This is a repository copy of Vocal Interactivity in-and-between Humans, Animals, and Robots.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/108727/

Version: Published Version

Article:

https://doi.org/10.3389/frobt.2016.00061

This Document is Protected by copyright and was first published by Frontiers. All rights reserved. It is reproduced with permission.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
Vocal Interactivity in-and-between Humans, Animals, and Robots

Roger K. Moore1*, Ricard Marxer1 and Serge Thill2

1 Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Sheffield, UK; 2 Interaction Lab, School of Informatics, University of Skövde, Skövde, Sweden

Almost all animals exploit vocal signals for a range of ecologically motivated purposes: detecting predators/prey and marking territory, expressing emotions, establishing social relations, and sharing information. Whether it is a bird raising an alarm, a whale calling to potential partners, a dog responding to human commands, a parent reading a story with a child, or a business-person accessing stock prices using Siri, vocalization provides a valuable communication channel through which behavior may be coordinated and controlled, and information may be distributed and acquired. Indeed, the ubiquity of vocal interaction has led to research across an extremely diverse array of fields, from assessing animal welfare, to understanding the precursors of human language, to developing voice-based human–machine interaction. Opportunities for cross-fertilization between these fields abound; for example, using artificial cognitive agents to investigate contemporary theories of language grounding, using machine learning to analyze different habitats or adding vocal expressivity to the next generation of language-enabled autonomous social agents. However, much of the research is conducted within well-defined disciplinary boundaries, and many fundamental issues remain. This paper attempts to redress the balance by presenting a comparative review of vocal interaction within-and-between humans, animals, and artificial agents (such as robots), and it identifies a rich set of open research questions that may benefit from an interdisciplinary analysis.

Keywords: vocal interaction, speech technology, spoken language, human–robot interaction, animal calls, vocal learning, language evolution, vocal expression

1. INTRODUCTION

Almost all living organisms make (and make use of) sounds – even plants (Appel and Cocroft, 2014) – and many animals have specialized biological apparatus that is adapted to the perception and production of sound (Hopp and Evans, 1998). For example, some fish vibrate their swim bladders, many arthropods stridulate,1 and the majority of birds and mammals “vocalize” (using a vocal organ known as a syrinx or a larynx, respectively). Predators may use vocal cues to detect their prey (and vice versa), and a variety of animals (such as birds, frogs, dogs, wolves, foxes, jackals, coyotes, etc.) use vocalization to mark or defend their territory. Social animals (including human beings) also use vocalization to express emotions, to establish social relations, and to share information. Human beings, in particular, have extended this behavior to a very high level of sophistication.

1Stridulation is the act of making sound by rubbing body parts together.
through the evolution of speech and language – a phenomenon that appears to be unique in the animal kingdom, but which shares many characteristics with the communication systems of other animals.

Likewise, auditory perception in many animals is adapted to their acoustic environment and the vocal behavior of other animals, especially conspecifics (Talkington et al., 2012). Vocalization thus sits alongside other modes (such as vision and olfaction) as a primary means by which living beings are able to sense their environment, influence the world around them, coordinate cooperative or competitive behavior with other organisms, and communicate information.

Alongside the study of vocal behavior, recent years have seen important developments in a range of technologies relating to vocalization. For example, systems have been created to analyze and playback animals calls, to investigate how vocal signaling might evolve in communicative agents, and to interact with users of spoken language technology. Indeed, the latter has witnessed huge commercial success in the past 10–20 years, particularly since the release of Naturally Speaking (Dragon's continuous speech dictation software for a PC) in 1997 and Siri (Apple’s voice-operated personal assistant and knowledge navigator for the iPhone) in 2011. Research interest in this area is now beginning to focus on voice-enabling autonomous social agents (such as robots).

Therefore, whether it is a bird raising an alarm, a whale calling to potential partners, a dog responding to human commands, a parent reading a story with a child, or a business-person accessing stock prices using an automated voice service on their mobile phone, vocalization provides a valuable communication channel through which behavior may be coordinated and controlled, and information may be distributed and acquired. Indeed, the ubiquity of vocal interaction has given rise to a wealth of research across an extremely diverse array of fields from the behavioral and language sciences to engineering, technology, and robotics.

Some of these fields, such as human spoken language or vocal interactivity between animals, have a long history of scientific research. Others, such as vocal interaction between artificial agents or between artificial agents and animals, are less well studied – mainly due to the relatively recent appearance of the relevant technology. This means that there is huge potential for cross-fertilization between the different disciplines involved in the study and exploitation of vocal interactivity. For example, it might be possible to use contemporary advances in machine learning to analyze animal activity in different habitats or to use artificial agents to investigate contemporary theories of language grounding. Likewise, an understanding of animal vocal behavior might inform how vocal expressivity might be integrated into the next generation of autonomous social agents.

This paper appraises our current level of understanding about vocal interactivity within-and-between humans, animals, and artificial agents (such as robots). In particular, we present a snapshot of our understanding in six key areas of vocal interaction: animal⇔animal, human⇔human, robot⇔robot, human⇔animal, human⇔robot, and animal⇔robot (see Figure 1) through the consideration of three aspects of vocal interactivity:

1. Vocal signals in interaction. This concerns properties of the signals themselves, including their structure, grammar, and semantic content where applicable. This topic contains a large body of research on both animal and human vocalizations.
2. Vocal interaction between agents. Here, we primarily discuss the functions and different types of vocally interactive behavior between animals, between human beings and animals, and between human beings and technology.
3. Technology-based research methodologies. Lastly, this paper reviews the use of technology in studying vocal interactivity. These are of interest since they provide relatively recent and novel means to further our understanding in the field while also contributing to the development of new technology capable of vocal interaction.

Given the vastness of the topics covered, we aim for snapshots that provide a good sense of the current state-of-the-art and allow us to identify some of the most pertinent open research questions that might benefit from a cross-disciplinary approach. In particular, when reviewing research on specific aspects of human and/or animal vocal interactivity, we also highlight questions pertaining to the design of future vocally interactive technologies that raise.

---

Figure 1: Illustration of the key types of vocal interactivity linking humans, animals, and artificial agents (such as robots).

---

1 A conspecific is a member of the same species.

2 Spoken language technology (SLT) includes voice-based human–computer interaction using components, such as automatic speech recognition, text-to-speech synthesis, and dialogue management.
2. VOCAL SIGNALS IN INTERACTION

2.1. Physiology and Morphology
A range of different neural and physical mechanisms are involved in the production, perception, and interpretation of vocal behavior in humans, animals, and artificial agents (Doupe and Kuhl, 1999; Jarvis, 2004; Ackermann et al., 2014; Andics et al., 2014). The physical apparatus for articulation and audition differs from species to species, as does the neural substrate for processing incoming signals and generating outgoing signals. In some species, it has also been hypothesized that exploiting the vocal production system in a form of analysis-by-synthesis may facilitate the understanding of vocal input (Arbib, 2005).

Human beings are mammals and, as such, the physical mechanisms for producing and perceiving vocalizations are constructed along the same lines as those possessed by all other land mammals. Air flow from the lungs excites resonances in the oral cavity (the “vocal tract”) by vibrating the vocal cords to produce a rich harmonic sound structure, by creating partial closures and using the resulting turbulence to generate noisy fricative sounds, or by closing the vocal tract completely and producing explosive sounds on releasing the air pressure. The spectral characteristics of the generated sounds are modified by the shape of the vocal tract and thus continually influenced by the movement and position of the main articulators – the tongue, the lips, and the jaw. As in other animals, body size influences the characteristics of the vocalizations that human beings are capable of producing. Hence, the pitch of the voice and the “formants” (the vocal tract resonances) are considerably higher in a small child than they are in an adult. In a recent review, Pisanski et al. (2016) suggest that the control of vocal aspects, such as height of formants and pitch, to convey body size information, could be an evolutionary step toward our ability of producing speech.

One difference between the human vocal tract and those of all other mammals is that it is bent into an “L” shape, primarily as a result of our upright vertical posture. This configuration gives rise to the so-called “descended larynx” in adult humans, and it has been hypothesized that this allows human beings to produce a much richer variety of sounds than other mammals (for example, a dog or a monkey) (Lieberman, 1984). This traditional view has been challenged (Fitch and Reby, 2001).

In general terms, however, much regarding the similarities/differences between the vocal systems (including brain organization) in different animals remain unknown and open to further research. Similarly, while morphology has an obvious influence on vocalization as just discussed, the precise nature of this influence and how vocal mechanisms are constrained (or indeed facilitated) by the morphology of the individual agents involved is a topic deserving further study.

2.2. Properties and Function of Animal Signals
Several works have been dedicated to studying how non-human animals adapt their vocalizations to the acoustic context and to the listeners’ perception. Potash (1972) showed how ambient noise modifies the intensity, rate, and type of calls of the Japanese quail. Experiments conducted by Nonaka et al. (1997) demonstrate that the brain stems of cats hold neuronal mechanisms for evoking the Lombard reflex (Lombard, 1911) of increasing speaker effort under the presence of noise. This effect has also been observed in many avian species (Cynx et al., 1998; Manabe et al., 1998; Kobayasi and Okanoya, 2003; Leonard and Horn, 2005) and in frogs (Hafner et al., 2016). Recent work has focused on how other aspects of the vocalizations, such as duration or frequency, are adapted and on the role of auditory feedback in such adaptations (Osmanski and Dooling, 2009; Hage et al., 2013).

Non-human animals have been shown to adapt their vocalizations depending on the audience. For instance, female vervet monkeys produce a higher rate of alarm calls in the presence of their offspring. Likewise, male vervet monkeys make more calls in the presence of adult females than when other dominant males are near (Cheney and Seyfarth, 1985). In some cases, animals may employ vocalizations targeted at individuals of a different species. The kleptoparasitic fork-tailed drongo, when following terrestrially foraging pied babblers, will even perform false alarm calls to make the babblers fly to cover, thereby giving the drongos an opportunity to steal food items (Ridley et al., 2007). Also, vocal communication between species is not confined to animals of the same class, e.g., hornbills (a tropical bird) are known to be capable of distinguishing between different primate alarm calls (Rainey et al., 2004).

Some alarm and mobbing calls serve as an example of the capacity of non-human animals to transmit semantic information referring to specific stimuli categories, to an associated risk, or to a particular amount of danger. Seyfarth et al. (1980) showed how vervet monkeys use and recognize different alarm calls for at least three predators: leopards, eagles, and snakes. Predator or danger-specific calls have been observed in many other species and situations (Blumstein and Armitage, 1997; Greene and Meagher, 1998; Zuberbühler, 2000, 2001; Manser, 2001; Templeton et al., 2005; Grieser, 2009; Yorzinski and Vehrencamp, 2009).

Perhaps, the most interesting recent development in the field of non-human animal vocal interaction is the evidence of syntactic and combinatory rules, grammar, and learning in certain species. For example, McCowan et al. (1999) showed (using bottlenose dolphin whistle repertoires) how an information-theoretic analysis could be used to compare the structural and organizational complexity of various animal communications systems. Ouattara et al. (2009) investigated the ability of non-human primates to form meaningful acoustic variation during call production – a behavior that is functionally equivalent to suffixation in human language when referring to specific external events. A study by Schel et al. (2010) on the alarm call sequences of colobus monkeys concluded that the monkeys attended to the compositional aspects of utterances. Clay and Zuberbühler (2011) showed the ability of bonobos to extract information about external events by attending to vocal sequences of other individuals, thus highlighting the importance of call combinations in their natural communication system. Candioti et al. (2012) describe how some non-human primates vary the acoustic structure of their basic call type and, through combination, create complex structures that increase the effective size of their vocal repertoire.
Zuberbühler (2002) shows that the semantic changes introduced by a combinatorial rule in the natural communication of a particular species of primate may be comprehended by members of another species. Arnold and Zuberbühler (2008) conclude that in the free-ranging putty-nosed monkeys, meaning is encoded by call sequences, not individual calls. Clarke et al. (2006) provide evidence of referential signaling in a free-ranging ape species, based on a communication system that utilizes combinatorial rules. Even though most work is focused on primates, this vocal behavior is also seen in others, e.g., Kershenbaum et al. (2012) provide evidence of complex syntactic vocalizations in a small social mammal: the rock hyrax. More recently, several quantitative approaches have been proposed to understanding the complex structure of bird songs (Sasahara et al., 2012; Weiss et al., 2014) and other non-human vocalization sequences (Kershenbaum et al., 2016).

Ten Cate and Okanoya (2012) review a series of studies and perceptual experiments using artificial grammars that confirm the capacity of non-human animals to generalize and categorize vocal sequences based on phonetic features. Another reviewed set of experiments show ability in non-humans to learn simple rules, such as co-occurrence or duplication of vocal units. However, the capacity of non-human animals to detect abstract rules or rules beyond finite-state grammars remains an open question.

Overall, establishing reliable communications in a challenging environment may therefore require additional effort on the part of the interlocutors, and there is good evidence that animals and human beings alter the characteristics (such as loudness, clarity, or timing) of their vocalizations as a function of the context and perceived communicative success (Brumm and Slater, 2006; Hooper et al., 2006; Candiotti et al., 2012; Hotchkine et al., 2013). Such adaptive behavior may be conditioned on the distance between the interlocutors, the ambient noise level, or the reverberant characteristics of the environment. In general, such behavior is an evidence for “negative feedback control” (Powers, 1974). What objective functions are being optimized? How are vocalizations manipulated to achieve the desired results, and is such behavior reactive or proactive? How should vocally interactive artificial agents be designed in this context?

Further, advanced vocal communication systems, such as language, seem to depend on an intimate connection between low-level sensorimotor processing and high-level cognitive processing. This appears to be necessary in order for contextual knowledge, priors, and predictions to constrain the interpretation of ambiguous and uncertain sensory inputs. “Theory of Mind” (ToM), in particular, is the mechanism by which agents can infer intentions and cognitive states underlying overt behavior. However, the degree to which non-human animals have ToM (Premack and Woodruff, 1978; Bugnyar et al., 2016) or how such insights are supported in different brains (Kirsch et al., 2008) remain unclear. As discussed further below, vocal interactivity is likely often teleological and is thus conditioned on underlying intentions. Does this imply that ToM is crucial for language-based interaction? What level of ToM do animals possess, and could this be used to predict the complexity of their vocal interactivity? Similarly, do artificial agents need ToM in order to interact effectively with human beings vocally?

Finally, although we focus vocal interactivity here, it is nonetheless worth mentioning that there are important issues arising from the relationship between vocal and non-vocal signals in various types of animal (and indeed human) modes of interaction. Indeed, vocal interaction almost always takes place in a multimodal context (Wermer et al., 2009; Liebal et al., 2013; Mavridis, 2014), and this means that vocalization may well be critically coordinated with other physical activities, such as gestures (Esposito and Esposito, 2011; Gillespie-Lynch et al., 2013; Wagner et al., 2014), gaze direction (Holler et al., 2014), and body posture (Morse et al., 2015). How are such multimodal behaviors orchestrated, especially in multi-agent situations? How is information distributed across the different modes, and what is the relationship between vocal and non-vocal (sign) language?

2.3. Properties of Human Language

2.3.1. Structure

The main difference between human and animal vocalization lies not in the physical mechanisms per se, but in how they are used. As Miyagawa et al. (2014) have pointed out, human beings still employ their vocal apparatus as an animal-like call system (primarily to communicate affect). However, humans have also evolved a remarkable system for very high-rate information transfer that appears to be vastly superior to that enjoyed by any other animals – language. Indeed, based on the work of Dawkins (1991) and Gopnik et al. (2001), it can be reasonably claimed that “Spoken language is the most sophisticated behaviour of the most complex organism in the known universe” (Moore, 2007b).

The “special” nature of human spoken language has been much discussed, and it has been hypothesized that it is distinguished from all other forms of animal communication systems through its use of “recursion,” especially in syntactic structure (Hauser et al., 2002). Although compelling, such a distinction was immediately questioned by Pinker and Jackendoff (2005). What is clear is that human language appears to be based on a “particular” (as opposed to “blending”) mechanism for combining elements in a hierarchical structure that exploits the combinatorial properties of compound systems (Abler, 1989). This means that the expressive power of human language is effectively unlimited – as von Humboldt (1836) famously said “Language makes infinite use of finite media.” Likewise, human spoken language appears to be organized as a “contrastive” communication system, which aims to minimize communicative effort (i.e., employs minimal sound distinctions) while at the same time preserving communicative effectiveness, thereby giving rise to language-dependent “phonemic” structure.

The traditional representational hierarchy for spoken language spans acoustics, phonetics, phonology, syntax, semantics, and pragmatics. There is insufficient space here to discuss all of these levels in detail. Suffice to say that each area has been the subject of intensive study for several hundred years, and these investigations have given rise to many schools of thought regarding the structure and function of the underlying mechanisms. Of particular interest are the perspectives provided by research into the phylogenetic roots and ontogenetic constraints that condition
spoken language in both the species and the individual (Stark, 1980; MacNeilage, 1998; Aitchison, 2000; Fitch, 2000, 2010).

2.3.2. Human Language Evolution and Development

The contemporary view is that language is based on the coevolution of two key traits – ostensive-inferential communication and recursive mind-reading (Scott-Phillips, 2015) – and that meaning is grounded in sensorimotor experience. For relatively concrete concepts, this is substantiated by a number of studies that show activations in sensorimotor areas of the brain during language processing [see, e.g. Chersi et al. (2010), for a discussion]. For more abstract concepts, if (and if so, how) they are grounded in sensorimotor experience is still a matter of debate [e.g., Thill et al. (2014)]. Metaphors have been put forward as one mechanism to achieve such grounding (Lakoff and Johnson, 1980; Feldman, 2008), but others argue that abstract concepts may (possibly in addition) build on linguistic or statistical information that is not directly grounded (Barsalou et al., 2008; Dove, 2011; Thill and Twomey, 2016).

There is also considerable interest in the developmental trajectory exhibited by young children while acquiring language (Gopnik et al., 2001), including long-term studies of word learning (Roy et al., 2015). It is well established that early babbling and vocal imitation serves to link perception and production, and that an adult addressing an infant will adapt their own speech to match that of the child (so-called “infant-directed speech”) (Kuhl, 2000). There is also evidence that children are sensitive to statistical and prosodic regularities allowing them to infer the structure and composition of continuous contextualized input (Saffran et al., 1996; Saffran, 2003; Kuhl, 2004; Smith and Yu, 2008).

Rather more controversial is the claim that children exhibit an acceleration in word learning around the age of 18 months – the so-called “vocabulary spurt” phenomenon (McCarty, 1954; Goldfield and Reznick, 1990; Nazzi and Bertoncini, 2003; Ganger and Brent, 2004). However, using data from almost 1800 children, Moore and ten Bosch (2009) found that the acquisition of a receptive/productive lexicon can be quite adequately modeled as a single mathematical growth function (with an ecologically well grounded and cognitively plausible interpretation) with little evidence for a vocabulary spurt.

2.3.3. Interlocutor Abilities

These perspectives on language not only place strong emphasis on the importance of top-down pragmatic constraints (Levinson, 1983) but they are also founded on an implicit assumption that interlocutors share significant priors. Indeed, evidence suggests that some animals draw on representations of their own abilities [expressed as predictive models (Friston and Kiebel, 2009)] in order to interpret the behaviors of others (Rizzolatti and Craighero, 2004; Wilson and Knoblich, 2005). For human beings, this is thought to be a key enabler for efficient recursive mind-reading (Scott-Phillips, 2015) and hence for language (Pickering and Garrod, 2007; Garrod et al., 2013).

A significant factor in the study of (spoken) language is that its complexity and sophistication tends to be masked by the apparent ease with which it is used. As a result, theories are often dominated by a somewhat naïve perspective involving the coding and decoding of messages passing from one brain (the sender) to another (the receiver). They also place a strong emphasis on “turn-taking,” and hence interaction, in spoken language dialog (Levinson, 2006, 2015). However, some researchers claim that “language” is better viewed as an emergent property of the dynamic coupling between cognitive unities that serves to facilitate distributed sense-making through cooperative (social) behaviors (Maturana and Varela, 1987; Bickhard, 2007; Cowley, 2011; Cummins, 2014; Fusaroli et al., 2014).

It is also important to consider the dependencies that exist between interlocutors and the effect such dependencies have on interactive behaviors. The degree to which a talker takes into account the perceived needs of the listener strongly conditions the resulting vocalizations. For example, it is well established that talkers adjust the volume and clarity of their speech in the presence of noise and interference (Lombard, 1911). This is the reason why there is a lack of so-called “invariance” in the vocal signals. Such adaptive behavior is ubiquitous, and speaker–listener coupling may be readily observed in interactions between adults and children (Fernald, 1985), between native and non-native speakers (Nguyen and Delvaux, 2015), and even between humans and machines (Moore and Morris, 1992). As observed by Lindblom (1990), such dependencies may be explained by the operation of control–feedback processes that maximize communicative effectiveness, while minimizing the energy expended in doing so.

2.3.4. Conveyance of Emotion

The formal study of emotion started with the observational work of Charles Darwin (Darwin, 1872) and has since grown into the field we know today as “Affective Science” [and its technical equivalent – “Affective Computing” (Picard, 1997)]. Emotion is a complex physiological, cognitive, and social phenomenon that is exhibited by both humans and animals. Plutchik (1980) hypothesized that emotions serve an adaptive role in helping organisms deal with key survival issues posed by the environment, and that, despite different forms of expression in different species, there are certain common elements, or prototype patterns, that can be identified. In particular, Plutchik (1980) claimed that there a small number of basic, primary, or prototype emotions – conceptualized in terms of pairs of polar opposites – and that all other emotions are mixed or derivative states. Ekman (1999) subsequently proposed six “basic emotions”: happiness, sadness, fear, anger, surprise, and disgust. More recent research favors a “dimensional” approach based on valence (positive vs. negative), arousal, and dominance [Mehrabian (1996), see also the circumplex model, Russell (1980)].

The expression of emotions can be of communicative value, and a number of theories exist regarding this value (Thill and Lowe, 2012). For example, it has been put forward that expressing emotions facilitates social harmony (Griffiths and Scarantino, 2005), while Camras (2011) suggests that emotion expression may even serve the need of the expressor in the sense that it can manipulate the perceiver to the benefit of the expressor’s needs.

In general, emotion is thought to be just one aspect of the various “affective states” that an animal or human being can exhibit, the others being personality, mood, interpersonal stances, and attitudes – all of which have the potential to influence vocalization.
One of the most important overarching sets of research questions relates to the special (or possibly unique) position of human language in relation to the signaling systems used by other living systems, and how we acquired it as a species (Fitch, 2000; Knight et al., 2000; MacNeille, 2008; Tomasello, 2008; Berwick et al., 2013; Ravignani et al., 2016; Vernes, 2016). Likewise, it is often asked whether the patterning of birdsong is similar to speech or, perhaps, more related to music (Shannon, 2016). As discussed earlier, human spoken language appears to have evolved to be a contrastive participant compositional communication system founded on ostensive-inferential recursive mind-reading. Some of these features are exhibited by non-human animals (Berwick et al., 2011; Arnold and Zuberbühler, 2012; ten Cate, 2014). In particular, Engesser et al. (2015) recently claimed evidence for “phonemic” structure in the song of a particular species of bird [although the value of this result was immediately questioned by Bowling and Fitch (2015)]. However, only humans appear to have evolved a system employing all these aspects, so there is considerable interest in comparative analyses of how communication systems can emerge in both living and artificial systems (Oller, 2004; Lyon et al., 2007; Nolfi and Mirolli, 2010). What, for example, is the relationship (if any) between language and the different signaling systems employed by non-human animals? To what degree is there a phonemic structure to animal communications, and how would one experimentally measure the complexity of vocal interactions (beyond information-theoretic analyses)? Bringing it all together, to what extent can different animals said to possess language and to what degree can human vocal interactivity be said to be signaling?

Similarly, vocal learning (especially imitation and mimicry) is thought to be a key precursor of high-order vocal communication systems, such as language (Jarvis, 2006a,b; Lipkina et al., 2013), and only a subset of species exhibits vocal learning: parrots, songbirds, humming birds, humans, bats, dolphins, whales, sea lions, and elephants (Reiss and McCowan, 1993; Tchernichovski et al., 2001; Poole et al., 2005; Pepperberg, 2010; King and Janik, 2013; Chen et al., 2016). More recently, Watson et al. (2015) have added chimpanzees to the list. However, the degree to which animals are capable of learning complex rules in vocal interaction remains an open question (ten Cate and Okanoya, 2012). What are the common features of vocal learning that these species share, and why is it restricted to only a few species? How does a young animal (such as a human child) solve the correspondence problem between the vocalizations that they hear and the sounds that they can produce? Who should adapt to whom in order to establish an effective channel [see, for example, Bohannon and Marquis (1977), for a study showing that adults adapt their vocal interactivity with children based on comprehension feedback by the children]? How are vocal referents acquired? What, precisely, are the mechanisms underlying vocal learning?

3. VOCAL INTERACTIVITY

3.1. Use of Vocalization

Cooperative, competitive, and communicative behaviors are ubiquitous in the animal kingdom, and vocalization provides a means through which such activities may be coordinated and managed in communities of multiple individuals [see, e.g., work by Fang et al. (2014); King et al. (2014); Volodin et al. (2014); Ma (2015)]. Recent years have also seen an emergence of interest in “social signal processing” (Pentland, 2008; Vinciarelli et al., 2009) and even in the characteristics of speech used during speed-dating (Ranganath et al., 2013). This in itself already raises a number of interesting questions. Does the existence (or absence) of prior relationships between agents impact on subsequent vocal activity? Do the characteristics of vocalizations carry information about the social relationship connecting the interactants (for example, how is group membership or social status signaled vocally)? This goes beyond conspecifics – humans and dogs are able to manage a productive and mutually supportive relationship despite the vocal communication being somewhat one-sided. What is it about the human–dog relationship that makes this one-sidedness sufficient, and conversely, what can biases in communication balancing say about social relationships? Finally, how is vocalization used to sustain long-term social relations?

Non-human animals make multiple uses of vocalizations – from signals warning of the presence of predators to social calls strengthening social bonding between individuals. Alarm vocalizations are characterized by being high pitched to avoid the predator localizing the caller (Greene and Meagher, 1998). Alarm calls have been extensively studied in a wide variety of species (Seyfarth et al., 1980; Cheney and Seyfarth, 1985; Blumstein, 1999; Manser, 2001; Fichtel and van Schaik, 2006; Arnold and Zuberbühler, 2008; Stephan and Zuberbühler, 2008; Schel et al., 2010).

The function of alarm calls is not limited to warning conspecifics. For example, Zuberbühler et al. (1999) observed that high rates of monkey alarm calls had an effect on the predator who gave up his hiding location faster once it was detected. Many other animals employ vocalizations to cooperatively attack or harass a predator; these are known as mobbing calls (Ficken and Popp, 1996; Hurd, 1996; Templeton and Greene, 2007; Clara et al., 2008; Grieser, 2009; Yorzinski and Vehrencamp, 2009).

Another role of vocal communication between animals is to inform individuals during the selection of mating partners. Mating or advertising calls have received much research attention in birds due to their complex vocalizations (McGregor, 1992; Scherer, 2003; Seyfarth and Cheney, 2003; Pongrácz et al., 2006; Solits et al., 2009; Perez et al., 2012). The research challenge, especially in respect of emotionally aware artificial agents, is to identify the degree to which affective states can be interpreted and expressed, and whether they should be treated as superficial or more deeply rooted aspects of behavior. What is the role of vocal affect in coordinating cooperative or competitive behavior? How do affective states influence communicative behavior? Interesting work in this direction includes, for example, the design of sound systems that are capable of conveying internal states of a robot through appropriate modulation of the vocal signals (Schwenk and Arras, 2014).
Searcy and Yasukawa, 1996; Vallet et al., 1998; Gil and Gahr, 2002; Mennill et al., 2003; Paff et al., 2007; Alonso Lopes et al., 2010; Bolund et al., 2012; Hall et al., 2013). However, many other species employ such vocal interaction during sexual selection (Brzoska, 1982; Gridi-Papp et al., 2006; Charlton et al., 2012).

Some species use vocalizations to advertise their territory or to maintain territorial exclusion, and the sounds emitted will usually travel long distances. For example, wolves use howls as means to control wolf pack spacing (Harrington and Mech, 1983). These types of vocalization are also used by frogs to advertise their willingness to defend their territory (Brzoska, 1982). Territorial calls also play an important role in sea lions during the breeding season (Peterson and Bartholomew, 1969; Schusterman, 1977). Sea lions and other pinnipeds are also commonly cited as animals that use vocalization between mothers and their offspring. Mothers employ a “pup-attraction call” that will often elicit a “mother-response call” in the pup (Trillmich, 1981; Hanggi and Schusterman, 1990; Gisiner and Schusterman, 1991; Insley, 2001). Mother–offspring calls are one of many examples of the transmission of identity information through animal vocalizations. This aspect has also been studied in the context of songbirds (Weary and Krebs, 1992; Lind et al., 1996), domestic horses (Proops et al., 2009), dolphins (Kershenbaum et al., 2013), and primates (Candiotti et al., 2013).

Overall, vocal signals are therefore arguably generated on purpose (Tomasello et al., 2005; Townsend et al., 2016) and serve to attract attention (Crockford et al., 2014) as well as to provide information (Schel et al., 2013) and support cooperation (Eskelinen et al., 2016). However, other agents can exploit unintentional vocalizations for their own purposes. Also, in living systems, the ultimate driver of behavior is thought to be a hierarchy of “needs” (with survival as the most basic) (Maslow, 1943). As a result, there is interest in the role of “intrinsic motivations,” especially learning (Moulin-Frier et al., 2013). To what extent are vocal signals teleological, and is it possible to distinguish between intentional and unintentional vocalizations? Can intentional vocal activity be simulated by technological means to explore animal behavior? Does a vocalization carry information about the underlying intention, and how can the latter be inferred from the former? How do motivational factors such as “urgency” impact on vocalization? What motivational framework would be appropriate for a voice-enabled autonomous social agent?

An interesting type of vocal interaction, which often occurs between mating pairs is duetting. This comprises a highly synchronized and temporally precise vocal display involving two individuals. Duets have been observed in several bird species (Grafe et al., 2004; Hall, 2004; Elie et al., 2010; Templeton et al., 2013; Dowling and Webster, 2016). There are a number of different hypotheses concerning the function of such behavior, e.g., territory defense, mate-guarding, and paternity-guarding (Mennill, 2006; Dowling and Webster, 2016). Duets also occur in other species and contexts, such as in the alarm calls of lemurs (Fichtel and van Schaik, 2006) and gibbons (Clarke et al., 2006). More generally, vocalizations are often carefully timed in relation to other events taking place in an environment (including other vocalizations) (Benichov et al., 2016). This may take the form of synchronized ritualistic behavior (such as rhythmic chanting, chorusing, or singing) or asynchronous turn-taking (which can be seen as a form of dialog) (Cummins, 2014; Fusaroli et al., 2014; Ravigani et al., 2014).

Of particular interest is the dynamics of such interactions in both humans and animals (Fitch, 2013; Takahashi et al., 2013; De Looze et al., 2014), especially between conspecifics (Friston and Frith, 2015). Is there a common physiological basis for such rhythmic vocal behavior, and how is vocal synchrony achieved between agents? What are the segmental and suprasegmental prosodic features that facilitate such timing relations? What are the dependencies between vocalizations and other events, and how would one characterize them? Given the crucial nature of synchrony and timing in interactivity between natural agents, to what extent does this importance carry over to human–machine dialog? How would one model the relevant dynamics (whether to study natural interactivity or to facilitate human–machine interaction)?

### 3.2. Vocal Interactivity between Non-Conspecifics

Vocal interaction normally takes place between conspecifics (that is, agents with similar capabilities), but what happens between mismatched entities – between humans and/or animals and/or artificial agents? For example, Joslin (1967) employed both human-simulated howls and playback recordings to study wolf behavior and, surprisingly, discovered that the wolves responded more to the human-simulated howls than to the playbacks. Also Kuhl (1981) conducted listening tests on chinchillas in order to determine their capacity for discriminating speech and to provide support for the existence of a relation between the mammalian auditory system and the evolution of different languages.

More recently, the study of domestic or domesticated animals has become a topic of interest in the field of human–animal vocal interaction. For example, Waiblinger et al. (2006) proposes considering vocal interaction in the assessment of human–animal relationships, especially in the context of farm animals’ welfare. Also, Kaminski et al. (2004) present a case study in which they demonstrate a dog’s capacity to “fast map,” i.e., forming quick and rough semantic hypotheses of a new word after a single presentation. Horowitz and Hecht (2016) investigated owner’s vocalizations in dog–human “play” sessions and found some identifiable characteristics associated with affect.

Research in this area also extends to wild animals. For example, McComb et al. (2014) show how elephants respond differently to playbacks of human speech depending on their gender and age – aspects that can greatly affect the predator risks that humans present to elephants.

Research on vocal interactivity between non-conspecifics is particularly pertinent to the design of vocally interactive artificial agents. For example, Jones et al. (2008) found differences in individual preferences when people interacted with dog-like robots. According to Moore (2015, 2016b), understanding this situation could be critical to the success of future speech-based interaction with “intelligent” artificial agents. For example, different bodies may lead to different sensorimotor experiences in which an agent’s concepts are grounded, which may impact the
degree to which two agents can communicate about the same things (Thill et al., 2014).

What, therefore, are the limitations (if any) of vocal interaction between non-conspecifics? What can be learned from attempts to teach animals, the human language (and vice versa)? How do conspecifics accommodate mismatches in temporal histories (for example, interaction between different aged agents) or cultural experience? How can insights from such questions inform the design of vocally interactive artificial agents beyond Siri? Is it possible to detect differences in how different agents ground concepts from their language use, and can artificial agents use such information in vocal interactivity with humans [as suggested by Thill et al. (2014)]?

3.3. Spoken Language Systems

On the technology front, recent years have seen significant advances in technologies that are capable of engaging in voice-based interaction with a human user. The performance of automatic speech recognition, text-to-speech synthesis, and dialog management has improved year-on-year, and this has led to a growth in the sophistication of the applications that are able to be supported, from the earliest military Command and Control Systems to contemporary commercial Interactive Voice Response (IVR) Systems and the latest Voice-Enabled Personal Assistants (such as Siri) – see Figure 2. Progress has been driven by the emergence of a data-driven probabilistic modeling paradigm in the 1980s (Gales and Young, 2007; Bellegarda and Monz, 2015) – recently supplemented by deep learning (Hinton et al., 2012) – coupled with an ongoing regime of government-sponsored benchmarking.4 Pieraccini (2012) presents a comprehensive review of the history of spoken language technology up to the release of Siri in 2011.

At the present time, research into spoken language technology is beginning to focus on the development of voice-based interaction with Embodied Conversational Agents (ECAs) and Autonomous Social Agents (such as robots). In these futuristic scenarios, it is envisioned that spoken language will provide a “natural” conversational interface between human beings and the so-called intelligent systems. However, many challenges need to be addressed in order to meet such a requirement (Baker et al., 2009a; Moore, 2013, 2015), not least how to evolve the complexity of voice-based interfaces from simple structured dialogs to more flexible conversational designs without confusing the user (Bernsen et al., 1998; McTear, 2004; Lopez Cozar Delgado and Araki, 2005; Phillips and Philips, 2006; Moore, 2016b). In particular, seminal work by Nass and Brave (2005) showed how attention needs to be paid to users’ expectations [e.g., selecting the “gender” of a system’s voice (Crowell et al., 2009)], and this has inspired work on “empathic” vocal robots (Breazeal, 2003; Fellous and Arbib, 2005; Haring et al., 2011; Eyssel et al., 2012; Lim and Okuno, 2014; Crumpton and Bethel, 2016). On the other hand, user interface experts, such as Valentine (2007), have argued that such agents should be clearly machines rather than emulations of human beings, particularly to avoid the “uncanny valley effect” (Mori, 1970), whereby mismatched perceptual cues can lead to feelings of repulsion (Moore, 2012). For a voice-enabled robot, this underpins the importance of matching the voice and face (Mitchell et al., 2011).

It has also been argued that the architecture of future spoken language systems needs to be more cognitively motivated if it is to engage meaningfully with human users (Moore, 2007a, 2010; Baker et al., 2009b), or that such systems should take inspiration from the way in which children acquire their communicative skills (ten Bosch et al., 2009).

4. TECHNOLOGY-BASED RESEARCH METHODS

The large number of disciplines concerned with vocal interactivity means that there is an equally wide variety of tools, techniques, and methodologies used in the different areas of research that are relatively novel and emergent, resulting in several avenues for further research, both concerning the development of these methodologies themselves and their use in future studies of vocal interactivity. For example, large-scale data collection is the norm in spoken language technology (Pieraccini, 2012), and several international agencies exist for sharing data between laboratories (for example, the Linguistic Data Consortium5 and the European Language Resource Association).6 Are there other opportunities for sharing data or for inserting technology into non-technological areas? Is it necessary to create new standards in order to facilitate more efficient sharing of research resources?

Likewise, technology for simulating vocalizations is already being used in studies of animal behavior, but different disciplines model vocal interactivity using different paradigms depending on whether they are interested in predicting the outcome of field experiments, eliciting (Benichov et al., 2016) and simulating (Webb, 1995) the behavior in the laboratory, or engineering

---

4 A graph illustrating the history of automatic speech recognition evaluations at the US National Information Technology Laboratory (NIST) can be found at http://www.itl.nist.gov/iad/mig/publications/ASRhistory/.

5 https://www.ldc.upenn.edu.

practical solutions (Moore, 2016a). Vocal interaction may be modeled within a variety of frameworks ranging from traditional behaviorist stimulus–response approaches (for example, using stochastic modeling or deep learning and artificial neural networks) to coupled dynamical systems (using mutual feedback control). In the latter case, vocal interaction is seen as an emergent phenomenon arising from a situated and embodied active relationship between cognitive unities, but how can these interactive behaviors be modeled computationally? Are there any mathematical modeling principles that may be applied to all forms of vocal interactivity, and is it possible to derive a common architecture or framework for describing vocal interactivity?

In addition, technological tools offer great potential for studying vocalization in the wild. As Webb (2008) argues, because robots that act in the world, including interacting with other agents, need to solve many of the same problems that natural autonomous agents need to solve, they provide an additional means by which to study natural behaviors of interest. An oft-cited example is that of cricket mating calls: a female will be attracted to the male of her own species who produces the loudest calls. Webb (1995) built a robot capable of reproducing this behavior using a mechanism of phase cancelation and latency comparison. This is noteworthy in that the potentially complex computational problem, of not just locating sounds but also identifying the loudest source and ensuring it is the correct species, can be solved without an explicit representation of any of these factors. This is discussed further by Wilson and Golonka (2013) as an example of embodied cognition: it is the particular morphology of the cricket’s ear channels and interneurons together with particular aspects of the environment (that males of different species will chirp at different frequencies) that solve this problem, foregoing the need for potentially complicated computations.

Robots can also help to elucidate necessary precursors and mechanisms for vocal interaction. For example, computational models have been used to investigate how children are able to solve the “correspondence problem” and map between their own perceptual and vocal experiences to those of the adult speakers with whom they interact (Howard and Messum, 2014; Messum and Howard, 2015). Even physical (robotic) models of a child’s vocal tract have been designed to understand how these early stages of spoken language acquisition might function (Yoshikawa et al., 2003; Ishihara et al., 2009; Miura et al., 2012).

Another prominent example of this line of research is the “symbol grounding problem” (Harnad, 1990), which, in brief, states that amodal symbols manipulated by a formal system, such as a computer program, have no meaning that is intrinsic to the system itself; whatever meaning may exist is instead attributed by an external observer. Some researchers [e.g., Stramandinoli et al. (2012)] argue that robots require such an intrinsic understanding of concepts to achieve natural vocal interaction with humans. Cangelosi (2006), in particular, distinguishes between physical and social symbol grounding: the former concerns the grounding of an individual’s internal representations in sensorimotor experience, while the latter refers to the determination of symbols to be shared between individuals, including their grounded meanings (in other words, social symbol grounding is the creation of a shared vocabulary of grounded symbols).

Both forms of symbol grounding are a problem that natural agents solve to a greater or lesser extent. Both forms have also been investigated in robots. Luc Steels’ language games, for instance, provide a seminal example of robotic investigations into social symbol grounding (Steels, 2001). These games investigated how artificial agents would “generate and self-organise a shared lexicon as well as the perceptually grounded categorisations of the world expressed by this lexicon, all without human intervention or prior specification” (Steels, 2003, p. 310).

Physical symbol grounding, as mentioned, is the problem of grounding an individual’s internal representations in sensorimotor experience. Implementations of these mechanisms are thus not always concerned with cognitive plausibility but rather with implementing a practical solution [see Coradeschi et al. (2013) for a recent review and Thill et al. (2014) for a longer discussion of how the simpler sensorimotor aspects considered in most robotics may affect the degree to which these can comment on human grounding]. Nonetheless, robots have, for example, been used to put forward theories of how abstract concepts can be grounded in a sensorimotor experience (Cangelosi and Riga, 2006). Stramandinoli et al. (2012), in particular, propose a hierarchical structure for concepts; some may be directly grounded in sensorimotor experience, whereas others are indirectly grounded via other concepts.

The previously mentioned review by Coradeschi et al. (2013) also follow Belpaeme and Cowley (2007) in highlighting that social symbol grounding has the necessary mechanisms for the acquisition of language and meaning as a prerequisite. Here, we want to reaffirm the overall implication; to use robots to study vocal interactivity requires the implementation of prerequisite mechanisms. It is, for instance, occasionally argued that a “mirror neuron system” is an evolutionary precursor to language abilities (Arbib, 2005). This opens the discussion to robot (and computational) models of mirror neuron systems, for which we refer to recent reviews (Oztop et al., 2006; Thill et al., 2013). It also follows from at least some theoretical positions on embodiment that the precise body of an agent may play a fundamental role in all matters of cognition, including symbol grounding (Thill and Twomey, 2016). Indeed, Thill et al. (2014) propose that robot implementations need to take this into account explicitly, suggesting that human usage of concepts, as characterized by appropriate analyses of human-produced texts may in fact yield insights into the underlying grounding. A robot, whose own body would ground these concepts in (possibly subtly) different ways, could make use of this information in interaction with human beings. Overall, then, the take-home message is that using robots in vocal interaction requires the researcher to be explicit about all aspects of the necessary model [see Morse et al. (2011), for a similar point].

Once an artificial agent that is capable of vocal interactivity has been created (whether it achieved this as a result of cognitively plausible modeling or not), it is interesting to ask how humans might actually interact with it. Branigan et al. (2011), for example, report on five experiments in which humans interacted either with other humans or (so they were told) a computer. The core behavior of interest was verbal alignment (in which participants, in a dialog, converge on certain linguistic behaviors). Their main
TABLE 1 | Summary of research questions identified in this paper that pertain to vocal signals in interaction, grouped by the sections of the paper in which they are discussed.

<table>
<thead>
<tr>
<th>Physiology and morphology</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the similarities/differences between the vocal systems (including brain organization) in different animals?</td>
<td></td>
</tr>
<tr>
<td>How are vocal mechanisms constrained or facilitated by the morphology of the individual agents involved?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties and function of animal signals</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What objective functions are being optimized in modulating signals to establish reliable communications?</td>
<td></td>
</tr>
<tr>
<td>How are vocalizations manipulated to achieve the desired results, and is such behavior reactive or proactive?</td>
<td></td>
</tr>
<tr>
<td>How should vocally interactive artificial agents be designed in this context?</td>
<td></td>
</tr>
<tr>
<td>Is ToM crucial for language-based interaction?</td>
<td></td>
</tr>
<tr>
<td>What level of ToM do animals possess, and could this be used to predict the complexity of their vocal interactivity?</td>
<td></td>
</tr>
<tr>
<td>Do artificial agents need ToM in order to interact effectively with human beings vocally?</td>
<td></td>
</tr>
<tr>
<td>How are multimodal behaviors orchestrated, especially in multi-agent situations?</td>
<td></td>
</tr>
<tr>
<td>How is information distributed across the different modes, and what is the relationship between vocal and non-vocal (sign) language?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conveyance of emotion</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>To what degree can affective states be interpreted and expressed, and should they be treated as superficial or more deeply rooted aspects of behavior?</td>
<td></td>
</tr>
<tr>
<td>What is the role of vocal affect in coordinating cooperative or competitive behavior?</td>
<td></td>
</tr>
<tr>
<td>How do affective states influence communicative behavior?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparative analysis of human and animal vocalization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the relationship (if any) between language and the different signaling systems employed by non-human animals?</td>
<td></td>
</tr>
<tr>
<td>To what degree is there a phonemic structure to animal communications, and how would one experimentally measure the complexity of vocal interactions (beyond information-theoretic analyses)?</td>
<td></td>
</tr>
<tr>
<td>To what extent can different animals said to possess language and to what degree can human vocal interactivity be said to be signaling?</td>
<td></td>
</tr>
<tr>
<td>What are the common features of vocal learning that species capable of it share, and why is it restricted to only a few species?</td>
<td></td>
</tr>
<tr>
<td>How does a young animal (such as a human child) solve the correspondence problem between the vocalizations that they hear and the sounds that they can produce?</td>
<td></td>
</tr>
<tr>
<td>Who should adapt to whom in order to establish an effective channel?</td>
<td></td>
</tr>
<tr>
<td>How are vocal referents acquired?</td>
<td></td>
</tr>
<tr>
<td>What, precisely, are the mechanisms underlying vocal learning?</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2 | Summary of research questions identified in this paper that pertain to vocal interactivity, grouped by the sections of the paper in which they are discussed.

<table>
<thead>
<tr>
<th>Use of vocalization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the existence (or absence) of prior relationships between agents impact on subsequent vocal activity?</td>
<td></td>
</tr>
<tr>
<td>Do the characteristics of vocalizations carry information about the social relationship connecting the interactants (for example, how is group membership or social status signaled vocally)?</td>
<td></td>
</tr>
<tr>
<td>What is it about the human–dog relationship that makes the one-sidedness of this relation sufficient, and conversely, what can biases in communication balancing say about social relationships?</td>
<td></td>
</tr>
<tr>
<td>How is vocalization used to sustain long-term social relations?</td>
<td></td>
</tr>
<tr>
<td>To what extent are vocal signals teleological, and is it possible to distinguish between intentional and unintentional vocalizations?</td>
<td></td>
</tr>
<tr>
<td>Can intentional vocal activity be simulated by technological means to explore animal behavior?</td>
<td></td>
</tr>
<tr>
<td>Does a vocalization carry information about the underlying intention, and how can the latter be inferred from the former?</td>
<td></td>
</tr>
<tr>
<td>How do motivational factors such as “urgency” impact on vocalization?</td>
<td></td>
</tr>
<tr>
<td>What motivational framework would be appropriate for a voice-enabled autonomous social agent?</td>
<td></td>
</tr>
<tr>
<td>What are the segmental and supra-segmental prosodic features that facilitate precise timing relations in vocal interaction?</td>
<td></td>
</tr>
<tr>
<td>What are the dependencies between vocalizations and other events, and how would one characterize them?</td>
<td></td>
</tr>
<tr>
<td>Given the crucial nature of synchrony and timing in interactivity between natural agents, to what extent does this importance carry over to human–machine dialog?</td>
<td></td>
</tr>
<tr>
<td>How would one model the relevant dynamics (whether to study natural interactivity or to facilitate human–machine interaction)?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vocal interactivity between non-conspecifics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the limitations (if any) of vocal interaction between non-conspecifics?</td>
<td></td>
</tr>
<tr>
<td>What can be learned from attempts to teach animals, the human language (and vice versa)?</td>
<td></td>
</tr>
<tr>
<td>How do conspecifics accommodate mismatches in temporal histories (for example, interaction between different aged agents) or cultural experience?</td>
<td></td>
</tr>
<tr>
<td>How can insights from such questions inform the design of vocally interactive artificial agents beyond Siri?</td>
<td></td>
</tr>
<tr>
<td>Is it possible to detect the differences in how different agents ground concepts from their language use, and can artificial agents use such information in vocal interactivity with humans?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spoken language systems</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>How does one evolve the complexity of voice-based interfaces from simple structured dialogs to more flexible conversational designs without confusing the user?</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 3 | Summary of research questions identified in this paper that pertain to technology-based research methods.

Technology-based research methods

- Are there novel opportunities for sharing data or for inserting technology into non-technical areas?
- Is it necessary to create new standards in order to facilitate more efficient sharing of research resources?
- How can vocal interactivity as an emergent phenomenon be modeled computationally?
- Are there any mathematical modeling principles that may be applied to all forms of vocal interactivity, and is it possible to derive a common architecture or framework for describing vocal interactivity?
- What tools might be needed in the future to study vocalization in the wild?

5. CONCLUSION

This paper satisfies two objectives. First, we have presented an appraisal of the state-of-the-art in research on vocal interactivity in-and-between humans, animals, and artificial agents (such as robots). Second, we have identified a set of open research questions, summarized again in Tables 1–3 for convenience. It is worth highlighting that many of these open research questions require an interdisciplinary approach – be it the use of artificial agents to study particular aspects of human or animal vocalization, the study of animal vocal behavior to better distinguish between signaling and language in human beings, or indeed the study of human and/or animal vocal interactivity (including between humans and animals) with a view to designing the next generation of vocally interactive technologies.

The questions we have raised thus serve a dual purpose. Not only do they highlight opportunities for future research aimed at increasing our understanding of the general principles of vocal interactivity per se but they also have the potential to impact on practical applications and the design of new technological solutions. Consider, to give but one example, how current technology is moving toward an increasing number of artifacts that offer both cognitive capabilities and voice-enabled interfaces. How the vocal interactivity of such artifacts should be designed is not obvious, since it is not clear how users might expect to interact with such interfaces. Would they prefer natural language or a more command–style interface? What are the precise underlying mechanisms needed for the artifact to offer the desired capabilities?

Finally, let us close by emphasizing again that addressing many of the questions we raise fully requires an interdisciplinary approach that cuts across the different fields that study different types of vocal interactivity in different types of agent. We believe that the time is now ripe to tackle these challenges, and we expect interdisciplinary efforts at the intersections of the fields that together make up the study of vocal interactivity (as outlined in Figure 1) to blossom in the coming years.

AUTHOR CONTRIBUTIONS

RKM, RM, and ST contributed more or less equally to the preparation of this manuscript.

FUNDING

This work was supported by the European Commission [grant numbers EU-FP6-507422, EU-FP6-034434, EU-FP7-231868, and EU-FP7-611971] and the UK Engineering and Physical Sciences Research Council [grant number EP/I013512/1].

REFERENCES


Walters et al. (2008) are perceived, inding that a high-pitched “exuberant” voice lead to a more positive perception of the overall interaction than a low-pitched “calm” voice, highlighting the importance of appropriate pitch in improving human–agent interaction. Recently, de Greeff and Belpaeme (2015) have demonstrated the relevance of such factors in robot learning, finding that a robot that uses appropriate social cues tends to learn faster.

Similarly, the properties of the vocal signals themselves have consequences for the overall interaction. For example, Niculescu et al. (2011) investigated the effects of voice pitch on how robots are perceived, finding that a high-pitched “exuberant” voice lead to a more positive perception of the overall interaction than a low-pitched “calm” voice, highlighting the importance of appropriate voice design for the overall quality of a human–robot interaction. Walters et al. (2008), similarly, found that the voice of the robot modulates physical approach behavior of humans to robots, and what distance is perceived as comfortable.

Finally, it is worth highlighting that such communicative systems need not always be inspired by insights from human or animal vocalization; for example, Schwenk and Arras (2014) present a flexible vocal synthesis system for HRI capable of modulating the sounds a robot makes based on both features of the ongoing interaction and internal states of the robot.


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Copyright © 2016 Moore, Marxer and Thill. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.**