate and generalize this grammatical rule to novel elements they were never trained on during a previous rule learning phase.

A fundamental strand of comparative research on animals' ability to process compositional structures has attempted to identify the cognitive abilities that enable humans (but not other animal species) to process more complex compositional structures in language. This research builds on the assumption that the human-specific ability to express an open-ended number of thoughts using a finite set of linguistic units relies on recursion (Everaert et al. 2015), i.e., the operation of embedding constituents within constituents of the same kind (Pinker and Jackendoff 2005; cf. Martins 2012). Building on this assumption, Hauser et al. (2002) proposed that the ability to use recursion might be *the* key computational ability that differentiates the syntactical competence of humans from combinatorial abilities found in animals (cf. Bolhuis et al. 2018). Within this conceptual framework, much research has relied on the so-called "Chomsky hierarchy" (Chomsky 1956, 1959) as a way to guide empirical work. This hierarchy provides a theoretical structure to identify and classify different levels of computational powers, each corresponding to a specific "grammar". Each grammar includes a finite number of symbols, rules, and operators to apply to these symbols. One of the aims of this classification is to identify the level of computational power that enables an automaton to process natural languages on a mere mathematical and abstract level, i.e., excluding aspects such as lexical semantics, interactional dynamics, or context. This highly formal character of grammars favors well-controlled cross-species investigations of computational abilities that are foundational to language (O'Donnell et al. 2005). Hence, this research framework enables the investigation of animals' computational capacities along a complexity axis, which includes the computational capacity underpinning natural language processing.

As Fitch and Friederici (2012) explain in their exhaustive and, at the same time, intuitive overview of the formal language theory at the base of Chomsky's hierarchy, a crucial distinction within this hierarchy is between "regular" and "supra-regular" grammars. This distinction is important because it provides a line of demarcation between the computational abilities that are necessary to process very simple structures and those that are necessary to process hierarchical syntactic structures in natural languages. Regular grammars can be computed by the simplest class of automata (called "finite state automata"), using basic computational rules, namely, transition probabilities between a finite number of "states" (e.g., phonemes, syllables, or words). Examples of strings that can be processed by regular grammars are "$(AB)^n$" - where the automaton has to accept an *n* number of "AB" bigrams, or "AB*A"—where any number of B units can occur between the A units at the edges. These basic rules are not enough to process the structural complexity of natural languages (Jäger and Rogers 2012). However, they might suffice to process phonological sequences, an ability that humans might share with other animals (Fitch 2018a). In contrast, "supra-regular" grammars, which include multiple subsets of grammars, rely on more complex rules and computational power

than that required for a finite state automaton. An example of a supra-regular grammar is a context-free grammar, which can be computed by a "pushdown automaton". For instance, the A^nB^n sequence—where a number of B elements follows the same number of A elements—can be processed by this type of automaton, but not by a finite state one (Fitch and Friederici 2012; O'Donnell et al. 2005), which is not able to count and compare (Jäger and Rogers 2012). Crucially, the set of supra-regular grammars vary in the amount of requested computational power that can be used to process dependencies between the constitutive elements of an expression. Importantly, this set includes grammars that can process dependencies within recursive structures, such as $A_1A_2A_3B_3B_2B_1$, where the same pattern AB is nested in itself, following a center-embedded structure (Chomsky 1956, 1959). As Jäger and Rogers (2012) explain, an example of nested dependencies is given by the English construction "neither-nor", repeated multiple times within the same sentence, as in "*Neither* did Mary think she would *neither* go to the cinema *nor* eat pizza, *nor* did I".

The first study to use the distinction between regular and supra-regular grammars to compare humans and animals' (specifically, cotton-top tamarins) was conducted by Fitch and Hauser (2004). In this study, the authors found that cotton-top tamarins are able to process AB^n sequences—i.e., regular grammars, but fail to process A^nB^n sequence - i.e., supra-regular grammars, while humans, as predicted, succeeded in processing both grammars. Following up this work, a number of studies have probed how phylogenetically widespread the ability to process regular and supra-regular grammars is. To date, the majority of studies have found that, multiple species of animals are able to process regular grammars, specifically, $(AB)^n$ sequences (ravens, *Corvus* corax, Reber et al. 2016; kea, *Nestor notabilis*, and pigeons, *Columba livia*, Stobbe et al. 2012; cf. ten Cate and Okanoya 2012) and perceptual dependencies between edge stimuli in AB^nA sequences both in the visual domain (chimpanzees, *Pan troglodytes*, Sonnweber et al. 2015; cotton-top tamarins, *Saguinus oedipus*, Versace et al. 2019) and in the auditory domain (squirrel monkeys, *Saimiri sciureus*, Ravignani et al. 2013; cotton-top tamarins, *Saguinus oedipus*, Newport et al. 2004; common marmosets, *Callithrix jacchus*, Reber et al. 2019). In addition, although some research has suggested that birds are able to process supra-regular grammars (Abe and Watanabe 2011; Gentner et al. 2006), subsequent studies have shown that these birds might have used simple strategies—that do not require any of the computational power at the level of supra-regular automata—to parse these structures (Ravignani et al. 2015; Van Heijningen et al. 2009). However, in a recent study, Jiang et al. (2018) provided, for the first time, compelling evidence that an animal species—specifically, the macaque monkey (*Macaca mulatta*)—is able not only to parse, but also to produce a sequence according to a supra-regular grammar, namely, a "mirror" (context-free) grammar of the form ABC-CBA. Here, the second part of the string is a mirror image of the first part, thus including a center-embedding organization. The authors tested pre-school children on the same task, and found that, compared to monkeys, who needed a massive amount of training to learn the grammar, humans learned to master the grammar with only a little training. These findings suggest that monkeys possess these computational competences, although they do not have the same human inclination to use them (Fitch 2018b).

Here, it is important to stress that much debate is currently ongoing regarding the assumption that recursion is *the* defining computational system of language (Christiansen and Chater 2015; Evans and Levinson 2009; Parker 2007; Perruchet and Rey 2005). Nevertheless, comparative research relying on grammars defined within Chomsky's hierarchy is effective for a systematic investigation of the ability of animals to process different levels

of structural complexities in the vocal domain. This, in turn, may provide key insights into the evolution of the human ability to parse compositional patterns.

But what is the evolutionary advantage of the animals' ability to produce and process compositional structures? Crucially, in animal communication systems, higher levels of structural complexity in compositional structures allows for the transmission of information with greater degrees of complexity compared to vocalizations with simpler structures (Nowicki and Searcy 2014). In this regard, research indicates that higher levels of vocal complexity typically co-occur with the predisposition to learn to articulate a signal by imitating (and modifying) someone else's signal (Nottebohm 2002). Hence, the tendency to learn vocally might have been a key factor in the evolution of the human ability to identify and produce syntactical structures in language. Extensive research has addressed the phylogenic path of the ability for vocal learning, and the selective pressures underpinning its evolution (cf. Martins and Boeckx 2020). In particular, animal research on this topic has mainly focused on three groups of birds (parrots, hummingbirds, and songbirds) (Beecher and Brenowitz 2005; Jarvis 2006). Recently, this line of research has been complemented by studies on phylogenetically distant mammalian species, including terrestrial and marine mammals (e.g., African elephants, *Loxodonta africana*, Poole et al. 2005; Egyptian fruit bat, *Rousettus aegyptiacus*, Prat et al. 2015; humpback whale, *Megaptera novaeangliae*, Cerchio et al. 2001; Californian sea lion, *Zalophus californianus*, Reichmuth and Casey 2014).

The Animal Ability to Associate Vocal Utterances with Meanings

Much research aimed at pinpointing the evolutionary precursors of the human ability for word-meaning association in animal communication systems has focused on animals' ability to understand the link between vocal utterances and their meaning—i.e., the information they express or refer to (Dawkins and Krebs 1978; Macedonia and Evans 1993; Marler et al. 1992; Wiley 1983). For instance, studies indicate a strong link between acoustic features of the signal and information related to the body size and the emotional state of the signaler (Owren and Rendall 2001). Body size has been demonstrated to be reliably cued by formant-structure of mammalian vocalizations. Specifically, individuals with bigger bodies have lower formant frequencies than smaller individuals (domestic piglets, *Sus scrofa domesticus*, Garcia et al. 2016; koala, *Phascolarctos cinereus*, Charlton et al. 2011; rhesus macaques, *Macaca mulatta*, Fitch 1997; humans, Pisanski et al. 2014; for cross-species studies, see: Bowling et al. 2017; Charlton and Reby 2016; Taylor and Reby 2010). In accordance with these studies, research on the perception of vocal indicators of body size suggests that formants are also the most reliable acoustic parameters for perception of size-related variation in animals (e.g., whooping cranes, *Grus americana*, Fitch and Kelley 2000; red deer, *Cervus elaphus*, Charlton et al. 2007a, b; dog, *Canis lupus familiaris*, Faragó et al. 2010), and between species (Taylor et al. 2008). Similar mechanisms seem to be at play in the perception of body size and related information through acoustic features of the voice in humans. Indeed, research shows that formants are linked to size perception (Ohala 1984; Pisanski et al. 2014; Rendall et al. 2009) and dominance (Puts et al. 2006) in humans, and suggest that back vowels (e.g., /o/, /a/) are associated with big objects and front vowels (e.g., /i/, /e/) are associated with small objects (see Lockwood and Dingemanse 2015a for a review). In addition, Auracher (2017) reports that human participants associate back vowels with larger sizes, aggression, strength, and social dominance, and front vowels with small sizes, weakness, fearfulness, and social subordination.

Interestingly, the author found that, in this association process, the semantic content of the pictures (e.g., elephant vs. rabbit) overwrites the actual size of the depicted objects in this association process—given, for instance, by using an image of the elephant that was relatively smaller than the image of the rabbit.

In addition, Bowling et al. (2017) showed that body size inversely correlates to F0 in a wide variety of mammalian species. This study is consistent with Morton's (1977) "motivational-structural rules" hypothesis, which states that in mammals and birds, harsh, low-frequency vocalizations are used in competitive contexts to signal physical dominance, whereas more tonal, high-frequency vocalizations are used in fearful or appeasing contexts to signal submission. Recent research has extended these findings, suggesting that larynx size (in particular, vocal fold length), which might not be proportional to body size, predicts F0 better than body size (Garcia et al. 2017).

Critically, research found evidence for the ability to process simple spoken sound-meaning associations in animals. Dogs (Kaminski et al. 2004), parrots (Pepperberg 2006), and chimpanzees (Savage-Rumbaugh et al. 1993) have all been shown able to infer which specific object a word refers to. Finally, comparative research on animal communication has described animal calls as "word-like" vocal units in that these calls are associated with specific objects or events akin to the referential nature of human words. For instance, in a very influential study, Seyfarth et al. (1980) suggested that the vervet monkey (*Chlorocebus pygerythrus*) have three distinct alarm calls, each associated with 'snake', 'eagle', and 'leopard' respectively. These calls elicit appropriate behaviors in the listeners, such as looking up upon hearing the call emitted by the signaler in response to the presence of an eagle. More recently, research has revisited these original findings and adopted state-of-the-art techniques for acoustic data analyses (Fischer and Price 2017; Price et al. 2015). These studies highlight that animal calls do not "carry" information on the basis of an arbitrary association between sounds and meanings, as in the case of human words. On the contrary, in primates, vocalizations are genetically determined and are triggered by emotional and cognitive states of the signalers, which are reflected in specific acoustic features of the signal. The perception of these acoustic features, combined with contextual cues, allows listeners to associate the signal with its eliciting stimuli, and subsequently select the appropriate responses (Wheeler and Fischer 2012).

Within the comparative approach proposed here, studies on emotional expression through voice intonation are particularly relevant to the study of the evolution of the ability to associate arbitrary vocal utterances with their meaning. Indeed, as I will describe in the next sections, emotional expressions are widespread across a wide variety of vocalizing animal species (Darwin 1872), and, within humans, across cultures (Barrett and Bryant 2008; Sauter et al. 2015; Scherer et al. 2001). This makes emotional expresssions a good candidate for enhancing our understanding of the dynamics underpinning the evolution of the human ability for speech processing and word-meaning associations.

Vocal Emotional Expression: A Cross-Species Comparative Approach

The study of emotional expression through voice intonation in animals may provide crucial insights to reconstruct the dynamics underpinning language evolution (Darwin 1871; Filippi 2016; Filippi and Gingras 2018; Filippi et al. 2019). Across animal species, emotions serve adaptive functions, favoring actions that promote survival, such as a fight-or-flight response to an attacking predator in the surroundings (Nesse 1990). In addition,

emotional stimuli engage selective attention (Kret et al. 2016) and favor associative learning in animals (McGaugh 2004; Seymour and Dolan 2008).

Importantly, changes in emotional states may create tension in the muscles involved in vocal phonation, as for instance, those involved in respiration (diaphragm and intercostal muscles) and, importantly, in the vocal folds (Ladefoged 1996; Titze 1994). These changes affect vocal sound production, generating audible differences between vocalizations emitted in intense emotional states and those emitted in less intense ones. In line with Darwin (1871, 1872), I assume that the mechanisms of production of emotional vocalizations might be evolutionary conserved across species (Filippi et al. 2017a), and were, presumably, in place at the time the first hominids diverged from the last common ancestor between humans and chimpanzees.

Multiple studies have addressed emotional communication in animals, focusing on discrete emotions, such as fear or rage (Camperio Ciani 2000; Forkman et al. 2007). However, as Mendl et al. (2010) observe, this approach may narrow down the range of emotions that can be assessed in animals within a comparative approach. In fact, a research framework that is best suited for comparative analyses is offered by the dimensional approach (Russell 1980), in which emotions are described according to two dimensions: arousal (low/calm or high/excited) and valence (positive or negative). Crucially, the investigation of arousal, which relies on quantitative measures of physiological correlates of emotional activation of signalers, serves cross-species comparison very well (Briefer 2012). In addition, this quantitative approach allows researchers to identify vocal indicators of arousal levels in the vocalizing animals. In an extensive review, Briefer (2012) reports that across the vast majority of studied mammalian species (including humans, see Banse and Scherer 1996; Johnstone and Scherer 2000), heightened levels of arousal are expressed through energy distribution towards higher frequencies, higher frequency-related parameters, amplitude contour, vocalization rate, and lower inter-vocalization interval. In addition to studies on emotional arousal expression in mammals, research reports that a songbird species, the black-capped chickadee (*Poecile atricapillus*) encodes the degree of threat posed by small or large predators—which presumably trigger low and high arousal emotional states, respectively—in their calls (Templeton et al. 2005). Specifically, the higher the threat, the higher the number of D notes at the end of their call.

The ability to identify emotional states in vocal signals, which may be produced within social interactions (Altenmüller et al. 2013; Bryant 2013), favors survival of conspecifics in contexts such as territory defense or predation (Cross and Rogers 2006; Desrochers et al. 2002; Owings and Morton 1998). In addition, survival chances may be favored by "eavesdropping" on another species' alarm calls although acoustically different from their own (de Boer et al. 2015; Fallow et al. 2011; Kitchen et al. 2010; Lea et al. 2008; Magrath et al. 2009). In line with these studies, recent work has found that humans and black-capped chickadees can discriminate high versus low arousal calls across a large variety of vocalizing species, spanning all classes of vocalizing vertebrates (Congdon et al. 2019; Filippi et al. 2017a, b).

Finally, the ability to identify emotional activation in the signaler (conspecific or heterospecific) may determine survival of newborns, who can express their needs very effectively through voice intonation, thus enabling their caregivers to respond appropriately (Marmoset monkey, *Callithrix jacchus*, Tchernichovski and Oller 2016; Zhang and Ghazanfar 2016; human, Fernald 1992). Interestingly, Lingle and Riede (2014) found that mule deer (*Odocoileus hemionus*) and white-tailed deer (*Odocoileus virginianus*) mothers are sensitive to high arousal, negatively-valenced vocalizations of infants of a variety of mammalian species (e.g., mule deer, *Odocoileus hemionus*, bighorn sheep, *Ovis canadensis*, marmots,

Marmota flaviventris, bats, *Lasionycteris noctivagans*, Australian sea lion, *Neophoca cinerea* and Subantarctic fur seals, *Arctocephalus tropicalis*), if the F0 values are within the frequency range produced by infants of their own species.

Taken together, these studies are consistent with Darwin's (1871) hypothesis that emotional communication in animals is produced through mechanisms underpinning voice production that are conserved across phylogenetically distant species. This hypothesis is in line with a growing body of studies attesting to the human ability to identify vocal emotions across widely different cultures (Barrett and Bryant 2008; Sauter et al. 2015; Scherer et al. 2001). Emotional communication might, therefore, be biologically ancient and immune to the influence of cultural dynamics.

In light of the evidence reviewed in this section, it is worth addressing how emotional intonation, as a communication code used across a wide variety of animal species, affects language processing in humans. This line of investigation will provide insights into the dynamics underlying the emergence of language from nonhuman animal communication systems.

Emotional Intonation: Facilitating Effect on Language Processing

The Human Ability to Identify and Produce Phonemes (Within Compositional Structures): Facilitating Effect of Emotional Intonation

"I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own instinctive cries. [...] we may conclude from a widely-spread analogy that this power would have been especially exerted during the courtship of the sexes, serving to express various emotions, as love, jealousy, triumph, and serving as a challenge to their rivals. The imitation by articulate sounds of musical cries might have given rise to words expressive of various complex emotions." (Darwin 1871, p. 56)

In accordance with Darwin's hypothesis on the origins of language, extensive research has identified a positive effect of emotional stimuli on cognition, particularly on attentional, perceptual, and memory resources, which are at the core of language processing (Dolan 2002; Kotz and Paulmann 2011; Storbeck and Clore 2008). Indeed, multiple studies indicate that the presentation of emotional written words, images, or sounds enhances the processing of target stimuli that are presented before or after the given emotional stimulus. This results in higher accuracy in recalling the target stimulus and facilitates associative learning between the emotional stimulus and the target one (Finn and Roediger 2011; Guillet and Arndt 2009; Riegel et al. 2016). For instance, high arousal auditory stimuli—independently from their valence—affect selective attention, favoring perception and memorization of salient visual stimuli (namely, letters with higher contrast font within a set of visually presented letters) (Sutherland and Mather 2018). Importantly, the effects of emotional stimuli on perception, attention, memory, and learning are mediated by primary brain networks in the limbic system that humans and animals share, in particular, the amygdala (Dolan 2002; Phelps and LeDoux 2005; Seymour and Dolan 2008).

Consistent with these studies, research on auditory emotional words show that, in adults, shifts towards higher F0 mean and F0 variation positively affect perceptual salience of these spoken words, engaging attention, and ultimately, favoring their intelligibility (Davis et al. 2017; Dupuis and Pichora-Fuller 2014; Nencheva et al. 2020). Numerous studies

have addressed this topic focusing on the special speech register that human caregivers use when addressing infants (hereafter infant-directed speech or IDS). Crucially, emotional intonation in this type of speech is prominent and effective in conveying communicative functions such as alerting, comforting, alarming, or disapproving (Fernald 1992; Trainor et al. 2000). Fernald et al. (1989) found that, compared to adult-directed speech (ADS), IDS is characterized by higher values related to F0, shorter utterances, and longer pauses. The authors found that this result applies to six different language groups (American English, British English, Japanese, German, Italian, and French), suggesting that voice modulation in IDS is shared across human societies. In addition, expanded pitch contours and longer vowel duration in IDS, compared to ADS (Andruski and Kuhl 1996; Fernald and Simon 1984; Kuhl et al. 1997), favor infants' discrimination of vowel categories (de Boer 2005; Trainor and Desjardins 2002; Werker et al. 2007). Similarly, voice onset time (VOT) in stops (namely, /b/, /d/, /t/, /g/, /k/) are longer in IDS than in ADS (Englund 2005). These findings are corroborated by research showing that 7–8-month-old infants are better at recognizing words spoken in IDS compared to words spoken in ADS (Singh et al. 2009).

Generally, in speech, and, in particular, in the case of IDS, spoken sound identification occurs within sentences, hence, within compositional structures. Indeed, a higher order of language processing at which emotional intonation may play a key role consists of producing and parsing well-formed connections between spoken words or phrases, according to compositional rules. To my knowledge, the effect of emotional intonation on these processes has never been investigated directly. In contrast, extensive research has focused on linguistic prosody, i.e., prosodic structure of utterances that is used, for instance, to recognize words within sentences, to emphasize a particular word in a sentence, or to distinguish a command from a statement or a question (Cutler et al. 1997). Studies suggest that linguistic prosody has a crucial role in bootstrapping syntax comprehension and on marking the beginning and the end of a phrase (Soderstrom et al. 2003). In addition, Gussenhoven (2002, 2016) has addressed anatomical and physiological factors affecting the use of voice intonation to mark questions, utterance start/end, topic continuity, or focus in speech. For instance, he suggested that high pitch typically signals the beginning of an utterance, and low pitches signal their ending. This is given by the fact that, when starting an utterance, the subglottal air pressure in the speaker is higher than towards its end. In accordance with these studies, previous research has shown positive effects of emphasized prosodic features in language comprehension (Klieve and Jeanes 2001) and in perception of interrogative forms (See et al. 2013) in children with hearing deficits.

The Human Ability to Associate Words with Meanings: Facilitating Effect of Emotional Intonation

A critical aspect of language to consider within the present research framework is the ability to associate arbitrary sequences of vocal sounds (i.e., words) with meanings. Notably, word-meaning association is an essential part of word learning, where categorical, conceptual and social factors come into play (Waxman and Gelman 2009). One of the most efficient paradigms to investigate word-meaning association is the cross-situational word learning paradigm, where participants are exposed to a series of visual images containing a target referent, while hearing a target word that always co-occurs exclusively with the corresponding referent (Yu and Smith 2007). Research applying this paradigm to an artificial language learning experiment suggests that marking a target word with IDS typical F0 exaggerated contours benefits the learners' ability to associate target word and target

visual referents into a word-meaning pairing (Filippi et al. 2014, 2017c). This research is consistent with previous work suggesting that IDS-typical F0 prominence facilitates word-meaning mapping in preverbal infants (Ma et al. 2011; cf. Fernald and Mazzie 1991). In addition, much work has focused on the relative prominence of emotional intonation and lexical content within a task of emotional meaning identification. For instance, in a recent study, Filippi et al. (2017d) adopted a Stroop task in which participants had to identify the meaning of an emotional word by either focusing on emotional intonation (hence, ignoring lexical content) or the other way around, by focusing on lexical content, while ignoring emotional intonation. In this task, the two channels can be congruent—as in the case of the word "happy" spoken with a happy intonation—or incongruent, as in the case of happy", spoken with a sad intonation. The authors found that, in the incongruent condition, when participants had to ignore emotional intonation and identify the emotional meaning focusing on lexical content, they were significantly less accurate than when they had to ignore lexical content and identify the emotional meaning conveyed by intonation. These findings are echoed by multiple studies reporting the higher salience of emotional intonation over lexical units also at a brain level (Schirmer et al. 2002; Schirmer and Kotz 2006). The attested prominence of emotional intonation over lexical content corroborates the hypothesis that the ability to process emotional content through voice intonation is older than phonetic processing, and might have favored its emergence. Within this research framework, Aryani and Jacobs (2018) addressed the interaction between semantic content and phonemes iconically associated with high emotional arousal, for instance plosives or hissing sibilants (as assessed in Aryani et al. 2018). The authors found that words where semantic content and constituent phonemes (e.g., the plosive consonant /k/ in "Krieg" [war]) are congruent in the expression of arousal are processed faster and more accurately. These findings are consistent with further studies showing facilitating effects at a neural level, provided by the interaction between phonemes and semantic content in emotional word processing tasks (Aryani et al. 2019).

Conclusions

Taken together, the studies reviewed here are in line with the hypothesis that emotional intonation scaffolded the emergence of the following core abilities involved in language: phonemes' identification and production (within compositional structures) and word-meaning association. In support of this hypothesis, firstly, I reviewed studies tracing the presence of simpler forms of these abilities across a variety of vocalizing animal species. These studies shed light on the evolutionary precursors of the corresponding abilities involved in language, and on the selective pressures driving their emergence. Secondly, I described studies providing compelling evidence of the use of emotional intonation across multiple animal species, including humans. Consistent with this literature, research indicates that emotional stimuli activate the most evolutionarily ancient components of the brain that are shared between humans and animals (Dolan 2002; Phelps and LeDoux 2005; Seymour and Dolan 2008). The link between emotional communication in animals and the linguistic abilities reviewed here is provided by research indicating the enhancing effect of emotional stimuli on cognitive processes involved in language, namely, perception, selective attention, memory and learning in humans (Dolan 2002; Storbeck and Clore 2008). Hence, it is plausible that emotional intonation might have boosted the evolution of abilities involved in language from comparable ones found in animal communication systems. This line of

argumentation was further corroborated by evidence that emotional intonation facilitates the perception of phonemes and spoken word-meaning association in human infants and adults (reviewed above).

This review is in accordance with previous studies that, building on Darwin's work (1871), and within a comparative approach to animal communication, suggest that the expression of emotion through prosodic (or "musical") voice modulation set the stage for the emergence of language (Altenmüller et al. 2013; Brown 2017; Darwin 1871; Filippi 2016; Filippi and Gingras 2018; Filippi et al. 2019; Panksepp 2009; Thompson et al. 2012). Furthermore, the present work extends previous research on the physiological foundations of the use of voice intonation for linguistic purposes (Gussenhoven 2002).

Importantly, this review supports previous research suggesting that the investigation of the dynamics underlying language evolution can take place by integrating empirical evidence from multiple disciplines (Fitch 2017). Within this approach it is beneficial to explore language as a complex ability made of core components whose evolutionary dynamics can be investigated separately, rather than as a monolithic block. In particular, this work opens the avenue for empirical investigation of specific research questions I plan to address in follow-up studies, namely, whether emotional intonation facilitates the following abilities in animals and preverbal humans: (1) producing and identifying phonemes; (2) processing and learning compositional rules in vocal utterances; (3) associating unfamiliar spoken words with their meaning.

In order to enhance our understanding of language evolution (which is an intrinsically multimodal system, Levinson and Holler 2014), studies on the vocal and emotional origins of language need to be integrated with research on primates' abilities for gestural communication (Meguerditchian et al. 2013). Furthermore, it is crucial to bridge this research with investigations on the role of time-coordinated interactions in the emergence of language processing abilities (Filippi 2016; Filippi et al. 2019; Levinson 2016; Ravignani et al. 2019), and on the evolution of pragmatic abilities such as, for instance, theory of mind (Fitch et al. 2010; Scarantino 2018). Finally, within this research framework, comparative work on animal communication abilities needs to be connected to studies tracking language evolution in the hominin line (e.g., Blasi et al. 2019).

To conclude, I would like to highlight that emotional intonation is strongly connected to the social dimension of linguistic communication (Sander et al. 2005), and might have driven the typically human "urge" to create socio-emotional bonds and share information with conspecifics (Fitch 2010). Hence, by investigating emotional communication as a communication code widely used across animal species, and which may have been critical in fostering the emergence of spoken language, we ultimately begin to elucidate a fundamental dimension of humans: the species-specific drive for interpersonal communication.

Acknowledgements Open access funding provided by University of Zurich. The author was supported by the Forschungskredit of the University of Zurich, Grant No. [FK-19-070]. The author is grateful to Sabrina Engesser, Stuart Watson, Simon W. Townsend, and to the anonymous reviewers for their insightful comments and suggestions.

References

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Publishing Group, 14*(8), 0–6. https://doi.org/10.1038/nn.2869.

Altenmüller, E., Schmidt, S., & Zimmermann, E. (Eds.). (2013). *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man*. Oxford: Oxford University Press.

Andruski, J. E., & Kuhl, P. K. (1996). The acoustic structure of vowels in mothers' speech to infants and adults. In *Proceeding of fourth international conference on spoken language processing. ICSLP'96* (vol. 3, pp. 1545–1548). IEEE.

Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why 'piss' is ruder than 'pee'? The role of sound in affective meaning making. *PLoS ONE, 13*(6), e0198430. https://doi.org/10.1371/journal.pone.0198430.

Aryani, A., Hsu, C. T., & Jacobs, A. M. (2019). Affective iconic words benefit from additional sound–meaning integration in the left amygdala. *Human Brain Mapping, 40*(18), 5289–5300. https://doi.org/10.1002/hbm.24772.

Aryani, A., & Jacobs, A. M. (2018). Affective congruence between sound and meaning of words facilitates semantic decision. *Behavioral Sciences, 8*(56), 1–11. https://doi.org/10.3390/bs8060056.

Auracher, J. (2017). Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance. PLoS ONE. https://doi.org/10.1371/journal.pone.0187196.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614.

Barrett, H. C., & Bryant, G. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8*(1), 135–148. https://doi.org/10.1163/156770908x289242.

Beecher, M. D., & Brenowitz, E. A. (2005). Functional aspects of song learning in songbirds. *Trends in Ecology and Evolution, 20*(3), 143–149. https://doi.org/10.1016/j.tree.2005.01.004.

Berwick, R. C., Okanoya, K., Beckers, G. J. L., & Bolhuis, J. J. (2011). Songs to syntax: The linguistics of birdsong. *Trends in Cognitive Sciences, 15*(3), 113–121. https://doi.org/10.1016/j.tics.2011.01.002.

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science, 363*, eaav3218. https://doi.org/10.1126/science.aav3218.

Boë, L. J., Berthommier, F., Legou, T., Captier, G., Kemp, C., Sawallis, T. R., Fagot, J. (2017). Evidence of a vocalic proto-system in the baboon (*Papio papio*) suggests pre-hominin speech precursors. *PLoS ONE, 12*(1), e0169321.

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018). Meaningful syntactic structure in songbird vocalizations? *PLoS Biology, 16*(6), e2005157. https://doi.org/10.1371/journal.pbio.2005157.

Bowling, D. L., Garcia, M., Dunn, J. C., Ruprecht, R., Stewart, A., Frommolt, K. H., & Fitch, W. T. (2017). Body size and vocalization in primates and carnivores. *Scientific Reports*. https://doi.org/10.1038/srep41070.

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

Brown, S. (2017). A joint prosodic origin of language and music. *Frontiers in Psychology, 8*, 1894. https://doi.org/10.3389/fpsyg.2017.01894.

Bryant, G. A. (2013). Animal signals and emotion in music: Coordinating affect across groups. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00990.

Camperio Ciani, A. (2000). When to get mad: Adaptive significance of rage in animals. *Psychopathology, 33*(4), 191–197. https://doi.org/10.1159/000029142.

Cerchio, S., Jacobsen, J. K., & Norris, T. F. (2001). Temporal and geographical variation in songs of humpback whales, *Megaptera novaeangliae*: Synchronous change in Hawaiian and Mexican breeding assemblages. *Animal Behaviour, 62*(2), 313–329. https://doi.org/10.1006/anbe.2001.1747.

Charlton, B. D. (2014). Menstrual cycle phase alters women's sexual preferences for composers of more complex music. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2014.0403.

Charlton, B. D., Ellis, W. A. H., McKinnon, A. J., Cowin, G. J., Brumm, J., Nilsson, K., & Fitch, W. T. (2011). Cues to body size in the formant spacing of male koala (*Phascolarctos cinereus*) bellows: Honesty in an exaggerated trait. *Journal of Experimental Biology, 214*(20), 3414–3422. https://doi.org/10.1242/jeb.061358.

Charlton, B. D., & Reby, D. (2016). The evolution of acoustic size exaggeration in terrestrial mammals. *Nature Communications*. 1https://doi.org/10.1038/ncomms12739.
Charlton, B. D., Reby, D., & McComb, K. (2007a). Female red deer prefer the roars of larger males. *Biology Letters, 3*(4), 382–385. https://doi.org/10.1098/rsbl.2007.0244.
Charlton, B. D., Reby, D., & McComb, K. (2007b). Female perception of size-related formant shifts in red deer, *Cervus elaphus*. *Animal Behaviour, 74*(4), 707–714. https://doi.org/10.1016/j.anbehav.2006.09.021.
Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*(3), 113–124. https://doi.org/10.1017/S0022226700002024.
Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2*(2), 137–167. https://doi.org/10.1016/S0019-9958(59)90362-6.
Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology. Frontiers*. https://doi.org/10.3389/fpsyg.2015.01182.
Collier, K., Bickel, B., van Schaik, C. P., Manser, M. B., & Townsend, S. W. (2014). Language evolution: Syntax before phonology? *Proceedings of the Royal Society B: Biological Sciences*, *281*(1788), 20140263. https://doi.org/10.1098/rspb.2014.0263.
Congdon, J. V., Hahn, A. H., Filippi, P., Campbell, K. A., Hoang, J., Scully, E. N., et al. (2019). Hear them roar: A comparison of black-capped chickadee (*Poecile atricapillus*) and human (*Homo sapiens*) perception of rousal in Vocalizations across all classes of terrestrial vertebrates. *Journal of Comparative Psychology, 133*(4), 520–541. https://doi.org/10.1037/com0000187.
Cross, N., & Rogers, L. J. (2006). Mobbing vocalizations as a coping response in the common marmoset. *Hormones and Behavior, 49*(2), 237–245. https://doi.org/10.1016/j.yhbeh.2005.07.007.
Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141–201. https://doi.org/10.1177/002383099704000203.
Darwin, C. (1871). *The descent of man and selection in relation to sex*. London: John Murray.
Darwin, C. (1872). *The expression of the emotions in man and animals*. New York: Oxford University Press.
Davis, C., Chong, C. S., & Kim, J. (2017). *The effect of spectral profile on the intelligibility of emotional speech in noise* (pp. 581–585). The MARCS Institute, Western Sydney University, Australia, Interspeech.
Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? *Behavioural Ecology: An Evolutionary Approach, 2*, 282–309.
de Boer, B. (2005). Infant-directed speech and evolution of language. In M. Tallerman (Ed.), *Evolutionary prerequisites for language* (pp. 100–121). Oxford: Oxford University Press.
de Boer, B., Wich, S. A., Hardus, M. E., & Lameira, A. R. (2015). Acoustic models of orangutan hand-assisted alarm calls. *The Journal of Experimental Biology, 218*, 907–914. https://doi.org/10.1242/jeb.110577.
Desrochers, A. U., Be, M., & Bourque, J. (2002). Do mobbing calls affect the perception of predation risk by forest birds? *Animal Behaviour, 64*, 709–714. https://doi.org/10.1006/anbe.2002.4013.
Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science, 298*(5596), 1191–1194. https://doi.org/10.1126/science.1076358.
Dooling, R. J., & Brown, S. D. (1990). Speech perception by budgerigars (*Melopsittacus undulatus*): Spoken vowels. *Perception and Psychophysics, 47*(6), 568–574. https://doi.org/10.3758/BF03203109.
Dunbar, R. I. M. (2003). The social brain: Mind, language, and society in evolutionary perspective. *Annual Review of Anthropology, 32*(1), 163–181. https://doi.org/10.1146/annurev.anthro.32.061002.093158.
Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of emotional speech in younger and older adults. *Ear and Hearing, 35*(6), 695–707. https://doi.org/10.1097/AUD.0000000000000082.
Engesser, S., Ridley, A. R., & Townsend, S. W. (2016). Meaningful call combinations and compositional processing in the southern pied babbler. *Proceedings of the National Academy of Sciences*, *113*(21), 5976–5981. https://doi.org/10.1073/pnas.1600970113.
Engesser, S., & Townsend, S. W. (2019). Combinatoriality in the vocal systems of nonhuman animals. *Wiley Interdisciplinary Reviews: Cognitive Science*. https://doi.org/10.1002/wcs.1493.
Englund, K. T. (2005). Voice onset time in infant directed speech over the first six months. *First Language, 25*(2), 219–234. https://doi.org/10.1177/0142723705050286.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences, 32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X.
Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences, 19*(12), 729–743. https://doi.org/10.1016/j.tics.2015.09.008.

Fallow, P. M., Gardner, J. L., & Magrath, R. D. (2011). Sound familiar? Acoustic similarity provokes responses to unfamiliar heterospecific alarm calls. *Behavioral Ecology, 22*(2), 401–410. https://doi.org/10.1093/beheco/arq221.

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

Faragó, T., Pongrácz, P., Miklósi, Á, Huber, L., Virányi, Z., & Range, F. (2010). Dogs' expectation about signalers' body size by virtue of their growls. *PLoS ONE, 5*(12), e15175. https://doi.org/10.1371/journal.pone.0015175.

Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. The Adapted Mind, 1:391–428. https://doi.org/10.1007/BF00852474.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology, 27*(2), 209–221.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20*(1), 104.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., De Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*(3), 477–501. https://doi.org/10.1017/S0305000900010679.

Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology, 7*, 1393. https://doi.org/10.3389/fpsyg.2016.01393.

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., et al. (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. https://doi.org/10.1098/rspb.2017.0990.

Filippi, P., & Gingras, B. (2018). Emotion communication in animal vocalizations, music and language: An evolutionary perspective. In E. M. Luef & M. M. Marin (Eds.), *The talking species* (pp. 105–125). Graz: Uni-Press Graz Verlag GmbH.

Filippi, P., Gingras, B., & Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.01468.

Filippi, P., Gogoleva, S. S., Volodina, E. V., Volodin, I. A., & Boer, B. D. (2017b). Humans identify negative (but not positive) arousal in silver fox vocalizations: Implications for the adaptive value of interspecific eavesdropping. *Current Zoology, 63*(4), 445–456.

Filippi, P., Hoeschele, M., Spierings, M., & Bowling, D. L. (2019). Temporal modulation in speech, music, and animal vocal communication: Evidence of conserved function. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.14228.

Filippi, P., Laaha, S., & Fitch, W. T. (2017c). Utterance-final position and pitch marking aid word learning in school-age children. *Royal Society Open Science*. https://doi.org/10.1098/rsos.161035.

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & de Boer, B. (2017d). More than words (and faces): Evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion*. https://doi.org/10.1080/02699931.2016.1177489.

Finn, B., & Roediger, H. L. (2011). Enhancing retention through reconsolidation: Negative emotional arousal following retrieval enhances later recall. *Psychological Science, 22*(6), 781–786. https://doi.org/10.1177/0956797611407932.

Fischer, J., & Price, T. (2017). Meaning, intention, and inference in primate vocal communication. *Neuroscience and Biobehavioral Reviews, 82*, 22–31. https://doi.org/10.1016/j.neubiorev.2016.10.014.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America, 102*(2), 1213–1222. https://doi.org/10.1121/1.421048.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(00)01494-7.

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

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic Bulletin and Review, 24*(1), 3–33. https://doi.org/10.3758/s13423-017-1236-5.

Fitch, W. T. (2018a). What animals can teach us about human language: The phonological continuity hypothesis. *Current Opinion in Behavioral Sciences, 21*, 68–75. https://doi.org/10.1016/j.cobeha.2018.01.014.

Fitch, W. T. (2018b). Bio-linguistics: Monkeys break through the syntax barrier. *Current Biology, 28*(12), R695–R697. https://doi.org/10.1016/j.cub.2018.04.087.

Fitch, W. T., De Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *Science Advances*. https://doi.org/10.1126/sciadv.1600723.

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 1933–1955. https://doi.org/10.1098/rstb.2012.0103.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science, 303*(5656), 377–380. https://doi.org/10.1126/science.1089401.

Fitch, W. T., Huber, L., & Bugnyar, T. (2010). Social cognition and the evolution of language: Constructing cognitive phylogenies. *Neuron, 65*(6), 795–814. https://doi.org/10.1016/j.neuron.2010.03.011.

Fitch, W. T., & Kelley, J. P. (2000). Perception of vocal tract resonances by whooping cranes *Grus americana*. *Ethology, 106*(6), 559–574. https://doi.org/10.1046/j.1439-0310.2000.00572.x.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings. Biological Sciences/The Royal Society*, *268*(1477), 1669–75. https://doi.org/10.1098/rspb.2001.1704.

Fitch, W. T., & Zuberbühler, K. (2013). Primate precursors to human language: Beyond discontinuity. In E. Altenmüller, S. Schmidt & E. Zimmermann (Eds.), *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man* (pp. 26–48). Oxford: Oxford University Press.

Forkman, B., Boissy, A., Meunier-Salaün, M.-C., Canali, E., & Jones, R. B. (2007). A critical review of fear tests used on cattle, pigs, sheep, poultry and horses. *Physiology and Behavior, 92*(3), 340–374. https://doi.org/10.1016/J.PHYSBEH.2007.03.016.

Garcia, M., Herbst, C. T., Bowling, D. L., Dunn, J. C., & Fitch, W. T. (2017). Acoustic allometry revisited: Morphological determinants of fundamental frequency in primate vocal production. *Scientific Reports, 7*(1), 10450. https://doi.org/10.1038/s41598-017-11000-x.

Garcia, M., Wondrak, M., Huber, L., & Fitch, W. T. (2016). Honest signaling in domestic piglets (*Sus scrofa domesticus*): Vocal allometry and the information content of grunt calls. *Journal of Experimental Biology, 219*(12), 1913–1921. https://doi.org/10.1242/jeb.138255.

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature, 440*(7088), 1204–1207. https://doi.org/10.1038/nature04675.

Guillet, R., & Arndt, J. (2009). Taboo words: The effect of emotion on memory for peripheral information. *Memory and Cognition, 37*(6), 866–879. https://doi.org/10.3758/MC.37.6.866.

Gussenhoven, C. (2002). Intonation and biology. *Liber Amicorum Bernard Bichakjian (Festschrift for Bernard Bichakjian)*, 59–82.

Gussenhoven, C. (2016). Foundations of intonational meaning: Anatomical and physiological factors. *Topics in Cognitive Science, 8*(2), 425–434. https://doi.org/10.1111/tops.12197.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569.

Heimbauer, L. A., Beran, M. J., & Owren, M. J. (2011). A chimpanzee recognizes synthetic speech with significantly reduced acoustic cues to phonetic content. *Current Biology, 21*(14), 1210–1214. https://doi.org/10.1016/J.CUB.2011.06.007.

Jäger, G., & Rogers, J. (2012). Formal language theory: Refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 1956–1970. https://doi.org/10.1098/rstb.2012.0077.

Jarvis, E. D. (2006). Selection for and against vocal learning in birds and mammals. *Ornithological Science, 5*(1), 5–14. https://doi.org/10.2326/osj.5.5.

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., & Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Current Biology, 28*(12), 1851–1859.e4. https://doi.org/10.1016/j.cub.2018.04.047.

Johnstone, T., & Scherer, K. R. (2000). *Vocal communication of emotion. The Handbook of Emotion* (pp. 220–235). https://doi.org/10.1016/S0167-6393(02)00084-5.

Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Reviews, 26*(2), 235–258.

Jürgens, U. (2009). The neural control of vocalization in mammals: A review. *Journal of Voice, 23*(1), 1–10. https://doi.org/10.1016/j.jvoice.2007.07.005.

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping. *Science, 304*(5677), 1682–1683. https://doi.org/10.1126/science.1097859.

Kitchen, D. M., Bergman, T. J., Cheney, D. L., Nicholson, J. R., & Seyfarth, R. M. (2010). Comparing responses of four ungulate species to playbacks of baboon alarm calls. *Animal Cognition, 13*(6), 861–870. https://doi.org/10.1007/s10071-010-0334-9.

Klieve, S., & Jeanes, R. C. (2001). Perception of prosodic features by children with cochlear implants: Is it sufficient for understanding meaning differences in language? *Deafness and Education International, 3*(1), 15–37. https://doi.org/10.1179/146431501790561061.

Kotz, S. A., & Paulmann, S. (2011). Emotion, language, and the brain. *Linguistics and Language Compass, 5*(3), 108–125. https://doi.org/10.1111/j.1749-818X.2010.00267.x.

Kret, M. E., Jaasma, L., Bionda, T., & Wijnen, J. G. (2016). Bonobos (*Pan paniscus*) show an attentional bias toward conspecifics' emotions. *Proceedings of the National Academy of Sciences*, *113*(14), 3761–3766. https://doi.org/10.1073/pnas.1522060113.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*(4209), 69–72. https://doi.org/10.1126/science.1166301.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America, 73*(3), 1003–1010. https://doi.org/10.1121/1.389148.

Kumar, S., Filipski, A., Swarna, V., Walker, A., & Hedges, S. B. (2005). Placing confidence limits on the molecular age of the human–chimpanzee divergence. *Proceedings of the National Academy of Sciences*, *102*(52), 18842–18847.

Ladefoged, P. (1996). *Elements of acoustic phonetics*. Chicago: The University of Chicago Press. https://doi.org/10.2307/417823.

Lea, A. J., Barrera, J. P., Tom, L. M., & Blumstein, D. T. (2008). Heterospecific eavesdropping in a nonsocial species. *Behavioral Ecology, 19*(5), 1041–1046. https://doi.org/10.1093/beheco/arn064.

Levinson, S. C. (2016). Turn-taking in human communication—Origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010.

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

Lieberman, P. H., Klatt, D. H., & Wilson, W. H. (1969). Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science, 164*(3884), 1185–1187. https://doi.org/10.1126/science.164.3884.1185.

Lingle, S., & Riede, T. (2014). Deer mothers are sensitive to infant distress vocalizations of diverse mammalian species. *The American Naturalist, 184*(4), 510–522.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology, 6*, 1246. https://doi.org/10.3389/fpsyg.2015.01246.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Language Learning and Development, 7*(3), 185–201. https://doi.org/10.1080/15475441.2011.579839.

Macedonia, J. M., & Evans, C. S. (1993). Essay on contemporary issues in ethology: Variation among mammalian alarm call systems and the problem of meaning in animal signals. *Ethology, 93*(3), 177–197. https://doi.org/10.1111/j.1439-0310.1993.tb00988.x.

Magrath, R. D., Pitcher, B. J., & Gardner, J. L. (2009). Recognition of other species' aerial alarm calls: Speaking the same language or learning another? *Proceedings of the Royal Society B: Biological Sciences*, *276*(1657), 769–774. https://doi.org/10.1098/rspb.2008.1368.

Marler, P., Evans, C. S., & Hauser, M. D. (1992). Animal signals: Motivational, referential, or both? In H. Papoušek, U. Jürgens, & M. Papoušek (Eds.), *Studies in emotion and social interaction Nonverbal vocal communication: Comparative and developmental approaches* (pp. 66–86). Cambridge: Cambridge University Press.

Martins, M. D. (2012). Distinctive signatures of recursion. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 2055–2064.

Martins, P. T., & Boeckx, C. (2020). Vocal learning: Beyond the continuum. *PLoS Biology, 18*(3), e3000672.

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27*(1), 1–28. https://doi.org/10.1146/annurev.neuro.27.070203.144157.

Meguerditchian, A., Vauclair, J., & Hopkins, W. D. (2013). On the origins of human handedness and language: A comparative review of hand preferences for bimanual coordinated actions and gestural communication in nonhuman primates. *Developmental Psychobiology, 55*(6), 637–650. https://doi.org/10.1002/dev.21150.

Mendl, M., Burman, O. H. P., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings. Biological sciences/The Royal Society*, *277*(1696), 2895–904. https://doi.org/10.1098/rspb.2010.0303.

Miller, G. (2000). Evolution of human music through sexual selection. In N. L. Wallin, B. Merker & S. Brown (Eds.), *The origins of music* (pp. 329–360). Cambridge: MIT Press. https://doi.org/10.7551/mitpress/5190.003.0025.

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. https://doi.org/10.1086/283219.

Nencheva, M. L., Piazza, E. A., & Lew-Williams, C. (2020). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Developmental Science*, e12997. https://doi.org/10.1111/desc.12997.

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature, 1*(3), 261–289.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology, 49*(2), 85–117. https://doi.org/10.1016/j.cogpsych.2003.12.002.

Nottebohm, F. (2002). The origins of vocal learning. *The American Naturalist, 106*(947), 116–140. https://doi.org/10.1086/282756.

Nowicki, S., & Searcy, W. A. (2014). The evolution of vocal learning. *Current Opinion in Neurobiology, 28*, 48–53. https://doi.org/10.1016/j.conb.2014.06.007.

O'Donnell, T. J., Hauser, M. D., & Fitch, W. T. (2005). Using mathematical models of language experimentally. *Trends in Cognitive Sciences, 9*(6), 284–289. https://doi.org/10.1016/j.tics.2005.04.011.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1–16. https://doi.org/10.1159/000261706.

Ouattara, K., Lemasson, A., & Zuberbühler, K. (2009). Campbell's monkeys use affixation to alter call meaning. *PLoS ONE, 4*(11), e7808. doi:https://doi.org/10.1371/journal.pone.0007808.

Owings, D. H., & Morton, E. S. (1998). *Animal vocal communication: A new approach*. Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/CBO9781139167901.

Owren, M. J., & Rendall, D. (2001). Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology: Issues, News, and Reviews, 10*(2), 58–71. doi:https://doi.org/10.1002/evan.1014.

Panksepp, J. (2009). The emotional antecedents to the evolution of music and language. *Musicae Scientiae, 13*(2_suppl), 229–259. https://doi.org/10.1177/1029864909013002111.

Parker, A. R. (2007). Evolving the narrow language faculty: was recursion the pivotal step? In *The evolution of language: Proceedings of the 6th international conference (EVOLANG06)* (pp. 239–246). https://doi.org/10.1142/9789812774262_0031.

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

Pepperberg, I. M. (2010). Vocal learning in Grey parrots: A brief review of perception, production, and cross-species comparisons. *Brain and Language, 115*(1), 81–91. https://doi.org/10.1016/j.bandl.2009.11.002.

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review, 12*(2), 307–313. https://doi.org/10.3758/BF03196377.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*(2), 175–187. https://doi.org/10.1016/j.neuron.2005.09.025.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition, 95*(2), 201–236.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., et al. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011.

Poole, J. H., Tyack, P. L., Stoeger-Horwath, A. S., & Watwood, S. (2005). Elephants are capable of vocal learning. *Nature, 434*(7032), 455–456. https://doi.org/10.1038/434455a.

Prat, Y., Taub, M., & Yovel, Y. (2015). Vocal learning in a social mammal: Demonstrated by isolation and playback experiments in bats. *Science Advances, 1*(2), e1500019. https://doi.org/10.1126/sciadv.1500019.

Price, T., Wadewitz, P., Cheney, D., Seyfarth, R., Hammerschmidt, K., & Fischer, J. (2015). Vervets revisited: A quantitative analysis of alarm call structure and context specificity. *Scientific Reports, 5*(13220), 1–11. https://doi.org/10.1038/srep13220.

Pulvermller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*, 576–582.

Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283–296. https://doi.org/10.1016/j.evolhumbehav.2005.11.003.

Ralls, K., Fiorelli, P., & Gish, S. (1985). Vocalizations and vocal mimicry in captive harbor seals, Phoca vitulina. *Canadian Journal of Zoology, 63*, 1050–1056. doi:https://doi.org/10.1139/z85-157.

Ravignani, A., Sonnweber, R. S., Stobbe, N., & Fitch, W. T. (2013). Action at a distance: Dependency sensitivity in a New World primate. *Biology Letters, 9*, 20130852. https://doi.org/10.1098/rsbl.2013.0852.

Ravignani, A., Verga, L., & Greenfield, M. D. (2019). Interactive rhythms across species: The evolutionary biology of animal chorusing and turn-taking. *Annals of the New York Academy of Sciences, 1453*(1), 12.

Ravignani, A., Westphal-Fitch, G., Aust, U., Schlumpp, M. M., & Fitch, W. T. (2015). More than one way to see it: Individual heuristics in avian visual computation. *Cognition, 143*, 13–24. https://doi.org/10.1016/j.cognition.2015.05.021.

Reber, S. A., Boeckle, M., Szipl, G., Janisch, J., Bugnyar, T., & Fitch, W. T. (2016). Territorial raven pairs are sensitive to structural changes in simulated acoustic displays of conspecifics. *Animal Behaviour, 116*, 153–162. https://doi.org/10.1016/j.anbehav.2016.04.005.

Reber, S. A., Šlipogor, V., Oh, J., Ravignani, A., Hoeschele, M., Bugnyar, T., & Fitch, W. T. (2019). Common marmosets are sensitive to simple dependencies at variable distances in an artificial grammar. *Evolution and Human Behavior, 40*(2), 214–221. https://doi.org/10.1016/j.evolhumbehav.2018.11.006.

Reichmuth, C., & Casey, C. (2014). Vocal learning in seals, sea lions, and walruses. *Current Opinion in Neurobiology, 28*, 66–71. https://doi.org/10.1016/j.conb.2014.06.011.

Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour, 78*(2), 233–240. https://doi.org/10.1016/j.anbehav.2009.06.007.

Riegel, M., Wierzba, M., Grabowska, A., Jednoróg, K., & Marchewka, A. (2016). Effect of emotion on memory for words and their context. *Journal of Comparative Neurology, 524*(8), 1636–1645. https://doi.org/10.1002/cne.23928.

Russell, A. F., & Townsend, S. W. (2017). Communication: Animal steps on the road to syntax? *Current Biology, 27*(15), R753–R755. https://doi.org/10.1016/j.cub.2017.06.066.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714.

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *NeuroImage, 28*(4), 848–858. https://doi.org/10.1016/j.neuroimage.2005.06.023.

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*(3), 354–356. https://doi.org/10.1177/0956797614560771.

Savage-Rumbaugh, E. S., Murphy, J., Sevcik, R. A., Brakke, K. E., Williams, S. L., & Rumbaugh, D. M. (1993). Language comprehension in ape and child. *Monographs of the Society for Research in Child Development, 58*, 1–221.

Scarantino, A. (2018). Emotional expressions as speech act analogs. *Philosophy of Science, 85*(5), 1038–1053. https://doi.org/10.1086/699667.

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. https://doi.org/10.1177/0022022101032001009.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24–30. https://doi.org/10.1016/j.tics.2005.11.009.

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research, 14*(2), 228–233. https://doi.org/10.1016/S0926-6410(02)00108-8.

Schmelz, M., Call, J., & Tomasello, M. (2011). Chimpanzees know that others make inferences. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 3077–3079. https://doi.org/10.1073/pnas.1000469108.

See, R. L., Driscoll, V. D., Gfeller, K., Kliethermes, S., & Oleson, J. (2013). Speech intonation and melodic contour recognition in children with cochlear implants and with normal hearing. *Otology and Neurotology, 34*(3), 490–498. https://doi.org/10.1097/MAO.0b013e318287c985.

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.

Seymour, B., & Dolan, R. (2008). Emotion, decision making, and the amygdala. *Neuron, 58*(5), 662–671. https://doi.org/10.1016/j.neuron.2008.05.020.

Simonyan, K. (2014). The laryngeal motor cortex: Its organization and connectivity. *Current Opinion in Neurobiology, 28*, 15–21. https://doi.org/10.1016/j.conb.2014.05.006.

Simonyan, K., & Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist, 17*(2), 197–208.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy, 14*(6), 654–666. https://doi.org/10.1080/15250000903263973.

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*(2), 249–267. https://doi.org/10.1016/S0749-596X(03)00024-X.

Sonnweber, R., Ravignani, A., & Fitch, W. T. (2015). Non-adjacent visual dependency learning in chimpanzees. *Animal Cognition, 18*(3), 733–745. https://doi.org/10.1007/s10071-015-0840-x.

Spierings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences, 113*(27), E3977–E3984. https://doi.org/10.1073/pnas.1600483113.

Stansbury, A. L., & Janik, V. M. (2019). Formant modification through vocal production learning in gray seals. *Current Biology, 29*(13), 2244–2249. https://doi.org/10.1016/j.cub.2019.05.071.

Stefanski, R., & Klatt, D. (1974). How does a mynah bird mimic human speech? *The Journal of the Acoustical Society of America, 55*(4), 822–832. https://doi.org/10.1121/1.3437210.

Stevens, N. J., Seiffert, E. R., O'Connor, P. M., Roberts, E. M., Schmitz, M. D., Krause, C., Temu, J. (2013). Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. *Nature, 497*(7451), 611–614.

Stobbe, N., Westphal-Fitch, G., Aust, U., & Tecumseh Fitch, W. (2012). Visual artificial grammar learning: Comparative research on humans, kea (*Nestor notabilis*) and pigeons (*Columba livia*). *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 1995–2006. https://doi.org/10.1098/rstb.2012.0096.

Stoeger, A. S., Mietchen, D., Oh, S., de Silva, S., Herbst, C. T., Kwon, S., & Fitch, W. T. (2012). An asian elephant imitates human speech. *Current Biology, 22*(22), 2144–2148. https://doi.org/10.1016/j.cub.2012.09.022.

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass, 2*(5), 1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x.

Sutherland, M. R., & Mather, M. (2018). Arousal (but not valence) amplifies the impact of salience. *Cognition and Emotion, 32*(3), 616–622. https://doi.org/10.1080/02699931.2017.1330189.

Suzuki, T. N., Wheatcroft, D., & Griesser, M. (2016). Experimental evidence for compositional syntax in bird calls. *Nature Communications, 7*, 1–7. https://doi.org/10.1038/ncomms10986.

Taylor, A., & Reby, D. (2010). The contribution of source-filter theory to mammal vocal communication research. *Journal of Zoology, 280*(3), 221–236. https://doi.org/10.1111/j.1469-7998.2009.00661.x.

Taylor, A. M., Reby, D., & McComb, K. (2008). Human listeners attend to size information in domestic dog growls. *The Journal of the Acoustical Society of America, 123*(5), 2903–2909. https://doi.org/10.1121/1.2896962.

Tchernichovski, O., & Oller, D. K. (2016). Vocal development: How marmoset infants express their feelings. *Current Biology, 26*(10), R422–R424. https://doi.org/10.1016/j.cub.2016.03.063.

Templeton, C. N., Greene, E., & Davis, K. (2005). Allometry of alarm calls: Black-capped chickadees encode information about predator size. *Science*, *308*(5730):1934-1937.

ten Cate, C., & Okanoya, K. (2012). Revisiting the syntactic abilities of nonhuman animals: Natural vocalizations and artificial grammar learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 1984–1994. https://doi.org/10.1098/rstb.2012.0055.

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 of the United States of America, 109*(46), 19027–19032. https://doi.org/10.1073/pnas.1210344109.

Titze, I. R. (1994). *Principles of voice production*. Upper Saddle River: Prentice Hall.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states–the question is which ones and to what extent. *Trends in Cognitive Sciences, 7*(4), 153–156.

Townsend, S. W., Engesser, S., Stoll, S., Zuberbühler, K., & Bickel, B. (2018). Compositionality in animals and humans. PLoS Biology. https://doi.org/10.1371/journal.pbio.2006425.

Trainor, L., & Desjardins, R. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin and Review, 9*(2), 335–340.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science, 11*(3), 188–195. https://doi.org/10.1111/1467-9280.00240.

Van Heijningen, C. A. A., De Visser, J., Zuidema, W., & Cate, T., C (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the United States of America, 106*(48), 20538–20543. https://doi.org/10.1073/pnas.0908113106.

Versace, E., Rogge, J. R., Shelton-May, N., & Ravignani, A. (2019). Positional encoding in cotton-top tamarins (*Saguinus oedipus*). *Animal Cognition, 22*(5), 825–838. https://doi.org/10.1007/s10071-019-01277-y.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences, 13*(6), 258–263. https://doi.org/10.1016/j.tics.2009.03.006.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese q. *Cognition, 103*, 147–162. https://doi.org/10.1016/j.cognition.2006.03.006.

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology, 21*(5), 195–205. https://doi.org/10.1002/evan.21319.

Wiley, R. H. (1983). The evolution of communication: information and manipulation. *Animal Behavior, 2*, 156–189. https://doi.org/10.1016/S0271-5309(97)00009-8.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics: Research article. *Psychological Science, 18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x.

Zhang, Y. S., & Ghazanfar, A. A. (2016). Perinatally influenced autonomic system fluctuations drive infant vocal sequences. *Current Biology, 26*(10), 1249–1260. https://doi.org/10.1016/j.cub.2016.03.023.

Zuberbühler, K. (2020). Syntax and compositionality in animal communication. *Philosophical Transactions of the Royal Society B, 375*(1789), 20190062.