Published in final edited form as:
Trends Cogn Sci. 2014 October ; 18(10): 543–553. doi:10.1016/j.tics.2014.06.004.

The evolution of speech: vision, rhythm, cooperation

Asif A. Ghazanfar[1,2,3,*] and **Daniel Y. Takahashi**[1,2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton NJ 08544, USA

[2]Department of Psychology, Princeton University, Princeton NJ 08544, USA

[3]Department of Ecology & Evolutionary Biology, Princeton University, Princeton NJ 08544, USA

Abstract

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 the associated facial postures and gain behavioral benefits from them. Some monkey vocalizations even have a speech-like acoustic rhythmicity, yet they lack the concomitant rhythmic facial motion that speech exhibits. We review data showing that facial expressions like lip-smacking may be an ancestral expression that was later linked to vocal output in order to produce 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 strips."
>
> —Thomas Huxley [1](pg 63, italics added).

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

*To whom correspondence should be addressed: asifg@princeton.edu.

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—genes (e.g., FOXP2 [2]), anatomy (e.g., laryngeal descent, [3]), increases in the size of the neocortex or particular neocortical areas [4, 5], the advent of peculiar neural circuitry (e.g., mirror neurons [6]; neocortical connections with brainstem nuclei [7]), or behavior (e.g., gestures [8] and cooperation [9])—were critical to our "infinite divergence" from other primates in the realm of communication.

To be sure, 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 integrates to produce uniquely human vocal output patterns and their perception. That is, like 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 since 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 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), couldn't 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: **(1)** Speech is produced by dynamically changing the shape of the vocal tract by making different facial expressions. Not surprisingly, humans recognize the correspondence between vocalizations and the facial postures associated with them. Since speech is inherently "multisensory", it is important to investigate the role of facial expressions in the vocalizations of other primates. **(2)** 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 without 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. **(3)** Finally, speech is a behavior that occurs between individuals and is thus a cooperative endeavor. Humans take turns during a conversation to be better heard 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, while language is a system for communicating complex concepts, irrespective of modality (e.g., writing, sign language as well as speech). In this review, we are focusing on the evolution of speech. Nevertheless, since speech is the default signal system for language in all human cultures, its evolution may have implications for linguistic evolution as well [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 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 as well [18, 19]. While facial and vocal expressions are typically treated separately in most studies, in fact, they are 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 not a capacity that was simply piggy-backed onto auditory speech perception later in the course of our evolution [26].

If audiovisual speech is the default mode, then this should be reflected in its evolution. Many species integrate audio-visual 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 for the production of a greater diversity of facial and vocal expressions in primates [33], with different patterns of facial motion uniquely linked to different vocal expressions [34, 35]. Vocalizations are the result of coordinated movements of the lungs, 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 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]. To put it 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]. While "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). Under such conditions, monkeys exhibited greater accuracy and faster reaction times to audiovisual vocalizations than to unisensory events (Figure 1B); similar to what was observed 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 unisensory conditions. This was the first evidence for a behavioral advantage for combining faces and voices in a primate.

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 bisensory speech rhythm evolve?

On the origins of the speech rhythm

Across all languages studied to date, both the mouth motion and the acoustic envelope of speech typically exhibits 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 critical to 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 as 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 shared low-frequency components of 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 doesn't 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 the one that is audiovisual, with both mouth movements and acoustics sharing the same rhythmicity (Figure 3C)?

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 a 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. While 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).

While 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 orofacial rhythms produced 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), though all three rhythmic orofacial motions use the same effectors. Yet it still may be that the 3 – 8 Hz range is large enough that the correspondence between the speech rhythm and the lip-smacking rhythm is coincidental. However, recent evidence from development, x-ray cineradiography, and perception dismiss 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 adult speech and involves rhythmic sequences of a mouth close-open alternation [74–76]. Babbling does not emerge with the same rhythmic structure as adult speech. It starts out slower and is more variable. Over development, the rhythmic frequency increases from ~ 3 Hz to ~5 Hz [21, 47, 77, 78], and the variability of this rhythm is 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 both movements use the same effectors.

The developmental trajectory of macaque monkey lip-smacking parallels speech development [81, 82]. It starts out slower and is more variable. Measurements of the rhythmic frequency and variability of lip-smacking across neonates, juveniles and adults revealed that young individuals produce slower, more variable mouth movements and as they get older, these movements become faster and less variable [82]. Moreover, the developmental trajectory for lip-smacking was 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, 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—an 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 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 4–6Hz [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 4 alternative outcomes in this experiement, 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

Just 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 is unique among baboons and other Old World primates as are a few of their vocalizations [18]. One of those 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 the 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 don’t produce anything like wobble vocalizations, it suggests that linking rhythmic facial expressions like 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 critical information about the human transition to rhythmic audiovisual vocal output—and, more generally, to the production of consonants (another evolutionary puzzle; see [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 primate to exhibit vocal cooperation in the form of turn-taking.

Box 1

Vocal coordination: other forms in other species

Many species of animals exchange vocalizations, but these usually take the form of a single “call-and-response” (also known as “antiphonal” calling) as opposed to an extended, structured sequence of vocal interactions. 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. Instances of extended, coordinated vocal exchanges include the chorusing behaviors of male anurans and insects in the competitive context of mate attraction [124] and duetting between pair-bonded songbirds (e.g., [125, 126]; for review, see [127]), titi monkeys [128] and gibbons (e.g., [129]; for review, see [130]). Duetting is usually associated with mate-guarding and/or cooperative defense of territory. Unlike vocal turn-taking in marmosets and humans, chorusing 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 [131].

One possibility is that "call-and-response" behavior, duetting and cooperative vocal turn-taking are evolutionarily related to one another [132]. For example, Yoshida & Okanoya [132] 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 derived from 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 like cooperative vocal turn-taking may have been driven by their cooperative breeding strategy [133]. Thus, cooperative vocal turn-taking may be an extension of "duetting-like" vocal coordination to any conspecific. More comparative data are needed to distinguish the most plausible evolutionary scenarios. Regardless of the initial conditions, cooperative vocal turn-taking in marmosets and humans is the result of convergent evolution, as even call-and-response vocal exchanges are not consistently observed among Old World primates. Convergent evolution of vocal behaviors is not uncommon: both vocal learning [134] and duetting [135] evolved multiple times in birds. The evolution of duetting in birds is related to a decline in migration and the formation of more stable social bonds between mates [135]. The cooperative breeding strategy of marmosets and humans also produce more stable social bonds, but beyond the mated pair.

Importantly, convergent evolution of vocal behaviors does not mean that new mechanisms must be deployed at each instance. For example, coupled oscillatory mechanisms can explain the chorusing behaviors of frogs [136], duetting in birds [125] and vocal turn-taking in marmosets [101] and humans [114]. Of course, it is impossible that the specific neural instantiation (the central pattern generators, their connectivity and modulation) of the coupled oscillator mechanisms is the same across all species. However, it may be the case that convergent of evolution vocal turn-taking in marmosets and humans is the outcome of a homologous neural circuit [100]. This is for two reasons: developmental trajectories are highly constrained across related species [137] and radically different behaviors (e.g., turn-taking versus no turn-taking) can hinge on differential neuromodulation of the same circuit [138].

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, present in the conversations of traditional indigenous peoples to 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 by

[9], but 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 in order to express more efficiently shared intentionality. No primate other than humans is thought to exhibit cooperative vocal communication. Does this mean that communication via turn-taking requires a big brain and complex cognitive mechanisms [100]? Not necessarily. Perhaps vocal turn-taking evolved through a voluble and prosocial ancestor without the prior scaffolding of a manual gestures or big brains. The vocal exchanges of the common marmoset monkey provide evidence for this alternative route [101].

Marmoset monkeys are part of the Callatrichinae subfamily of Cebidae family of New World primates. Marmosets display little evidence of shared intentionality nor do they produce manual gestures. Like humans, they are cooperatively breeding and voluble. Marmosets 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 others' signals [103]. Besides humans, and perhaps to some extent in 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 information [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, while there are other animal species which exhibit vocal coordination over an extended time period (as opposed to a simple call-and-response), these behaviors are confined to competitive chorusing among males of the species or duetting strictly between pair-bonded mates (Box 1).

Dynamical 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: 1) periodic coupling in the timing of utterances across two interacting individuals (Figure 4A–B), and 2) 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 since 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 marmoset monkey's small encephalization quotient compared to great apes and humans [115].

The split between the New World primate lineage and the Old World primate lineage occured around 40 million years ago [116], and since no other Old World monkey or ape has been observed to vocally cooperate with conspecifics outside of a pair-bond, it is unlikely that the cooperative vocal behavior exhibited by both humans and marmosets are 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.

Conclusions

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—their similarities and differences—would also be illuminating.

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. As 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 to either bolster or falsify this claim. 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 oscillators 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 NIH R01NS054898 (AAG), the James S. McDonnell Scholar Award (AAG), a Pew Latin American Fellowship (DYT) and a Brazilian Science without Borders Fellowship (DYT).

References

1. Huxley, TH. Evidences as to man’s place in nature. Williams and Norgate; 1863.
2. Vargha-Khadem F, et al. FOXP2 and the neuroanatomy of speech and language. Nature Reviews Neuroscience. 2005; 6:131–138.
3. Lieberman P, et al. Vocal tract limitations on the vowel repertoires of rhesus monkey and other non-human primates. Science. 1969; 164:1185–1187. [PubMed: 4976883]
4. Deacon, TW. The symbolic species: The coevolution of language and the brain. W.W. Norton & Company; 1997.
5. Dunbar, R. Grooming, gossip, and the evolution of language. Harvard University Press; 1998.
6. Rizzolatti G, Arbib MA. Language within our grasp. Trends in Neurosciences. 1998; 21:188–194. [PubMed: 9610880]
7. Jarvis ED. Learned birdsong and the neurobiology of human language. Ann NY Acad Sci. 2004; 1016:749–777. [PubMed: 15313804]
8. Arbib MA, et al. Primate vocalization, gesture, and the evolution of human language. Current Anthropology. 2008; 49:1053–1076. [PubMed: 19391445]
9. Tomasello, M. Origins of human communication. MIT Press; 2008.
10. Bates, E. Children’s language and communication: The Minnesota Symposium on Child Psychology. 1979. The emergence of symbols: Ontogeny and phylogeny; p. 121-157.
11. Pinker S, Jackendoff R. The faculty of language: what’s special about it? Cognition. 2005; 95:201–236. [PubMed: 15694646]
12. Fitch, WT. The evolution of language. Cambridge University Press; 2010.
13. Radick, G. The simian tongue: the long debate about animal language. University of Chicago Press; 2007.
14. Lowell, MS. A Rage to Live: A Biography of Richard and Isabel Burton. Little, Brown and Company; 1998.
15. Fitch WT. The evolution of speech: a comparative review. Trends Cogn Sci. 2000; 4:258–267. [PubMed: 10859570]
16. Ghazanfar AA, Santos LR. Primate brains in the wild: The sensory bases for social interactions. Nature Reviews Neuroscience. 2004; 5:603–616.

17. Dobson SD. Socioecological correlates of facial mobility in nonhuman anthropoids. American Journal of Physical Anthropology. 2009; 138:413–420. [PubMed: 19235791]
18. Gustison ML, et al. Derived vocalizations of geladas (*Theropithecus gelada*) and the evolution of vocal complexity in primates. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012; 367:1847–1859.
19. McComb K, Semple S. Coevolution of vocal communication and sociality in primates. Biology Letters. 2005; 1:381–385. [PubMed: 17148212]
20. Yehia H, et al. Linking facial animation, head motion and speech acoustics. Journal of Phonetics. 2002; 30:555–568.
21. Chandrasekaran C, et al. The natural statistics of audiovisual speech. PLoS Computational Biology. 2009; 5:e1000436. [PubMed: 19609344]
22. Sumby WH, Pollack I. Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America. 1954; 26:212–215.
23. Ross LA, et al. Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. Cereb Cortex. 2007; 17:1147–1153. [PubMed: 16785256]
24. van Wassenhove V, et al. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:1181–1186. [PubMed: 15647358]
25. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]
26. Rosenblum, LD. Primacy of Multimodal Speech Perception. In: Pisoni, DB.; Remez, RE., editors. The Handbook of Speech Perception. Blackwell publishing; 2005.
27. Narins PM, et al. Bimodal signal requisite for agonistic behavior in a dart-poison frog, Epipedobates femoralis. Proc Natl Acad Sci U S A. 2003; 100:577–580. [PubMed: 12515862]
28. Taylor RC, et al. Multimodal signal variation in space and time: how important is matching a signal with its signaler? J Exp Biol. 2011; 214:815–820. [PubMed: 21307068]
29. Uetz GW, Roberts JA. Multisensory cues and multimodal communication in spiders: insights from video/audio playback studies. Brain, Behavior and Evolution. 2002; 59:222–230.
30. Burrows AM, et al. Facial musculature in the rhesus macaque (Macaca mulatta): evolutionary and functional contexts with comparisons to chimpanzees and humans. Journal of Anatomy. 2009; 215:320–334. [PubMed: 19563473]
31. Sherwood CC. Comparative anatomy of the facial motor nucleus in mammals, with an analysis of neuron numbers in primates. Anatomical Record Part a-Discoveries in Molecular Cellular and Evolutionary Biology. 2005; 287A:1067–1079.
32. Sherwood CC, et al. Cortical orofacial motor representation in old world monkeys, great apes, and humans - II. Stereologic analysis of chemoarchitecture. Brain Behavior And Evolution. 2004; 63:82–106.
33. Andrew RJ. The origin and evolution of the calls and facial expressions of the primates. Behaviour. 1962; 20:1–109.
34. Hauser MD, et al. The role of articulation in the production of rhesus monkey, *Macaca mulatta*, vocalizations. Anim Behav. 1993; 45:423–433.
35. Partan SR. Single and Multichannel facial composition: Facial Expressions and Vocalizations Of Rhesus Macaques(Macaca Mulata). Behaviour. 2002; 139:993–1027.
36. Fitch WT, Hauser MD. Vocal production in nonhuman primates - acoustics, physiology, and functional constraints on honest advertisement. American Journal of Primatology. 1995; 37:191–219.
37. Ghazanfar AA, Rendall D. Evolution of human vocal production. Curr Biol. 2008; 18:R457–R460. [PubMed: 18522811]
38. Ghazanfar AA, Logothetis NK. Facial expressions linked to monkey calls. Nature. 2003; 423:937–938. [PubMed: 12827188]
39. Parr LA. Perceptual biases for multimodal cues in chimpanzee (Pan troglodytes) affect recognition. Animal Cognition. 2004; 7:171–178. [PubMed: 14997361]

40. Jordan KE, et al. Monkeys match the number of voices they hear with the number of faces they see. Current Biology. 2005; 15:1034–1038. [PubMed: 15936274]
41. Ghazanfar AA, et al. Vocal tract resonances as indexical cues in rhesus monkeys. Current Biology. 2007; 17:425–430. [PubMed: 17320389]
42. Sliwa J, et al. Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. Proc Natl Acad Sci U S A. 2011; 108:1735–1740. [PubMed: 21220340]
43. Adachi I, Hampton RR. Rhesus monkeys see who they hear: spontaneous crossmodal memory for familiar conspecifics. PLoS One. 2011; 6:e23345. [PubMed: 21887244]
44. Habbershon HM, et al. Rhesus macaques recognize unique multimodal face-voice relations of familiar individuals and not of unfamiliar ones. Brain, Behavior and Evolution. 2013; 81:219–225.
45. Chandrasekaran C, et al. Monkeys and humans share a common computation for face/voice integration. PLoS Comput Biol. 2011; 7:e1002165. [PubMed: 21998576]
46. Ghitza O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Frontiers in Psychology. 2011; 2:130. [PubMed: 21743809]
47. Greenberg S, et al. Temporal properties of spontaneous speech--a syllable-centric perspective. Journal of Phonetics. 2003; 31:465–485.
48. Saberi K, Perrott DR. Cognitive restoration of reversed speech. Nature. 1999; 398:760–760. [PubMed: 10235257]
49. Smith ZM, et al. Chimaeric sounds reveal dichotomies in auditory perception. Nature. 2002; 416:87–90. [PubMed: 11882898]
50. Elliot TM, Theunissen FE. The modulation transfer function for speech intelligibility. PLoS Computational Biology. 2009; 5:e1000302. [PubMed: 19266016]
51. Ghitza O, Greenberg S. On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. Phonetica. 2009; 66:113–126. [PubMed: 19390234]
52. Vitkovitch M, Barber P. Visible Speech as a Function of Image Quality: Effects of Display Parameters on Lipreading Ability. Applied Cognitive Psychology. 1996; 10:121–140.
53. Ghazanfar, AA.; Poeppel, D. The neurophysiology and evolution of the speech rhythm. In: Gazzaniga, MS., editor. The cognitive neurosciences V. The MIT Press; 2014.
54. Giraud AL, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. Nature neuroscience. 2012; 15:511–517.
55. Ghazanfar AA, Takahashi DY. Facial expressions and the evolution of the speech rhythm. Journal of cognitive neuroscience. 2014 In press.
56. Cohen YE, et al. Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. Journal Of Neurophysiology. 2007; 97:1470–1484. [PubMed: 17135477]
57. MacNeilage PF. The frame/content theory of evolution of speech production. Behavioral and Brain Sciences. 1998; 21:499. [PubMed: 10097020]
58. Preuschoft S. Primate faces and facial expressions. Soc Res. 2000; 67:245–271.
59. Hinde RA, Rowell TE. Communication by posture and facial expressions in the rhesus monkey (Macaca mulatta). Proceedings of the Zoological Society London. 1962; 138:1–21.
60. Redican, WK. Facial expressions in nonhuman primates. In: Rosenblum, LA., editor. Primate behavior: developments in field and laboratory research. Academic Press; 1975. p. 103-194.
61. Van Hooff JARAM. Facial expressions of higher primates. Symposium of the Zoological Society, London. 1962; 8:97–125.
62. Parr LA, et al. Influence of Social Context on the Use of Blended and Graded Facial Displays in Chimpanzees. International Journal of Primatology. 2005; 26:73–103.
63. Kemp C, Kaplan G. Facial expressions in common marmosets (*Callithrix jacchus*) and their use by conspecifics. Animal Cognition. 2013; 16:773–788. [PubMed: 23412667]
64. De Marco A, Visalberghi E. Facial displays in young tufted capuchin monkeys (Cebus apella): Appearance, meaning, context and target. Folia Primatologica. 2007; 78:118–137.
65. Newell TG. Social encounters in two prosimian species. Psychonomic Science. 1971; 24:128–130.

66. Ferrari PF, et al. Reciprical face-to-face communication between rhesus macaque mothers and their newborn infants. Current Biology. 2009; 19:1768–1772. [PubMed: 19818617]
67. Livneh U, et al. Self-monitoring of social facial expressions in the primate amygdala and cingulate cortex. Proc Natl Acad Sci U S A. 2012; 109:18956–18961. [PubMed: 23112157]
68. Shepherd SV, et al. Facial muscle coordination during rhythmic facial expression and ingestive movement. Journal Of Neuroscience. 2012; 32:6105–6116. [PubMed: 22553017]
69. Ghazanfar AA, et al. Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: implications for the evolution of audiovisual speech. European Journal of Neuroscience. 2010; 31:1807–1817. [PubMed: 20584185]
70. Gottlieb, G. Individual development & evolution: the genesis of novel behavior. Oxford University Press; 1992.
71. Locke, JL. The child's path to spoken language. Harvard University Press; 1993.
72. Smith A, Zelaznik HN. Development of functional synergies for speech motor coordination in childhood and adolescence. Developmental Psychobiology. 2004; 45:22–33. [PubMed: 15229873]
73. Preuschoff K, et al. Human Insula Activation Reflects Risk Prediction Errors As Well As Risk. Journal Of Neuroscience. 2008; 28:2745–2752. [PubMed: 18337404]
74. Davis BL, MacNeilage PF. The Articulatory Basis of Babbling. J Speech Hear Res. 1995; 38:1199–1211. [PubMed: 8747814]
75. Lindblom, B., et al. Phonetic systems and phonological development. In: de Boysson-Bardies, B., et al., editors. Developmental Neurocognition: Speech and Face Processing in the First Year of Life. Kluwer Academic Publishers; 1996.
76. Oller, DK. The emergence of the speech capacity. Lawrence Erlbaum; 2000.
77. Dolata JK, et al. Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm. Infant Behav Dev. 2008; 31:422–431. [PubMed: 18289693]
78. Nathani S, et al. Final syllable lengthening (FSL) in infant vocalizations. Journal of Child Language. 2003; 30:3–25. [PubMed: 12718291]
79. Green JR, et al. Development of chewing in children from 12 to 48 months: Longitudinal study of EMG patterns. Journal Of Neurophysiology. 1997; 77:2704–2727. [PubMed: 9163386]
80. Kiliaridis S, et al. Characteristics of masticatory mandibular movements and velocity in growing individuals and young adults. Journal of Dental Research. 1991; 70:1367–1370. [PubMed: 1939831]
81. Locke, JL. Lipsmacking and babbling: Syllables, sociality, and survival. In: Davis, BL.; Zajdo, K., editors. The syllable in speech production. Lawrence Erlbaum Associates; 2008. p. 111-129.
82. Morrill RJ, et al. Monkey lip-smacking develops like the human speech rhythm. Developmental Science. 2012; 15:557–568. [PubMed: 22709404]
83. Moore CA, Ruark JL. Does speech emerge from earlier appearing motor behaviors? J Speech Hear Res. 1996; 39:1034–1047. [PubMed: 8898256]
84. Steeve RW. Babbling and chewing: Jaw kinematics from 8 to 22 months. Journal of Phonetics. 2010; 38:445–458. [PubMed: 20725590]
85. Steeve RW, et al. Babbling, Chewing, and Sucking: Oromandibular Coordination at 9 Months. J Speech Lang Hear R. 2008; 51:1390–1404.
86. Hiiemae KM, Palmer JB. Tongue movements in feeding and speech. Crit Rev Oral Biol M. 2003; 14:413–429. [PubMed: 14656897]
87. Hiiemae KM, et al. Hyoid and tongue surface movements in speaking and eating. Arch Oral Biol. 2002; 47:11–27. [PubMed: 11743928]
88. Ostry DJ, Munhall KG. Control of Jaw Orientation and Position in Mastication and Speech. Journal Of Neurophysiology. 1994; 71:1528–1545. [PubMed: 8035233]
89. Matsuo K, Palmer JB. Kinematic linkage of the tongue, jaw, and hyoid during eating and speech. Arch Oral Biol. 2010; 55:325–331. [PubMed: 20236625]
90. Takahashi DY, et al. Information theoretic interpretation of frequency domain connectivity measures. Biological Cybernetics. 2010; 103:463–469. [PubMed: 21153835]
91. Ghazanfar AA, et al. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Curr Biol. 2012; 22:1176–1182. [PubMed: 22658603]

92. Ghazanfar AA, et al. Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. Proc Natl Acad Sci U S A. 2013; 110:1959–1963. [PubMed: 23319616]
93. Steckenfinger SA, Ghazanfar AA. Monkey visual behavior falls into the uncanny valley. Proc Natl Acad Sci USA. 2009; 106:18362–18466. [PubMed: 19822765]
94. Bergman TJ. Speech-like vocalized lip-smacking in geladas. Curr Biol. 2013; 23:R268–R269. [PubMed: 23578870]
95. Lameira AR, et al. Primate feedstock for the evolution of consonants. Trends Cogn Sci. 2014; 18:60–62. [PubMed: 24238780]
96. Levinson, SC. On the human interactional engine. In: Enfield, NJ.; Levinson, SC., editors. Roots of human sociality. Berg Publishers; 2006. p. 39-69.
97. Sacks H, et al. Simplest Systematics for Organization of Turn-Taking for Conversation. Language. 1974; 50:696–735.
98. Stivers T, et al. Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, USA. 2009; 106:10587–10592.
99. Hewes GW. Primate communication and the gestural origin of language. Current Anthropology. 1973; 14:5–24.
100. Borjon JI, Ghazanfar AA. Convergent evolution of vocal cooperation without convergent evolution of brain size. Brain, Behavior and Evolution. 2014 In press.
101. Takahashi DY, et al. Coupled Oscillator Dynamics of Vocal Turn-Taking in Monkeys. Curr Biol. 2013; 23:2162–2168. [PubMed: 24139740]
102. Zahed SR, et al. Male parenting and response to infant stimuli in the common marmoset (Callithrix jacchus). American Journal of Primatology. 2008; 70:84–92. [PubMed: 17607701]
103. Burkart JM, van Schaik CP. Cognitive consequences of cooperative breeding in primates? Animal Cognition. 2010; 13:1–19. [PubMed: 19629551]
104. Hare B, et al. Tolerance allows bonobos to outperform chimpanzees on a cooperative task. Current Biology. 2007; 17:619–623. [PubMed: 17346970]
105. Chen HC, et al. Contact calls of common marmosets (Callithrix jacchus): influence of age of caller on antiphonal calling and other vocal responses. American journal of primatology. 2009; 71:165–170. [PubMed: 19026011]
106. Ghazanfar AA, et al. The units of perception in the antiphonal calling behavior of cotton-top tamarins (Saguinus oedipus): playback experiments with long calls. J Comp Physiol A. 2001; 187:27–35. [PubMed: 11318375]
107. Ghazanfar AA, et al. Temporal cues in the antiphonal long-calling behaviour of cottontop tamarins. Animal Behaviour. 2002; 64:427–438.
108. Miller CT, Wang X. Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology. 2006; 192:27–38.
109. Bezerra BM, Souto A. Structure and usage of the vocal repertoire of Callithrix jacchus. International Journal of Primatology. 2008; 29:671–701.
110. Miller CT, et al. The communicative content of the common marmoset phee call during antiphonal calling. American journal of primatology. 2010; 72:974–980. [PubMed: 20549761]
111. Norcross JL, Newman JD. Context and gender-specific differences in the acoustic structure of common marmoset (Callithrix jacchus) phee calls. American journal of primatology. 1993; 30:37–54.
112. Oullier O, et al. Social coordination dynamics: Measuring human bonding. Soc Neurosci. 2008; 3:178–192. [PubMed: 18552971]
113. Schmidt R, Morr S. Coordination dynamics of natural social interactions. Int J Sport Psychol. 2010; 41:105–106.
114. O'Dell ML, et al. Modeling Turn-Taking Rhythms with Oscillators. Linguist Ural. 2012; 48:218–227.
115. Jerison, HJ. Evolution of the brain and intelligence. Academic Press; 1973.
116. Steiper ME, Young NM. Primate molecular divergence dates. Molecular phylogenetics and evolution. 2006; 41:384–394. [PubMed: 16815047]

117. Yosida S, et al. Antiphonal Vocalization of a Subterranean Rodent, the Naked Mole-Rat (Heterocephalus glaber). Ethology. 2007; 113:703–710.
118. Masataka, N.; Biben, M. Behaviour. Brill; 1987. Temporal Rules Regulating Affiliative Vocal Exchanges of Squirrel Monkeys; p. 311-319.
119. Sugiura H. Matching of acoustic features during the vocal exchange of coo calls by Japanese macaques. Animal Behaviour. 1998; 55:673–687. [PubMed: 9514664]
120. Kondo N, et al. A Temporal Rule in Vocal Exchange Among Large-Billed Crows Corvus macrorhynchos in Japan. Ornithological Science. 2010; 9:83–91.
121. Nakahara F, Miyazaki N. Vocal exchanges of signature whistles in bottlenose dolphins (Tursiops truncatus). Journal of Ethology. 2011; 29:309–320.
122. Grafe TU. The function of call alternation in the African reed frog (Hyperolius marmoratus): precise call timing prevents auditory masking. Behavioral Ecology and Sociobiology. 1996; 38:149–158.
123. Zelick R, Narins PM. Characterization of the advertisement call oscillator in the frogEleutherodactylus coqui. Journal of Comparative Physiology A. 1985; 156:223–229.
124. Greenfield MD. Synchronous and Alternating Choruses in Insects and Anurans: Common Mechanisms and Diverse Functions. Integrative and Comparative Biology. 1994; 34:605–615.
125. Laje R, Mindlin GB. Highly structured duets in the song of the South American hornero. Physical Review Letters. 2003; 91:258104. [PubMed: 14754163]
126. Logue DM, et al. The behavioural mechanisms underlying temporal coordination in black-bellied wren duets. Animal Behaviour. 2008; 75:1803–1808.
127. Hall ML. A review of vocal duetting in birds. Advances in the Study of Behavior. 2009; 40:67–121.
128. Caselli CB, et al. Vocal behavior of black-fronted titi monkeys (*Callicebus nigrifrons*): acoustic properties and behavioral contexts of loud calls. Am J Primatol. 2014 In press.
129. Mitani JC. Gibbon Song Duets and Intergroup Spacing. Behaviour. 1985; 92:59–96.
130. Geissmann T. Duet-splitting and the evolution of gibbon songs. Biological Reviews. 2002; 77:57–76. [PubMed: 11911374]
131. Takahashi DY, et al. Coupled Oscillator Dynamics of Vocal Turn-Taking in Monkeys. Current biology : CB. 2013; 23:2162–2168. [PubMed: 24139740]
132. Yoshida S, Okanoya K. Evolution of turn-taking: A bio-cognitive perspective. Cognitive Studies. 2005; 12:153–165.
133. Burkart JM, et al. Cooperative breeding and human cognitive evolution. Evolutionary Anthropology. 2009; 18:175–186.
134. Jarvis ED. Selection for and against vocal learning in birds and mammals. Ornithological Science. 2006; 5:5–14.
135. Logue DM, Hall ML. Migration and the evolution of duetting in songbirds. Proceedings of the Royal Society B. 2014 In press.
136. Greenfield MD. Synchronous and Alternating Choruses in Insects and Anurans: Common Mechanisms and Diverse Functions. American Zoologist. 1994; 34:605–615.
137. Krubitzer LA, Seelke AM. Cortical evolution in mammals: the bane and beauty of phenotypic variability. Proc Natl Acad Sci U S A. 2012; 109 (Supplement 1):10647–10654. [PubMed: 22723368]
138. Marder E. Neuromodulation of neuronal circuits: back to the future. Neuron. 2012; 76:1–11. [PubMed: 23040802]

Outstanding questions

Beyond the advantages that facial motion provides for vocal detection in noisy environments, do non-human primate species also use facial motion to discriminate different call types?

What neural mechanisms and/or biomechanical structures link rhythmic facial motion with rhythmic vocal acoustics?

Is cooperative vocal turn-taking evident in species closely related to marmoset monkeys and humans, but without prosocial tendencies and/or cooperative breeding strategies (*e.g.*, squirrel monkeys and chimpanzees)?

What are the neural bases for the coupled oscillator dynamics during vocal turn-taking, and are these mechanisms the same across, for example, marmoset monkeys and humans? Are the neural bases the same or similar to those exhibited by duetting birds?

What changes in neural circuitry (or in its modulation) lead to changes in prosociality and/or cooperative vocal communication? Is this neural mechanism shared across all species that exhibit some form of vocal coordination (e.g., duetting) with conspecifics?

Highlights

- Human speech is multisensory, rhythmic, and cooperative in nature.
- Like humans, monkeys benefit from integrating faces and voices.
- The rhythmic nature of speech likely originated in ancestral rhythmic facial expressions
- Vocal cooperation that may have arisen through a combination of volubility and prosociality.

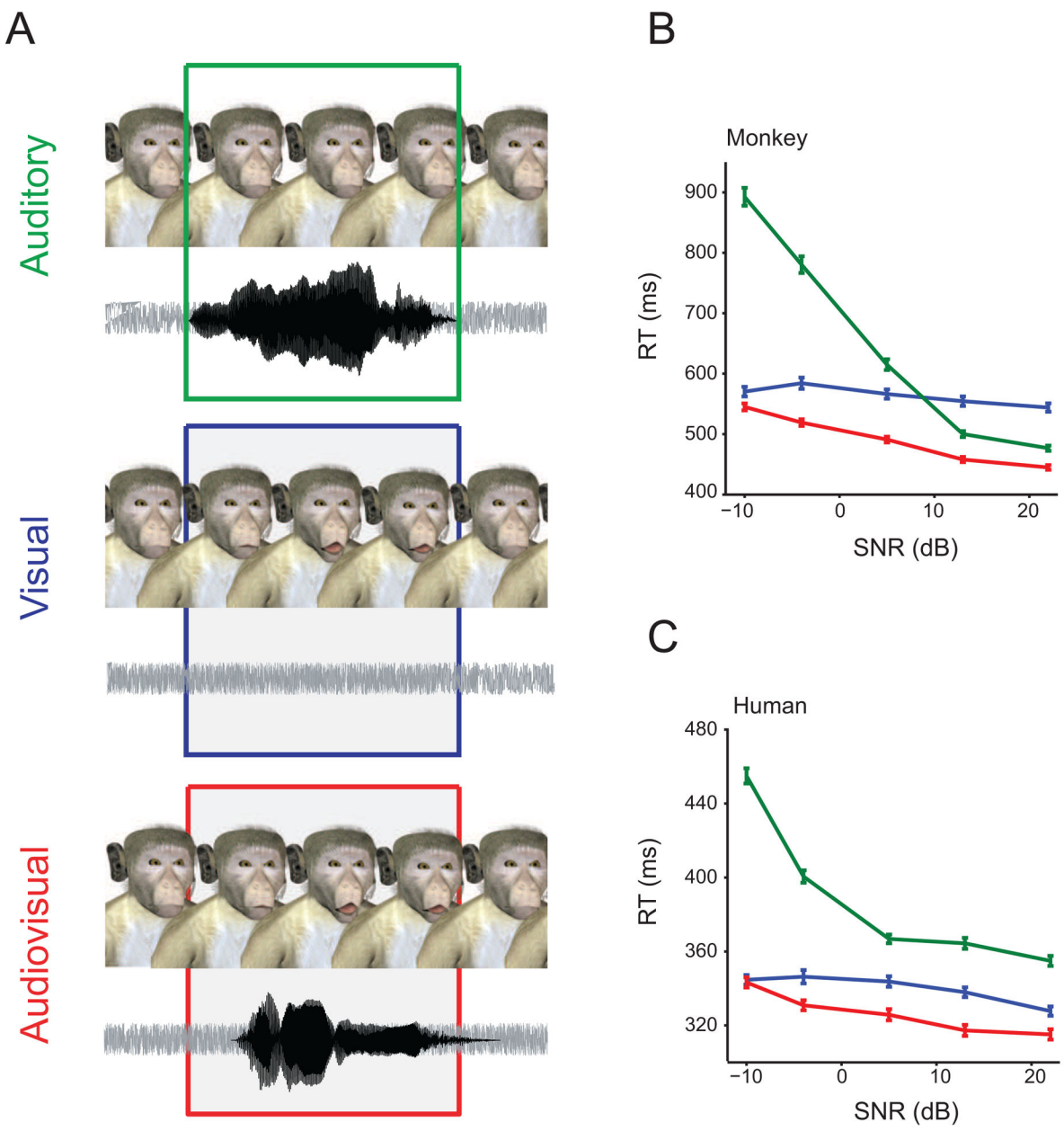

Figure 1.

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 fast and as accurately as possible. An avatar and background noise was 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 RT in milliseconds. **C.**

An analogous experiment with human avatar and speech was done in humans. The graph represents the mean reaction times as a function of SNR for the unisensory and multisensory conditions for one human. Conventions as in **B**.

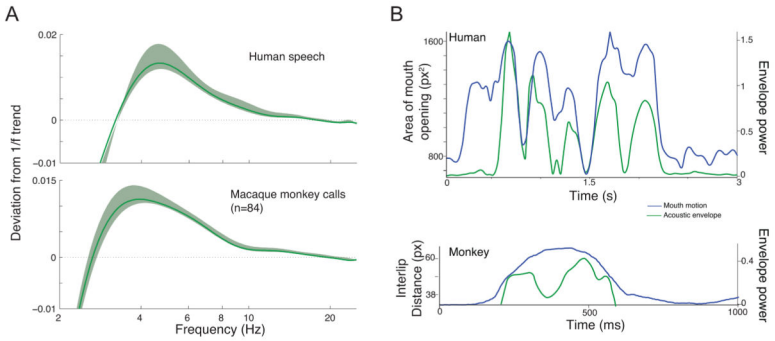

Figure 2.

A. Speech and macaque monkey calls have similar rhythmic structure in their acoustic envelopes. Modulation spectra for human speech and long duration (>400 ms) macaque monkey calls. X-axes represent frequency in log Hz; y-axes depict power deviations from a 1/f trend. **B.** Mouth motion and auditory envelope for a single sentence produced by human (top panel). X–axis depicts time in seconds; y–axis on the left depict the area of the mouth opening in pixel squared; y-axis on the right depict the acoustic envelope in Hilbert units. Bottom panel shows mouth motion and the auditory envelope for a single coo vocalization produced by a macaque monkey. X–axis depicts time in milliseconds; y–axis on the left depict the distance between lips in pixels; y-axis on the right depict the acoustic envelope power in Hilbert units.

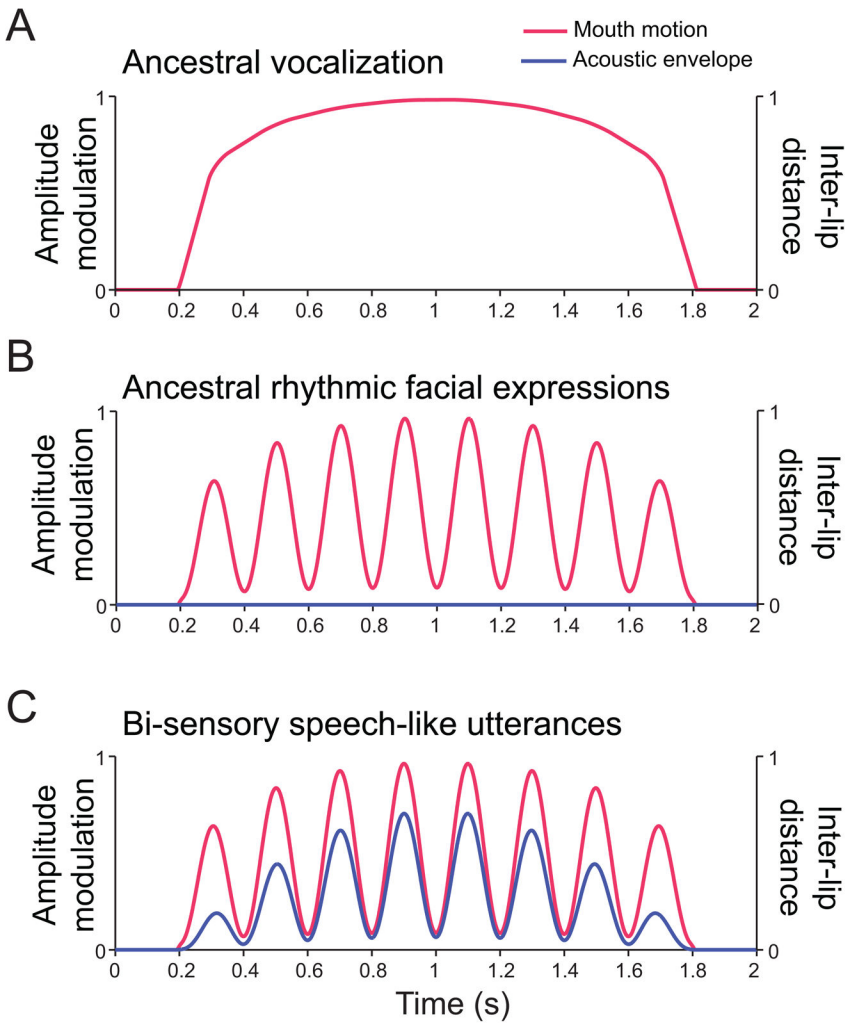

Figure 3.

Hypothetical transition from an ancestral unisensory, acoustic-only vocal rhythm to the one that is audiovisual, with both mouth movements and acoustics sharing the same rhythmicity. **A.** Schematic of a presumptive ancestral vocalization with 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.

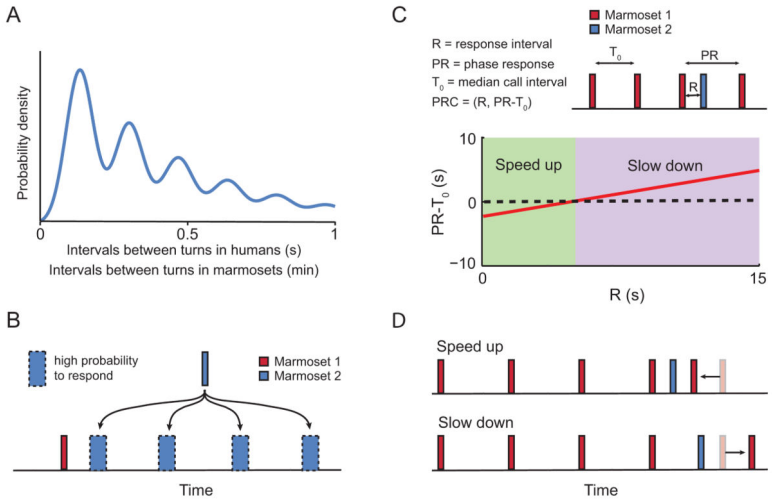

Figure 4.

Coupled oscillators dynamic of vocal turn-taking. **A.** Schematic of the probability distribution of the interval duration between turns during vocal exchanges in humans and marmosets. The same pattern of distribution of the intervals is observed in humans and marmosets, but with a difference in the time scale. **B.** Coupled rhythmicity implies that once a marmoset calls (red rectangle), the responses from a second marmoset (blue rectangle) will arrive with high probability at one of the intervals regularly spaced from each other (blue rectangle with dotted outline). **C.** Illustration of the correlation between response intervals (R) and phase response (PR) when there is an entrainment between call exchanges of Marmoset 1 (red rectangle) and Marmoset 2 (blue rectangle). When R is short (green area) PR is shorter than the median call interval (T_0), therefore there is a speed up in the call interval of Marmoset 1. When R is long (purple area) PR is longer than the median call interval (T_0), therefore there is a slow down in the call interval of Marmoset 1. **D.** Schematic of the effect of short and long R on PR. Convention as in C. The transparent red rectangle indicates where the call from Marmoset 1 would be produced had Marmoset 2 not responded.