The evolution of speech: vision, rhythm, cooperation

Asif A. Ghazanfar¹ and Daniel Y. Takahashi²

¹Princeton Neuroscience Institute, Departments of Psychology and Ecology & Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
²Princeton Neuroscience Institute, Department of Psychology, Princeton University, Princeton, NJ 08544, USA

A full account of human speech evolution must consider its multisensory, rhythmic, and cooperative characteristics. Humans, apes, and monkeys recognize the correspondence between vocalizations and their associated facial postures, and gain behavioral benefits from them. Some monkey vocalizations even have a speech-like acoustic rhythm but lack the concomitant rhythmic facial motion that speech exhibits. We review data showing that rhythmic facial expressions such as lip-smacking may have been linked to vocal output to produce an ancestral form of rhythmic audiovisual speech. Finally, we argue that human vocal cooperation (turn-taking) may have arisen through a combination of volubility and prosociality, and provide comparative evidence from one species to support this hypothesis.

Introduction

Believing, as I do . . ., that the possession of articulate speech is the grand distinctive character of man . . ., I find it very easy to comprehend that some . . . inconspicuous structural differences may have been the primary cause of the immeasurable and practically infinite divergence of the Human form from the simian strip. – Thomas Huxley [1] (p. 63).

The uniqueness of speech to humans is indisputable, but the question of how it came to be in humans and no other animal remains a source of contention. Did speech evolve gradually via communication precursors in the primate lineage or did it arise ‘spontaneously’ through a fortuitous confluence of genetic and/or neuroanatomical changes found only in humans? Some argue that, unlike traits such as opposable thumbs or color vision, where there is clear evidence for a gradual evolution, speech essentially arose suddenly, almost de novo. Even Thomas Huxley, Darwin’s irascible promoter of the theory of evolution by natural selection, found the idea that speech could evolve gradually – with many factors at play – through animal precursors too difficult to swallow. Huxley’s attitude is shared by modern scientists who continue to argue for ‘primary causes’ whereby key changes in one factor were pivotal for our ‘infinite divergence’ from other primates in the realm of communication. There are advocates for human-specific expression of genes (e.g., FOXP2 [2]), changes in anatomy (e.g., laryngeal descent [3]), increases in the size of the neocortex or particular neocortical areas [4,5], use of peculiar neural circuitry (e.g., mirror neurons [6], neocortical connections with brainstem nuclei [7]), or expression of unique behavioral precursors (e.g., gestures [8] and cooperation [9]).

Each of these factors may have played an important role in the evolution of human communication, but certainly none can be considered a Lynch-pin. This is largely because the problem of speech evolution is one about how a whole suite of features integrate to produce uniquely human vocal output patterns and their perception. That is, similarly to language [10–12], speech is a complex adaptation that evolved in a piecemeal fashion. As such, determining the many substrates required for the evolution of human speech is a difficult task, particularly because most traits thought to give rise to it – the vocal production apparatus and the brain – do not fossilize. We are left with one robust method of inquiry: comparing our vocal behaviors and brain organization with those of other extant mammals, and of primates in particular. Humans have long had a fascination with the utterances of other animals and how their vocal signals may or may not relate to our speech [13]. Even the daring adventurer and master linguist, Sir Richard Burton (1821–1890), could not resist investigating whether monkeys communicated using speech-like vocalizations [14]. Our interest in monkey and other animal vocalizations, and their putative relation to human speech, continues unabated because it is our only path to understanding how human vocal communication evolved.

We will explore three complex phenotypes that are part and parcel of human speech and universal across all languages, but that are typically ignored when considering speech origins: its audiovisual nature, its rhythmicity, and its coordination during conversations. In brief, here are the motivations: (i) speech is produced by making different facial expressions which change the shape of the vocal tract. Not surprisingly, humans recognize the correspondence between vocalizations and the facial postures associated with them. Because speech is thus inherently ‘multisensory’, it is important to investigate the role of facial expressions in the vocalizations of other primates.

(ii) One key characteristic of audiovisual speech is that the acoustic output and associated movements of the mouth are both rhythmic and tightly coordinated. Some monkey
vocalizations have similar acoustic rhythmicity but lack the concomitant rhythmic facial motion. This raises the question of how we evolved from a presumptive ancestral acoustic-only vocal rhythm to one that is audiovisual. (iii) Finally, speech is a behavior that occurs between individuals and is thus a cooperative endeavor. Humans take turns during a conversation to be heard clearly and to facilitate social interactions. Because of its importance and obvious communicative advantage, how vocal cooperation evolved is of great interest. We explore one possible evolutionary trajectory—a combination of prosociality and volubility—for the origin of vocal turn-taking, and use data from marmoset monkeys to explore this idea.

Before we begin we would like to address two caveats. First, speech and language are two separable phenomena that need not have evolved in parallel [12,15]. Speech is an audiovisual signaling system whereas language is a system for communicating complex concepts, irrespective of modality (e.g., writing and sign language as well as speech). In this review we focus on the evolution of speech. Nevertheless, because speech is the default signal system for language in all human cultures, its evolution may also have implications for linguistic evolution [12], but we do not explore these implications. The second caveat is that, as in any review on the evolutionary origins of a behavior, our arguments below are only as good as the amount of comparative evidence available (i.e., the number of species tested). Thus, we hope that even if what we suggest seems too speculative it will spur more experiments in other species (and potentially falsify our claims).

On the origins of multisensory speech

As with humans, many of the signals that nonhuman primates (hereafter, primates) exchange to mediate social interactions take the forms of facial expressions and vocalizations [16]. Indeed, in anthropoid primates, as social group size grows, the complexity of facial expressions [17] and vocal expressions grows in parallel [18,19]. Although facial and vocal expressions are typically treated separately in most studies, they are in fact often inextricably linked: a vocal expression typically cannot be produced without concomitant movements of the face. When we speak our face moves and deforms around the mouth and other regions [20,21]. These dynamics and deformations lead to a variety of visual motion cues related to the auditory components of speech. In noisy, real-world environments, these visual cues increase speech intelligibility [22,23], increase detection speeds [24], and are hard to ignore—visual cues integrate readily and automatically with auditory speech [25]. In light of this, audiovisual (or ‘multisensory’) speech is really the primary mode of speech perception, and is not a capacity that was simply piggybacked onto auditory speech perception later in the course of our evolution [26].

If audiovisual speech is our default mode of communication, then this should be reflected in its evolution. Many species integrate audiovisual signals during communication, including frogs [27,28] and spiders [29]. Moreover, any vertebrate organism that produces vocalizations will have a simple, concomitant visual motion in the area of the mouth. However, in the primate lineage both the number and diversity of muscles innervating the face [30] and the amount of neural control related to facial movement [31,32] increased over time relative to other mammals. This ultimately allowed the production of a greater diversity of facial and vocal expressions in primates [33], with different patterns of facial motion being uniquely linked to different vocal expressions [34,35]. Vocalizations are the result of coordinated movements of the larynx, larynx (vocal folds), and the vocal tract [36,37]. The vocal tract consists of the column of air that extends from the vocal folds to the mouth and nasal passages. Changing the shape of the vocal tract not only allows different sounds to be produced (by modifying the resonance frequencies of the vocal tract), but also results in the predictable deformation of the face around the mouth and other parts of the face [20,34]. Put another way, different facial expressions can result in different-sounding vocalizations.

Given that vocalizations are physically linked to different facial expressions, it is perhaps not surprising that many primates other than humans recognize the correspondence between the visual and auditory components of vocal signals. Both macaque monkeys (Macaca mulatta) and chimpanzees (Pan troglodytes) recognize auditory—visual correspondences between their vocalizations under various contextual and experiential constraints [38–44]. Although these may ‘matching’ experiments show that monkeys and apes can recognize the correspondence between visual and auditory signals, they do not demonstrate directly whether such recognition leads to a behavioral advantage—one that would lead to the natural selection of multisensory processes. In a recent vocal detection study, macaque monkeys were trained to detect auditory, visual, or audiovisual vocalizations embedded in noise as quickly and accurately as possible [45] (Figure 1A). Monkeys exhibited greater accuracy and faster reaction times to audiovisual vocalizations than to unsensory events (Figure 1B), as also seen in humans (Figure 1C). Under these task conditions, monkeys truly integrated faces and voices; that is, they combined them in such a way that behavioral performance was significantly better than either of the unsensory conditions. This was the first evidence for a behavioral advantage for combining faces and voices in a primate [45].

There are also some very important differences in how humans versus primates produce their utterances [37], and these differences further enhance human multisensory communication above and beyond what monkeys can do. One universal feature of speech—typically lacking in at least macaque monkey vocalizations—is its bi-sensory rhythm. That is, when humans speak both the acoustic output and the movements of the mouth are highly rhythmic and tightly correlated with each other [21]. This enhances perception and the parsing of long-duration vocal signals [46]. How did this bисensory speech rhythm evolve?

On the origins of the speech rhythm

Across all languages studied to date, both mouth motion and the acoustic envelope of speech typically exhibit a 3–8 Hz rhythm that is, for the most part, related to the rate of syllable production [21,47]. This 3–8 Hz rhythm is crucial for speech perception. Disrupting the acoustic
component [48–51] or the visual component arising from facial movements [52] decreases intelligibility. It is thought that the speech rhythm parses the signal into basic units from which information on a finer (faster) temporal scale can be extracted [46]. Given the importance of this rhythm in speech and its underlying neurophysiology [53,54], understanding how speech evolved requires investigating the origins of its bi-sensory rhythmic structure.

Unfortunately, not much is known about the rhythmicity of primate vocalizations. We do know that macaque monkey vocalizations have a similar acoustic rhythmicity to human speech, but without the concomitant and temporally correlated rhythmic facial motion [55]. Modulation spectra analyses of the acoustic rhythmicity of macaque monkey vocalizations reveal that their rhythmicity is strikingly similar to that of the acoustic envelope for speech [55] (Figure 2A). Both signals fall within the 3–8 Hz range (see also [56] for evidence of shared low-frequency components between macaque monkey calls and speech). Figure 2B shows that, unlike human speech (top panel), macaque coo vocalizations (bottom panel) are typically produced with a single ballistic facial motion – a motion that does not correspond to the amplitude modulation of the produced sound beyond its onset and offset. Thus, one key evolutionary question is – how did we evolve from a presumptive ancestral unisensory, acoustic-only vocal rhythm (Figure 3A) to one that is audiovisual, with both mouth movements and acoustics sharing the same rhythmicity (Figure 3C)?

Figure 1: Auditory, visual, and audiovisual vocalization detection. (A) Monkeys were trained to detect auditory (green box), visual (blue box), or audiovisual (red box) vocalizations embedded in noise as quickly and as accurately as possible. An avatar and background noise were continuously presented. In the auditory condition, a coo call was presented. In the visual condition, the mouth of the avatar moved without any corresponding vocalization. In the audiovisual, a coo call with a corresponding mouth movement was presented. Each stimulus was presented with four different signal-to-noise ratios (SNR). (B) Mean reaction times as a function of SNR for the unisensory and multisensory conditions for one monkey. The color code is the same as in (A); x axes denote SNR in dB; y axes depict reaction times (RT) in ms. (C) An analogous experiment with human avatar and speech was carried out in humans. The graph represents the mean reaction times as a function of SNR for the unisensory and multisensory conditions for one individual. Conventions as in (B).
One theory posits that the speech rhythm evolved through the modification of rhythmic facial movements in ancestral primates [57] (Figure 3B). In extant primates such facial movements are extremely common as visual communicative gestures. Lip-smacking, for example, is an affiliative signal commonly observed in many genera of primates including virtually every species of Old World monkey [58–61], chimpanzees [62], and in the few New World monkey species whose facial expressions have been studied (common marmosets, Callithrix jacchus [63], and capuchins, Cebus apella [64]). There are no reports of lip-smacking behavior in prosimian primates [65]. Lip-smacking is characterized by regular cycles of vertical jaw movement, often involving a parting of the lips, but sometimes occurring with closed, puckered lips. Although lip-smacking by both monkeys and chimpanzees is often produced during grooming interactions, macaque monkeys (at least) also exchange lip-smacking bouts during face-to-face interactions [61,66–68]. According to MacNeilage [57], during the course of speech evolution, such non-vocal rhythmic facial expressions were coupled with vocalizations to produce the audiovisual components of babbling-like (i.e., consonant-vowel-like) speech expressions in the human lineage (Figure 3C).

Although direct tests of such an evolutionary hypothesis are usually impossible, in this case one can use the 3–8 Hz rhythmic signature of speech as a foundation to explore its veracity. There are now many lines of evidence that demonstrate that the production of lip-smacking in macaque monkeys is similar to the production of orofacial rhythms during speech. First and foremost, lip-smacking exhibits a speech-like rhythm in the 3–8 Hz frequency range [69]. This rhythmic frequency range is distinct from that of chewing and teeth-grinding (an anxiety-driven expression), although all three rhythmic orofacial motions use the same effectors. Nevertheless, it may be that the 3–8 Hz range is large enough that the correspondence between the speech rhythm and the lip-smacking rhythm is not surprising. However, recent evidence from development, X-ray cineradiography, and perception dismiss the possibility that the similarities between lip-smacking and visual speech rhythm are coincidental.

**Development**

If the underlying mechanisms that produce the rhythm in monkey lip-smacking and human speech are homologous, then their developmental trajectories should be similar [70]. In humans, babbling – the earliest form of rhythmic and voluntary vocal behavior [71–73] – is characterized by the production of canonical syllables that have acoustic characteristics similar to those to adult speech and involves rhythmic sequences of mouth close–open alternation [74–76]. Babbling does not emerge with the same rhythmic structure as adult speech. It starts slower and is more variable. During development, the rhythmic frequency increases from ∼3 Hz to ∼5 Hz [21,47,77,78], and the variability of this rhythm starts out very high [77] and does not become fully adult-like until post-pubescence [72]. Importantly, this developmental trajectory from babbling to speech is distinct from that of another cyclical mouth movement, that of chewing. The frequency of chewing movements in humans is highly stereotyped and slow in frequency, remaining unchanged from early infancy into adulthood [79,80]. Chewing movements are often used as a reference movement in speech production studies because, again, both movements use the same effectors.

The developmental trajectory of macaque monkey lip-smacking parallels speech development [81,82]. Measurements of the rhythmic frequency and variability of lip-smacking across neonates, juveniles, and adults revealed...
that young individuals produce slower and more variable mouth movements and, as they get older, these movements become faster and less variable [82]. Moreover, the developmental trajectory for lip-smacking is distinct from that of chewing. As in humans [79,80], macaque monkey chewing had the same slow frequency and consistent low variability across age groups [82]. Thus, in terms of rhythmicity, the trajectory of lip-smacking development is identical to that of babbling-to-consonant-vowel production in humans. The differences in the developmental trajectories between lip-smacking and chewing are also identical to those reported in humans for speech and chewing [77,83–85].

The coordination of effectors

If human speech and monkey lip-smacking have a shared neural basis, one would expect commonalities in the coordination of the effectors involved. During speech, different sounds are produced through the functional coordination between key vocal tract anatomical structures: the jaw/lips, tongue, and hyoid. The hyoid is a bony structure to which the laryngeal muscles attach. These effectors are more loosely coupled during speech movements than during chewing movements [86–89]. X-ray cineradiography (X-ray movies) used to visualize the internal dynamics of the macaque monkey vocal tract during lip-smacking and chewing revealed that lips, tongue, and hyoid move during lip-smacking (as in speech) and do so with a speech-like 3–8 Hz rhythm. Relative to lip-smacking, movements during chewing were significantly slower for each of these structures. Importantly, the temporal coordination of these structures was distinct for each behavior. Partial directed coherence measures—a analysis that measures to what extent one time series can predict another [90]—revealed that although the hyoid moves continuously during lip-smacking there is no coupling of the hyoid with lips and tongue movements, whereas during chewing more coordination was observed between the three structures. These patterns are consistent with what is observed during human speech and chewing [86,87]: the effectors are more

Figure 3. Hypothetical transition from an ancestral unisensory, acoustic-only vocal rhythm to one that is audiovisual, with both mouth movements and acoustics sharing the same rhythmicity. (A) Schematic of a presumptive ancestral vocalization with a rhythmic auditory component (blue line) and non-rhythmic visual component (red line). (B) Graphical representation of a presumptive ancestral rhythmic facial expression without any vocal component; convention as in (A). (C) Illustration of a speech-like utterance with rhythmic and coupled audiovisual components.
Box 1. Vocal coordination: other forms in other species

Many species of animals exchange vocalizations often taking the form of a single ‘call-and-response’. For example, naked mole-rats [117], squirrel monkeys [118], female Japanese macaques [119], large-billed crows [120], bottlenose dolphins [121], and some anurans [122,123] are all capable of simple call-and-response behaviors. It is not known how many animals engage in extended, structured sequences of vocal interactions. Instances of extended, coordinated vocal exchanges include the chorous behaviors of male anurans and insects in the competitive context of mate attraction [124] and duetting between pair-bonded songbirds (e.g., [125,126]; [127] for review), titi monkeys [128], and gibbons (e.g., [129]; [130] for review). Duetting is usually associated with cooperative defense of territory and perhaps mate-guarding. Unlike vocal turn-taking in marmosets and humans, chorousing and duetting occur within the limited contexts of competitive interactions or pair-bonds, respectively. Marmosets and humans are able to flexibly coordinate extended vocal exchanges with any conspecific, regardless of pair-bonding status or relatedness [101].

One possibility is that ‘call-and-response’ behavior, duetting, and cooperative vocal turn-taking are evolutionarily related to one another [131]. For example, Yoshida and Okanoya [131] argue that the more general call-and-response behavior was derived from duetting behavior. Another possibility is that cooperative vocal turn-taking exhibited by marmoset monkeys and humans is like duetting, which has at its foundation a strong social bond between a mated pair. In the case of marmosets and humans, both of which exhibit stable social bonds with unrelated individuals, prosocial behaviors such as cooperative vocal turn-taking may have been driven by their loosely coupled during lip-smacking than during chewing. Furthermore, the spatial displacement of the lips, tongue, and hyoid is greater during chewing than for lip-smacking [91], again similar to what is observed in human speech versus chewing [87].

Perceptual tuning

In speech, disrupting the auditory or visual component of the 3–8 Hz rhythm significantly reduces intelligibility [48–52]. To test whether macaque monkeys were differentially sensitive to lip-smacking produced with a rhythmic frequency in the species-typical range (mean 4–6 Hz [69,82,91]), a preferential-looking procedure was used [92]. Computer-generated monkey avatars were used to produce stimuli varying in lip-smacking frequency within (6 Hz) and outside (3 and 10 Hz) the species-typical range but with otherwise identical features [45,93]. Although there were at least four alternative outcomes in this experiment, monkeys showed a preference for the 6 Hz lip-smacking over the 3 and 10 Hz. This lends behavioral support for the hypothesis that perceptual processes are similarly tuned to the natural frequencies of communication signals as they are for the speech rhythm in humans.

Bridging the gap

How easy would it be to link vocalizations to a rhythmic facial expression during the course of evolution? Recent work on gelada baboons (Theropithecus gelada) proves to be illuminating. Geladas are a highly-specialized type of baboon. Their social structure and habitat are unique among baboons and other Old World primates, as are some of their vocalizations [18]. One of these unique vocalizations, known as a ‘wobble’, is produced only by males of this species and during close affiliative interactions with females. Wobbles are essentially lip-smacking expressions produced concurrently with vocalization [94]. Moreover, their rhythmicity falls within the range of speech rhythm and lip-smacking by macaque monkeys. Given that gelada baboons are very closely related to yellow baboons (their taxa are separated by 4 million years), who do not produce anything resembling wobble vocalizations, it suggests that linking rhythmic facial expressions such as lip-smacking to vocal output may not be a complex evolutionary process. How geladas achieved this feat at the level of neural circuits is unknown, but finding out could reveal key information about the human transition to rhythmic audiovisual vocal output – and, more generally, to the production of consonants (another evolutionary puzzle [95]) – during the course of our evolution.

In humans, this rhythmic signal perception and production is often nested in another rhythm – the extended exchanges of speech across two individuals during a conversation. The evolution of such vocal cooperation between subjects is, of course, as important as the coupling between the visual and auditory modalities within a subject. Effective and efficient vocal communication is achieved by minimizing signal interference. Taking turns is one mechanism that reduces interference. To be conversation-like, such turn-taking would involve multiple exchanges, not simply a call-and-response (Box 1). Until recently humans were thought to be the only primates to exhibit vocal cooperation in the form of turn-taking.

On the origins of cooperative vocal communication

Cooperation is central to human communication [9,96]. Conversation, a form of vocal cooperation, proceeds smoothly because of turn-taking. Typically, speech exchanges between two individuals occur without any
explicit agreement on how the talk may flow [97]. A smooth speech interaction consists of vocal exchanges with gaps of silence and minimal overlaps. These features are universal, being present in the conversations of traditional indigenous peoples as well as those speaking any of the major world languages [98]. Given its central importance in everyday human social interactions, it is natural to ask how conversational, vocal turn-taking evolved. It has been argued that human cooperative vocal communication is unique and evolved in essentially three steps (put forth most cogently in [9]; see also [6,99] for similar scenarios). First, an ape-like ancestor used manual gestures to point and direct the attention of others. Second, later ancestors with prosocial tendencies used manual gestures in communications to mediate shared intentionality. Finally, and most mysteriously, a transition from primarily gestural to primarily vocal forms of cooperative communication formed, perhaps to more efficiently express shared intentionality. Implicit in this idea is that a large brain is required for these behaviors. Is this the only plausible scenario? Not necessarily. Perhaps vocal turn-taking evolved through a voluble and prosocial ancestor without the prior scaffolding of a manual gestures or big brains [100]. The vocal exchanges of the common marmoset monkey provide evidence for this alternative route [101].

Marmoset monkeys are part of the Callatrichinae subfamily of the Cebidae family of New World primates. Marmosets display little evidence of shared intentionality nor do they produce manual gestures. Similarly to humans, they are cooperatively breeding and voluble. Marmosets and humans are among the very few primate species that form pair bonds and exhibit bi-parental and allo-parental care of infants [102]. These cooperative care behaviors are thought to scaffold prosocial motivational and cognitive processes – such as attentional biases toward monitoring others, the ability to coordinate actions, increased social tolerance, and increased responsiveness to the signals of others [103]. Apart from marmosets and humans, and perhaps to some extent bonobos [104], this suite of prosocial behaviors is not typically seen in other primate species. Importantly, when out of visual contact, marmoset monkeys and other callitrichid primates will participate in vocal exchanges with out-of-sight conspecifics [105–108].

In the laboratory and in the wild, marmosets typically use phee calls, a high-pitched call that can be monosyllabic
or multisyllabic, as their contact call [109]. A phee call contains information about gender, identity, and social group [110,111]. Marmoset vocal exchanges can last as long as 30 minutes [101] and have a temporal structure that is strikingly similar to the turn-taking rules that humans use in informal, polite conversations [98]. First, there are rarely, if ever, overlapping calls (i.e., no interruptions, and thus no interference). Second, there is a consistent silent interval between utterances across two individuals. Importantly, as in human conversations, marmoset vocal turn-taking occurs spontaneously with another conspecific regardless of pair-bonding status or relatedness [101]. Thus, although some other animal species exhibit vocal coordination over an extended time-period (as opposed to a simple call-and-response), these behaviors are typically confined to competitive chouring among males of the species or to duetting between pair-bonded mates (Box 1).

Dynamic system models incorporating coupled oscillator-like mechanisms are thought to account for the temporal structure of conversational turn-taking and other social interactions in humans [112,113] (Figure 4A). Such a mechanism would have two basic features: (i) periodic coupling in the timing of utterances across two interacting individuals (Figure 4A,B), and (ii) entrainment, where if the timing of one individual’s vocal output quickens or slows, the other follows suit (Figure 4C,D). The vocal exchanges of marmoset monkeys share both of these features [101]. Thus, marmoset vocal communication, like human speech communication [114], can be modeled as loosely coupled oscillators. As a mechanistic description of vocal turn-taking, coupled oscillators are advantageous because they are consistent with the functions of brain oscillations underlying speech processing [54] and its evolution [55]. Further, such oscillations do not require any higher-order cognitive capacities to function [101]. In other words, a coupled oscillator can occur without the involvement of a big brain [100], something worth considering given the small encephalization quotient of the marmoset monkey compared to great apes and humans [115].

The split between the New World primate lineage and the Old World primate lineage occurred around 40 million years ago [116] and, because no other Old World monkey or ape has been observed to vocally cooperate with conspecifics outside of pair-bond, it is unlikely that the cooperative vocal behavior exhibited by both humans and marmosets is shared with a common ancestor. Thus, it is an example of convergent evolution. However, we argue that such convergent evolution of turn-taking behavior may occur through similar or identical modulation of a homologous neuronal circuit [100] (Box 1). Such modulation is driven by the two behavioral features shared by both humans and marmosets: prosociality and volubility. This hypothesis is consistent with the available data on cooperative vocal behaviors in other taxa, in which the strength of social bonds correlates with frequency and complexity of vocal interaction (Box 1). Given that marmosets engage in vocal cooperation in a manner similar to what we observe in humans, it suggests that cooperative vocal communication could have evolved in a manner very different than gestural-origins hypotheses predict [6,9,99]. Instead of taking an evolutionary route that requires the elaboration of manual gestures and shared intentionality, cooperative vocal communication could have evolved in a more direct fashion. In this alternative scenario, existing vocal repertoires were used in a cooperative, turn-taking manner when prosocial behaviors in general emerged. They developed in both humans and callitrichid primates when they evolved a cooperative breeding strategy.

Concluding remarks and future directions

The default mode of communication in many primates is multisensory. Humans, apes and monkeys all recognize the correspondence between vocalizations and the facial postures associated with them. One striking dissimilarity between some monkey vocalizations and human speech is that the latter has a unique bi-sensory rhythmic structure in that both the acoustic output and the movements of the mouth are rhythmic and tightly correlated. According to one hypothesis, this bimodal speech rhythm evolved through the rhythmic facial expressions of ancestral primates. Developmental, cineradiographic, electromyographic, and perceptual data from macaque monkeys all support the notion that a rhythmic facial expression common among many primate species – lip-smacking – may have been one such ancestral expression. Further explorations of this hypothesis must include a broader comparative sample, especially investigations of the temporal dynamics of facial and vocal expressions in the great apes. Understanding the neural basis of both lip-smacking and speech production – and their similarities and differences – would also be illuminating (Box 2).

In parallel to the evolution of audiovisual coordination within a subject, the evolution of temporal coordination between subjects would need to take place to achieve speech-like behavior. One pragmatic underlying successful speech communication is the ability to take turns. Until recently no nonhuman primate had been observed to naturally take turns using vocalizations in an extended manner with any conspecific. However, such behavior was
recently documented in the common marmoset. Because the common marmoset is distantly related to humans, we argue that turn-taking arose as an instance of convergent evolution and is part of a suite of prosocial behaviors. Such behaviors in both humans and marmosets may be, at least in part, the outcome of a cooperative breeding strategy. Here again more comparative evidence is needed either to bolster or falsify this claim (Box 2). Importantly, marmoset vocal turn-taking demonstrates that a large brain size and complex cognitive machinery is not needed for vocal cooperation to occur. Consistent with this idea, the structure of marmoset vocal exchanges can be described in terms of coupled oscillator dynamics that are similar to those used to describe human conversations.

Acknowledgments

We thank Diego Cordero, Lauren Kelly, and our two anonymous reviewers for their thoughtful comments on this manuscript. We thank David Logue for information on, and insights into, duetting in songbirds. This work was supported by National Institutes of Health grant R01NS054898 (A.A.G.), a James S. McDonnell Scholar Award (A.A.G.), a Pew Latin American Fellowship (D.Y.T.), and a Brazilian Science without Borders Fellowship (D.Y.T.).

References

40. Jordan, K.E. et al. (2005) Monkeys match the number of voices they hear with the number of faces they see. Curr. Biol. 15, 1034–1038
44. Habershon, H.M. et al. (2013) Rhesus macaques recognize unique multimodal face-voice relations of familiar individuals and not of unfamiliar ones. Brain Behav. Evol. 81, 219–225
558 Lindblom, B. et al. (1996) Phonetic systems and phonological development. In Developmental Neurocognition: Speech and Face...
117 Yoshida, S. et al. (2007) Antiphonal vocalization of a subterranean rodent, the naked mole-rat (Heterocephalus glaber). Ethology 113, 703–710