Published in final edited form as:
Behav Ecol Sociobiol. 2013 September 1; 67(9): . doi:10.1007/s00265-013-1491-z.

Multisensory vocal communication in primates and the evolution of rhythmic speech

Asif A. Ghazanfar[1,2,3,*]

[1]Neuroscience Institute, Princeton University, Princeton NJ 08540, USA

[2]Department of Psychology, Princeton University, Princeton NJ 08540, USA

[3]Department of Ecology & Evolutionary Biology, Princeton University, Princeton NJ 08540, USA

Abstract

The integration of the visual and auditory modalities during human speech perception is the default mode of speech processing. That is, visual speech perception is not a capacity that is "piggybacked" on to auditory-only speech perception. Visual information from the mouth and other parts of the face is used by all perceivers to enhance auditory speech. This integration is ubiquitous and automatic and is similar across all individuals across all cultures. The two modalities seem to be integrated even at the earliest stages of human cognitive development. If multisensory speech is the default mode of perception, then this should be reflected in the evolution of vocal communication. The purpose of this review is to describe the data that reveal that human speech is not uniquely multisensory. In fact, the default mode of communication is multisensory in nonhuman primates as well but perhaps emerging with a different developmental trajectory. Speech production, however, exhibits a unique bimodal rhythmic structure in that both the acoustic output and the movements of the mouth are rhythmic and tightly correlated. This structure is absent in most monkey vocalizations. One hypothesis is that the bimodal speech rhythm may have evolved through the rhythmic facial expressions of ancestral primates, as indicated by mounting comparative evidence focusing on the lip-smacking gesture.

Most, but not all, primates typically live in large groups. While other mammals may also live in very large groups (e.g. herds of wildebeests), primates uniquely maintain cohesion in their groups with moment-to-moment social interactions and the specialized signaling that such interactions require. In a dynamic social environment, it is essential that primates are well equipped for detecting, learning and discriminating relevant information from communication signals. Primates need to be able to produce signals accurately (both in terms of signal structure and context) and they need to be able to interpret these signals correctly. Many of the signals that primates exchange take the form of facial expressions and vocalizations (Ghazanfar and Santos 2004). Indeed, in anthropoid primates, as group size grows, the complexity of facial expressions (Dobson 2009) and vocal expressions grow as well (McComb and Semple 2005; Gustison et al. 2012). While facial and vocal expressions are typically treated separately in most studies, in fact, they are often inextricably linked: a vocal expression cannot be produced without concomitant movements of the face.

Primate (including human) vocalizations are produced by coordinated movements of the lungs, larynx (vocal folds), and the supralaryngeal vocal tract (Ghazanfar and Rendall 2008). The vocal tract consists of the column of air derived from the pharynx, mouth and nasal cavity. Vocal tract motion not only changes the acoustics of vocalizations by changing their

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

resonance frequencies but also results in the predictable deformation of the face around the mouth and other parts of the face (Hauser et al. 1993; Hauser and Ybarra 1994; Yehia et al. 1998; Yehia et al. 2002). Different macaque monkey (*Macaca spp.*) vocalizations are produced with unique lip configurations and mandibular positions, and the motion of such articulators influences the acoustics of the signal (Hauser et al. 1993; Hauser and Ybarra 1994). For example, coo calls, like/u/in speech, are produced with the lips protruded, while screams, like the/i/in speech, are produced with the lips retracted (Fig. 1). Thus, like many of the facial motion cues that humans use for speech-reading, such cues are present in primate vocalizations as well. In light of this, one way to increase the robustness of social signals in noisy, dynamic environments is to combine the two modalities—visual and auditory—together.

Monkeys match facial expressions to vocal expressions

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. Macaque monkeys (*Macaca mulatta*), capuchins (*Cebus apella*) and chimpanzees (*Pan troglodytes*) all recognize auditory-visual correspondences between their various vocalizations (Ghazanfar and Logothetis 2003; Izumi and Kojima 2004; Parr 2004; Evans et al. 2005). For example, when tested in a preferential looking paradigm that requires no training or reward, rhesus monkeys readily match the facial expressions of ‘coo’ and ‘threat’ calls with their associated vocal components (Ghazanfar and Logothetis 2003). Rhesus monkeys can also segregate competing voices in a chorus of coos and match them to the correct number of individuals seen cooing on a video screen (Jordan et al. 2005). Finally, macaque monkeys can use vocal tract resonances (‘formants’) as acoustic cues to assess age-related body size differences among conspecifics (Ghazanfar et al. 2007). They do so by linking the body size information embedded in the formant spacing of vocalizations (Fitch 1997) with the visual size of animals who are likely to produce such vocalizations (Ghazanfar et al. 2007). These experiments demonstrate that multisensory cues could be used to help discriminate emotional signals, group size and body condition. While these represent important social information, perhaps the most important cue in social interactions is knowing *who* is signaling.

In two recent experiments, rhesus monkeys demonstrated that they could recognize familiar individuals across modalities (Adachi and Hampton 2011; Sliwa et al. 2011). In the first experiment, monkeys had daily exposure to both conspecifics and human individuals from infancy and were familiarized with both the humans and other rhesus monkeys serving as stimuli in the experiment via recent real life daily exposure (housing “roommates,” caregivers, and researchers) (Sliwa et al. 2011). In a preferential looking time paradigm, monkeys spontaneously matched the faces of known individuals to their voices, regardless of species. Their known preferences for interacting with particular individuals were also apparent in the strength of their multisensory recognition. In the second study, the evidence is rather indirect and involved training (Adachi and Hampton 2011). Monkeys performed a visual delayed match-to-sample task, where they were required to match a video of a conspecific to its photograph presented among several other photos of monkeys after a short interval. When a coo vocalization was played during this interval, it biased the monkey’s performance on this visual task towards the identity of the caller the subject heard as opposed to the individual seen in the sample video. Overall, these experiments demonstrate the multisensory recognition of individuals.

Development of face-voice matching

While there are numerous studies on the development of multisensory processes in humans and non-primate animals, there is only a handful of studies for nonhuman primates (Gunderson 1983; Gunderson et al. 1990; Adachi et al. 2006; Batterson et al. 2008; Zangenehpour et al. 2009). Understanding development is important because different species develop at different rates. Old World monkeys are neurologically precocial relative to human infants. For example, at birth, the rhesus monkey brain is heavily myelinated whereas the human brain is only moderately myelinated (Gibson 1991) and in terms of overall brain size at birth, rhesus monkeys are among the most precocial of all mammals (Sacher and Staffeldt 1974), possessing ~65% of their brain size at birth compared to only ~25% for human infants (Sacher and Staffeldt 1974; Malkova et al. 2006). If a relatively immature postnatal state of neural development leaves a developing human infant more "open" to the effects of early sensory experience then it stands to reason that the more advanced state of neural development in monkeys might result in a different outcome when it comes to multisensory behaviors (Turkewitz and Kenny 1982).

Human infants go through an experience-dependent process of "perceptual narrowing" in their processing of unisensory as well as multisensory information; that is, where initially they exhibit broad sensory tuning, they later exhibit narrower tuning. For example, 4–6 month-old human infants can match rhesus monkey faces and voices, but 8–10 month-old infants no longer do so (Lewkowicz and Ghazanfar 2006). These findings suggest that as human infants acquire increasingly greater experience with conspecific human faces and vocalizations and native multisensory speech information —but none with heterospecific faces and vocalizations and nonnative multisensory speech—their perceptual tuning narrows to match their early experience (for a review, see Lewkowicz and Ghazanfar 2009). Do precocious monkeys go through a similar cross-species developmental narrowing process for face and voice matching?

This possibility was investigated in developing infant vervet monkeys (an Old World monkey species; *Chlorocebus pygerethrus*) by testing whether they can match the faces and vocalizations of another species with which (like the human infants above) they had no prior experience (Zangenehpour et al. 2009). As in the human infant study (Lewkowicz and Ghazanfar 2006), infant vervets ranging in age from 23 to 65 weeks (~6 to 16 months) were tested in a preference task in which they viewed pairs of the same adult rhesus monkey face producing a coo call on one side and a grunt call on the other side and while hearing one of the calls at the same time. Importantly, adult rhesus monkeys look very different from adult vervet monkeys (e.g., pink or reddish face versus black face) and vervet monkeys do not produce 'coo' calls. Even though the infant vervets had no prior exposure to rhesus monkey faces and vocalizations, they matched them. That is, they exhibited cross-species matching well beyond the age of perceptual narrowing in human infants.

Why do infant vervets continue to match hetero-specific faces and voices at a postnatal and neurological age that, relative to human infants, is beyond the time when multisensory perceptual narrowing should have occurred? One possibility is that while both young human infants and monkeys start with a broad range of sensitivity, the monkeys may be "stuck" with this broad range because of the more precocial state of their nervous system. The other possibility is that monkeys' precocial brains are not stuck *per se* but, rather, are less plastic because of their more advanced developmental state (Kaas 1991). In this scenario, infant vervets may still be sensitive to social experience, but it may take them longer to incorporate the effects of such experience and, consequently, to exhibit perceptual narrowing. The latter possibility is consistent with the development of vocal behavior in vervets in that their ability to produce vocalizations, use them in appropriate contexts, and respond appropriately

to the vocalizations of conspecifics emerges gradually during the first four years of life (Seyfarth and Cheney 1986). For example, 3-month old infant vervets produce different alarm calls according to three general categories: “terrestrial predator”, “aerial predator” and “snake-like object”, but they do not distinguish between real predators and non-predators. Only over the course of years do they restrict their alarm-calling to the small number of genuine predators within each category. It is also consistent with the fact that in Japanese macaques (another Old World monkey species), unisensory and multisensory representations of faces and voices are influenced by the amount of exposure they have to conspecifics and heterospecifics (Sugita 2008; Adachi et al. 2009).

A behavioral advantage for integrating faces and voices

The matching experiments described in the sections above show that monkeys and apes can recognize the correspondence between visual and auditory signals but do not demonstrate directly whether such recognition leads to a behavioral advantage. In a vocal detection study, two monkeys were trained to detect auditory, visual or audiovisual vocalizations embedded in noise as fast and as accurately as possible (Chandrasekaran et al. 2011). Under such conditions, monkeys exhibited faster reaction times to audiovisual vocalizations than to unisensory events. The task was a free-response paradigm designed to approximate a natural face-to-face vocal communication whereby the vocal components of the communication signals are degraded by environmental noise but the face and its motion are perceived clearly. In the task, monkeys had to detect ‘coo’ calls with different levels of sound intensity and embedded in a constant background noise. For dynamic faces, two computer-generated monkey avatars allowed exquisite control, including the restriction of facial motion to the mouth region, constant lighting and background, and parameterization of the size of mouth opening while keeping eye and head positions constant. The degree of mouth-opening was in accordance with the intensity of the associated vocalization: greater sound intensity was coupled to larger mouth openings by the dynamic face.

During the task, the face of Avatar 1 was continuously on the screen for a block of 60 trials; the background noise was also continuous (Fig. 2A). In the “visual only (V)” condition, this avatar moved its mouth without any corresponding auditory component; that is, it silently produced a coo facial expression. In the “auditory-only (A)” condition, the vocalization normally paired with the Avatar 2 (which is not on the screen) was presented with the *static* face of Avatar 1. Finally, in the “audiovisual (AV)” condition, Avatar 1 moved its mouth in accord with its vocalization and with an aperture in accordance with its intensity. Each condition (V, A, or AV) is presented after a variable interval between 1 and 3 seconds drawn from a uniform distribution. Subjects indicate the detection of a V, A, or AV event by pressing a lever within two seconds following its onset. At the end of every block, a brief pause (~10 to 12 s) is imposed followed by the start of a new block in which the avatar face and the identity of the coo used for the auditory-only condition are switched.

Under these task conditions, monkeys integrated faces and voices; that is, they combined them in such a way that behavioral performance was significantly better than the unisensory conditions. This was true for accuracy and especially for reaction times (Chandrasekaran et al. 2011) (Fig. 2B). This is the first evidence for a behavioral advantage for combining faces and voices in a nonhuman primate species.

Rhythmic facial expressions: A plausible scenario for the multisensory origins of speech

As reviewed above, there are many similarities in multisensory vocal communication between monkeys and humans (and in some cases, apes). Like us, monkeys match

individual identity and expression types across modalities, can segregate competing voices in noisy conditions using vision, use formant frequencies to estimate the body size of conspecifics, and use facial motion to speed up their reaction times to vocalizations. However, there are also some important differences in how humans produce speech (Ghazanfar and Rendall 2008) and how these differences further enhance multisensory communication above and beyond what monkeys can do. One universal feature of speech—lacking in monkey vocalizations—is it's bimodal rhythm (Fig. 3). That is, when humans speak both the acoustic output and the movements of the mouth are highly rhythmic and tightly correlated with each other.

Across all languages studied to date, speech typically exhibits a 3 – 8 Hz rhythm that is, for the most part, related to the rate of syllable production (Malecot et al. 1972; Crystal and House 1982; Greenberg et al. 2003; Chandrasekaran et al. 2009) (Fig. 3A). This 3 – 8 Hz rhythm is critical to speech perception: Disrupting the auditory component of this rhythm significantly reduces intelligibility (Drullman et al. 1994; Shannon et al. 1995; Saberi and Perrott 1999; Smith et al. 2002; Elliot and Theunissen 2009), as does disrupting the visual component arising from mouth and facial movements (Vitkovitch and Barber 1996). Given the importance of this rhythm in speech, understanding how speech evolved requires investigating the origins of its rhythmic structure.

As monkey vocalizations are most often produced with a single ballistic motion (Fig. 3B), one theory posits that the rhythm of speech evolved through the modification of rhythmic facial movements in ancestral primates (MacNeilage 1998). Such facial movements are extremely common as visual communicative gestures in primates. Lip-smacking, for example, is an affiliative signal observed in many genera of primates (Hinde and Rowell 1962; Van Hooff 1962; Redican 1975), including chimpanzees (Parr et al. 2005). It 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, monkeys also exchange lipsmacking bouts during face-to-face interactions (Van Hooff 1962). Lipsmacks are among the first facial expressions produced by infant monkeys (Ferrari et al. 2006; De Marco and Visalberghi 2007) and used during mother-infant interactions (Ferrari et al. 2009). According to MacNeilage (MacNeilage 1998), during the course of speech evolution, such non-vocal rhythmic facial expressions were coupled to vocalizations to produce the audiovisual components of babbling-like (i.e., consonant-vowel-like) speech expressions.

While direct tests of such evolutionary hypotheses are difficult, there are four lines of evidence that demonstrate that the production of lip-smacking in macaque monkeys is, indeed, strikingly similar to the orofacial rhythms produced during speech. First, both speech and lip-smacking are distinct from chewing, another rhythmic orofacial motion that uses the same effectors. Importantly, in contrast to chewing movements (which are slower), lip-smacking exhibits a speech-like rhythm in the 3 – 8 Hz frequency range (Ghazanfar et al. 2010).

Second, the developmental trajectory of monkey lip-smacking also parallels speech development (Locke 2008; Morrill et al. 2012). Measurements of the rhythmic frequency and variability of lip-smacking across individuals in three different age groups (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 (Fig. 4) (Morrill et al. 2012)—this is exactly as speech develops, from babbling to adult consonant-vowel production (Dolata et al. 2008). Furthermore, as in human speech development (Smith and Zelaznik 2004), the variability and frequency changes in lip-smacking are independent in that juveniles have the same rhythmic lip-smacking frequency as adult

monkeys, but the lip-smacking is much more variable. Importantly, the developmental trajectory for lip-smacking was different from that of chewing (Morrill et al. 2012). Chewing had the same slow frequency as in humans and consistent low variability across age groups. These differences in developmental trajectories between lip-smacking and chewing are identical to those reported in humans for speech and chewing (Moore and Ruark 1996; Steeve et al. 2008; Steeve 2010).

The third line of evidence that links human speech and monkey lip-smacking comes from motor control. During speech, the functional coordination between key vocal tract anatomical structures (the jaw/lips, tongue and hyoid) is more loosely coupled during speech movements than during chewing movements (Moore et al. 1988; Ostry and Munhall 1994; Hiiemae et al. 2002; Hiiemae and Palmer 2003; Matsuo and Palmer 2010). 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-smacks (as in speech) and do so with a speech-like 3 – 8 Hz rhythm (Fig. 5A,B)(Ghazanfar et al. 2012). Relative to lip-smacking, movements during chewing were significantly slower for each of these structures. Most importantly, the temporal coordination of these structures was distinct for each behavior (as it is for human speech versus chewing) (Fig. 5C). Facial electromyographic studies of muscle coordination during lip-smacking and chewing also revealed very distinct activity patterns associated with each behavior (Shepherd et al. 2012). Thus, the production of lip-smacking and speech is strikingly similar at the level of the rhythmicity and functional coordination of effectors.

Finally, monkeys seemed to be perceptually tuned to lip-smacking with a natural 3 – 8 Hz rhythm (AAG, RJ Morrill, C Kayser unpubl. data). Artificial perturbation (e.g., speeding it up) of the speech rhythm outside the natural range reduces speech intelligibility, demonstrating a perceptual tuning to this frequency band in humans. To investigate whether monkeys also exhibit a perceptual tuning to the natural rhythms of lip-smacking, we tested rhesus monkeys in a preferential looking paradigm, measuring the time spent looking at each of two side-by-side computer-generated monkey avatars lip-smacking at natural versus sped-up or slowed-down rhythms. Monkeys showed an overall preference for the natural rhythm when compared to the perturbed rhythms. This lends behavioral support for the hypothesis that perceptual processes are similarly tuned to the natural frequencies of communication signals across primate species.

Conclusion

Human speech is not uniquely multisensory. The default mode of communication in many primates is multisensory. Apes and monkeys recognize the correspondence between vocalizations and the facial postures associated with them. While this behavioral capacity is similar to those exhibited by human infants and adults, the developmental trajectory of the underlying mechanisms may differ across species. One striking dissimilarity between monkey vocalizations and human speech is that the latter has a unique bimodal 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. Cineradiographic, developmental and perceptual data from macaque monkeys all support the notion that lip-smacking may have been one such ancestral expression. In general, more comparative data are needed. Here are two important directions to pursue. First, a study quantifying rhythmic structure of lipsmacking and vocalizations (such as the chimpanzee pant-hoot which has a clear rhythm) by great apes would be necessary to establish that the lipsmacking-to-visual speech rhythm hypothesis is a continuous trait in the Primate clade. Second, all nonhuman primate studies of multisensory processes thus far have focused on the auditory and visual domains. Yet,

some species rely on other modalities to a greater degree, such as olfaction in strepsirrhines. Thus, it would lend greater and more complete insights into the use of multisensory cues by primates if we explored, for example, olfactory-auditory recognition in lemurs.

Acknowledgments

The authors gratefully acknowledge the scientific contributions and numerous discussions with the following people: Adrian Bartlett, Chand Chandrasekaran, Ipek Kulahci, Darshana Narayanan, Stephen Shepherd, Daniel Takahashi and Hjalmar Turesson. This work was supported by NIH R01NS054898 and the James S. McDonnell Scholar Award.

References

Adachi I, Hampton RR. Rhesus monkeys see who they hear: spontaneous crossmodal memory for familiar conspecifics. PLoS ONE. 2011; 6:e23345. [PubMed: 21887244]

Adachi I, Kuwahata H, Fujita K, Tomonaga M, Matsuzawa T. Japanese macaques form a cross-modal representation of their own species in their first year of life. Primates. 2006; 47:350–354. [PubMed: 16636747]

Adachi I, Kuwahata H, Fujita K, Tomonaga M, Matsuzawa T. Plasticity of the ability to form cross-modal representations in infant Japanese macaques. Dev Sci. 2009; 12:446–452. [PubMed: 19371369]

Batterson VG, Rose SA, Yonas A, Grant KS, Sackett GP. The effect of experience on the development of tactual-visual transfer in pigtailed macaque monkeys. Dev Psychobiol. 2008; 50:88–96. [PubMed: 18085561]

Chandrasekaran C, Lemus L, Trubanova A, Gondan M, Ghazanfar AA. Monkeys and humans share a common computation for face/voice integration. PLoS Comput Biol. 2011; 7:e1002165. [PubMed: 21998576]

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS Comput Biol. 2009; 5:e1000436. [PubMed: 19609344]

Crystal T, House A. Segmental durations in connected speech signals: Preliminary results. J Acoust Soc Am. 1982; 72:705–716. [PubMed: 7130529]

De Marco A, Visalberghi E. Facial displays in young tufted capuchin monkeys (*Cebus apella*): Appearance, meaning, context and target. Folia Primatol. 2007; 78:118–137. [PubMed: 17303940]

Dobson SD. Socioecological correlates of facial mobility in nonhuman anthropoids. Am J Phys Anthropol. 2009; 138:413–420. [PubMed: 19235791]

Dolata JK, Davis BL, MacNeilage PF. Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm. Inf Beh Dev. 2008; 31:422–431.

Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am. 1994; 95:2670–80. [PubMed: 8207140]

Elliot TM, Theunissen FE. The modulation transfer function for speech intelligibility. PLoS Comp Biol. 2009; 5:e1000302.

Evans TA, Howell S, Westergaard GC. Auditory-visual cross-modal perception of communicative stimuli in tufted capuchin monkeys (*Cebus apella*). J Exp Psychol-Anim Beh Process. 2005; 31:399–406.

Ferrari P, Visalberghi E, Paukner A, Fogassi L, Ruggiero A, Suomi S. Neonatal imitation in rhesus macaques. PLoS Biol. 2006; 4:1501.

Ferrari PF, Paukner A, Ionica C, Suomi S. Reciprical face-to-face communication between rhesus macaque mothers and their newborn infants. Curr Biol. 2009; 19:1768–1772. [PubMed: 19818617]

Fitch WT. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J Acoust Soc Am. 1997; 102:1213–1222. [PubMed: 9265764]

Ghazanfar AA, Chandrasekaran C, Morrill RJ. Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: implications for the evolution of audiovisual speech. Eur J Neurosci. 2010; 31:1807–1817. [PubMed: 20584185]

Ghazanfar AA, Logothetis NK. Facial expressions linked to monkey calls. Nature. 2003; 423:937–938. [PubMed: 12827188]

Ghazanfar AA, Rendall D. Evolution of human vocal production. Curr Biol. 2008; 18:R457–R460. [PubMed: 18522811]

Ghazanfar AA, Santos LR. Primate brains in the wild: The sensory bases for social interactions. Nat Rev Neurosci. 2004; 5:603–616. [PubMed: 15263891]

Ghazanfar AA, Takahashi DY, Mathur N, Fitch WT. Cineradiography of monkey lipsmacking reveals the putative origins of speech dynamics. Curr Biol. 2012; 22:1176–1182. [PubMed: 22658603]

Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. Vocal tract resonances as indexical cues in rhesus monkeys. Curr Biol. 2007; 17:425–430. [PubMed: 17320389]

Gibson, KR. Myelination and behavioral development: A comparative perspective on questions of neoteny, altriciality and intelligence. In: Gibson, KR.; Peterson, AC., editors. Brain maturation and cognitive development: comparative and cross-cultural perspectives. Aldine de Gruyter; New York: 1991. p. 29-63.

Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech--a syllable-centric perspective. J Phon. 2003; 31:465–485.

Gunderson VM. Development of cross-modal recognition in infant pigtail monkeys (*Macaca nemestrina*). Dev Psych. 1983; 19:398–404.

Gunderson VM, Rose SA, Grantwebster KS. Cross-modal transfer in high-risk and low-risk infant pigtailed macaque monkeys. Dev Psych. 1990; 26:576–581.

Gustison ML, le Roux A, Bergman TJ. Derived vocalizations of geladas (*Theropithecus gelada*) and the evolution of vocal complexity in primates. Philos T Roy Soc B. 2012; 367:1847–1859.

Hauser MD, Evans CS, Marler P. The role of articulation in the production of rhesus-monkey, *Macaca mulatta*, vocalizations. Anim Behav. 1993; 45:423–433.

Hauser MD, Ybarra MS. The role of lip configuration in monkey vocalizations - Experiments using xylocaine as a nerve block. Brain Lang. 1994; 46:232–244. [PubMed: 8137144]

Hiiemae KM, Palmer JB. Tongue movements in feeding and speech. Crit Rev Oral Biol Med. 2003; 14:413–429. [PubMed: 14656897]

Hiiemae KM, Palmer JB, Medicis SW, Hegener J, Jackson BS, Lieberman DE. Hyoid and tongue surface movements in speaking and eating. Arch Oral Biol. 2002; 47:11–27. [PubMed: 11743928]

Hinde RA, Rowell TE. Communication by posture and facial expressions in the rhesus monkey (Macaca mulatta). Proc Zool Soc Lond. 1962; 138:1–21.

Izumi A, Kojima S. Matching vocalizations to vocalizing faces in a chimpanzee (*Pan troglodytes*). Anim Cogn. 2004; 7:179–184. [PubMed: 15015035]

Jordan KE, Brannon EM, Logothetis NK, Ghazanfar AA. Monkeys match the number of voices they hear with the number of faces they see. Curr Biol. 2005; 15:1034–1038. [PubMed: 15936274]

Kaas JH. Plasticity of sensory and motor maps in adult animals. Annu Rev Neurosci. 1991; 5:137–167. [PubMed: 2031570]

Lewkowicz DJ, Ghazanfar AA. The decline of cross-species intersensory perception in human infants. P Natl Acad Sci USA. 2006; 103:6771–4.

Lewkowicz DJ, Ghazanfar AA. The emergence of multisensory systems through perceptual narrowing. Trends Cogn Sci. 2009; 13:470–478. [PubMed: 19748305]

Locke, JL. Lipsmacking and babbling: Syllables, sociality, and survival. In: Davis, BL.; Zajdo, K., editors. The syllable in speech production. Lawrence Erlbaum Associates; New York: 2008. p. 111-129.

MacNeilage PF. The frame/content theory of evolution of speech production. Behav Brain Sci. 1998; 21:499–511. [PubMed: 10097020]

Malecot A, Johonson R, Kizziar P-A. Syllable rate and utterance length in French. Phonetica. 1972; 26:235–251. [PubMed: 4670762]

Malkova L, Heuer E, Saunders RC. Longitudinal magnetic resonance imaging study of rhesus monkey brain development. Eur J Neurosci. 2006; 24:3204–3212. [PubMed: 17156381]

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]

McComb K, Semple S. Coevolution of vocal communication and sociality in primates. Biol Lett. 2005; 1:381–385. [PubMed: 17148212]

Moore CA, Ruark JL. Does speech emerge from earlier appearing motor behaviors? J Speech Hear Res. 1996; 39:1034–1047. [PubMed: 8898256]

Moore CA, Smith A, Ringel RL. Task specific organization of activity in human jaw muscles. J Speech Hear Res. 1988; 31:670–680. [PubMed: 3230897]

Morrill RJ, Paukner A, Ferrari PF, Ghazanfar AA. Monkey lip-smacking develops like the human speech rhythm. Dev Sci. 2012; 15:557–568. [PubMed: 22709404]

Ostry DJ, Munhall KG. Control of Jaw Orientation and Position in Mastication and Speech. J Neurophysiol. 1994; 71:1528–1545. [PubMed: 8035233]

Parr LA. Perceptual biases for multimodal cues in chimpanzee (*Pan troglodytes*) affect recognition. Anim Cogn. 2004; 7:171–178. [PubMed: 14997361]

Parr LA, Cohen M, de Waal F. Influence of social context on the use of blended and graded facial displays in chimpanzees. Int J Primat. 2005; 26:73–103.

Redican, WK. Facial expressions in nonhuman primates. In: Rosenblum, LA., editor. Primate behavior: developments in field and laboratory research. Academic Press; New York: 1975. p. 103-194.

Saberi K, Perrott DR. Cognitive restoration of reversed speech. Nature. 1999; 398:760–760. [PubMed: 10235257]

Sacher GA, Staffeldt EF. Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth. Am Nat. 1974; 108:593–615.

Seyfarth RM, Cheney DL. Vocal development in vervet monkeys. Anim Behav. 1986; 34:1640–1658.

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270:303–304. [PubMed: 7569981]

Shepherd SV, Lanzilotto M, Ghazanfar AA. Facial muscle coordination during rhythmic facial expression and ingestive movement. J Neurosci. 2012; 32:6105–6116. [PubMed: 22553017]

Sliwa J, Duhamel JR, Pascalis O, Wirth S. Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. P Natl Acad Sci USA. 2011; 108:1735–1740.

Smith A, Zelaznik HN. Development of functional synergies for speech motor coordination in childhood and adolescence. Dev Psychobiol. 2004; 45:22–33. [PubMed: 15229873]

Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. Nature. 2002; 416:87–90. [PubMed: 11882898]

Steeve RW. Babbling and chewing: Jaw kinematics from 8 to 22 months. J Phon. 2010; 38:445–458. [PubMed: 20725590]

Steeve RW, Moore CA, Green JR, Reilly KJ, McMurtrey JR. Babbling, chewing, and sucking: Oromandibular coordination at 9 months. J Speech Lang Hear Res. 2008; 51:1390–1404. [PubMed: 18664699]

Sugita Y. Face perception in monkeys reared with no exposure to faces. P Natl Acad Sci USA. 2008; 105:394–8.

Turkewitz G, Kenny PA. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. Dev Psychobiol. 1982; 15:357–368. [PubMed: 7106395]

Van Hooff JARAM. Facial expressions of higher primates. Symp Zool Soc Lond. 1962; 8:97–125.

Vitkovitch M, Barber P. Visible speech as a function of image quality: Effects of display parameters on lipreading ability. App Cogn Psychol. 1996; 10:121–140.

Yehia H, Rubin P, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. Speech Comm. 1998; 26:23–43.

Yehia HC, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. J Phon. 2002; 30:555–568.

Zangenehpour S, Ghazanfar AA, Lewkowicz DJ, Zatorre RJ. Heterochrony and cross-species intersensory matching by infant vervet monkeys. PLoS ONE. 2009; 4:e4302. [PubMed: 19172998]

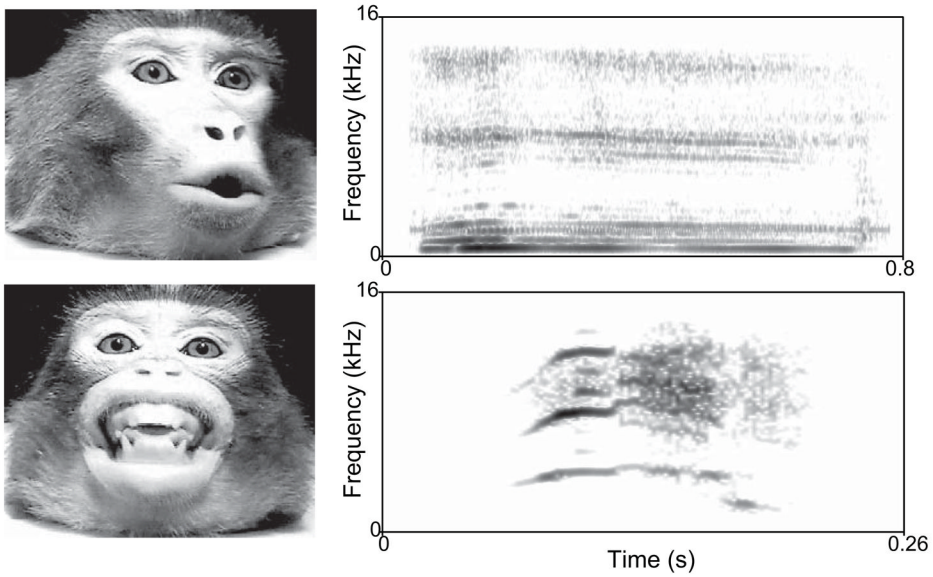

Fig. 1.
Exemplars of the facial expressions produced concomitantly with vocalizations. Rhesus monkey coo and scream calls taken at the midpoint of the expressions with their corresponding spectrograms

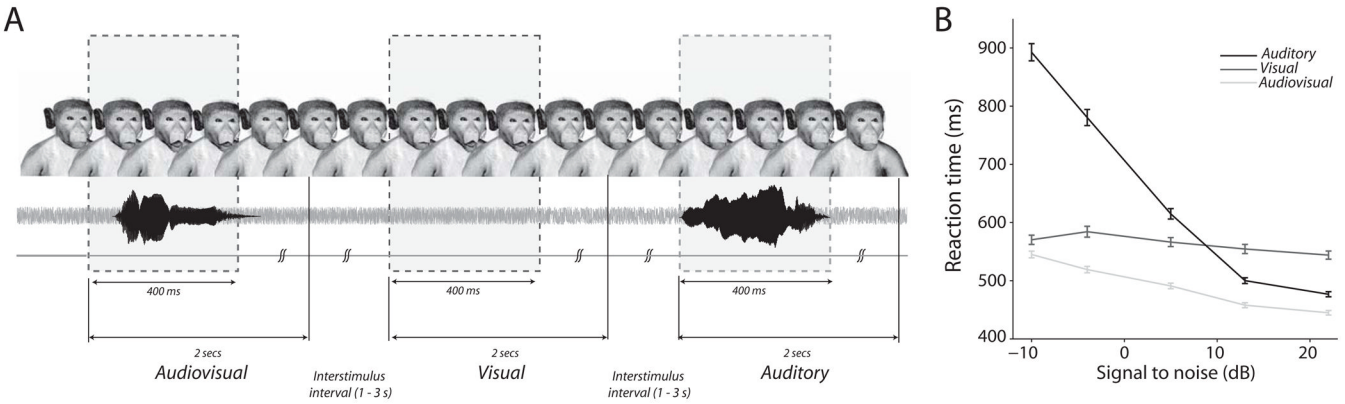

Fig. 2.

A. Free-response paradigm task structure. An avatar face was always on the screen. Visual, auditory and audiovisual stimuli were randomly presented with an inter stimulus interval of 1–3 seconds drawn from a uniform distribution. Responses within a 2 second window after stimulus onset were considered to be hits. Responses in the inter-stimulus interval are considered to be false alarms and led to timeouts. **B.** Mean reaction times obtained by pooling across all sessions as a function of SNR for the unisensory and multisensory conditions for one monkey. Error bars denote standard error of the mean estimated using bootstrapping. X-axes denote SNR in dB. Y-axes depict RT in milliseconds. Figures reprinted from Chandrasekaran et al. 2011

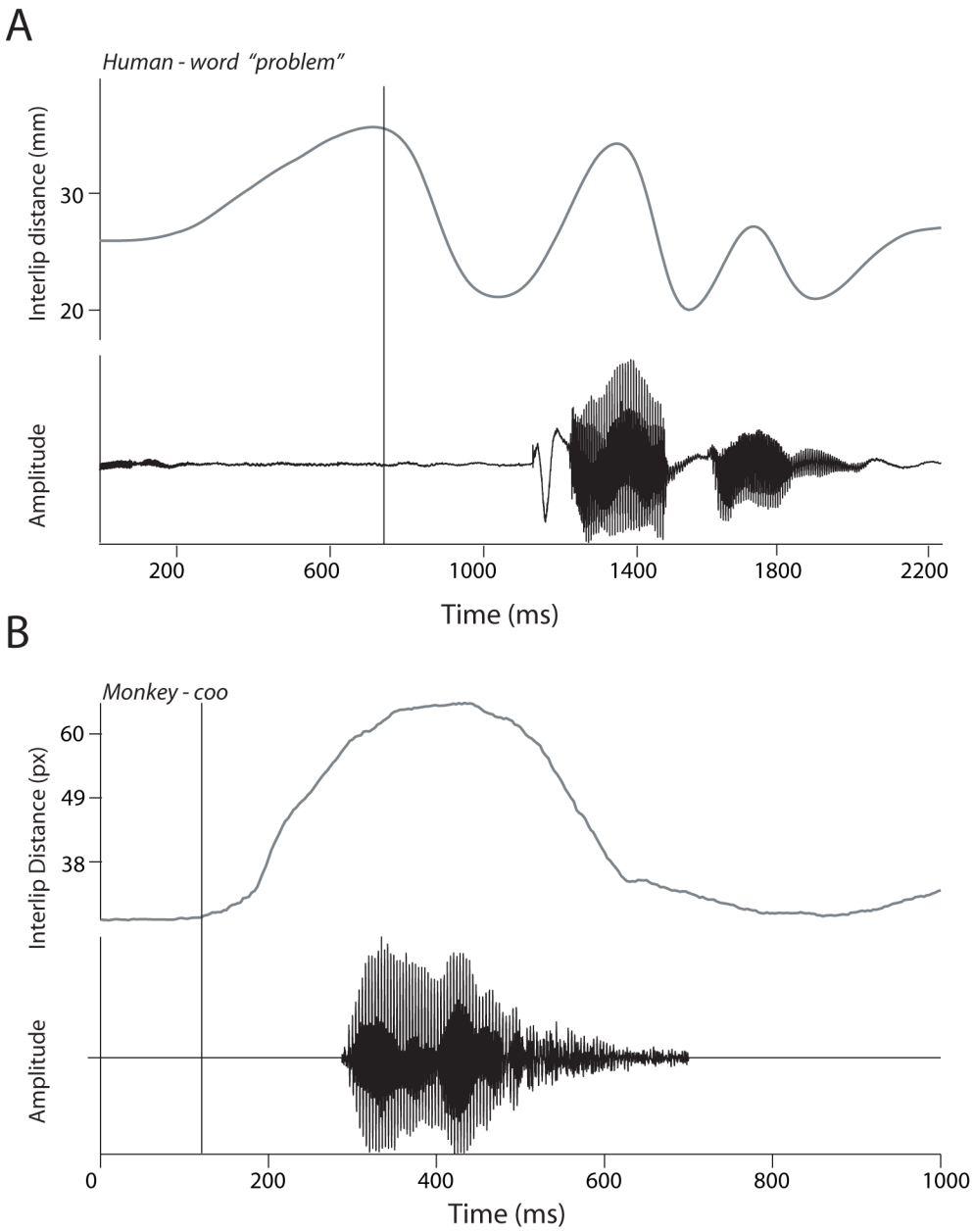

Fig. 3.

A. Visual and auditory dynamics during the production of the word "problem" by a single speaker. Upper panel denotes the inter-lip distance as a function of time. The lower panel shows waveform of the sound. Figure reprinted from Chandrasekaran et al. 2009. **B**. Visual and auditory dynamics during a coo vocalization. Figure conventions as in A

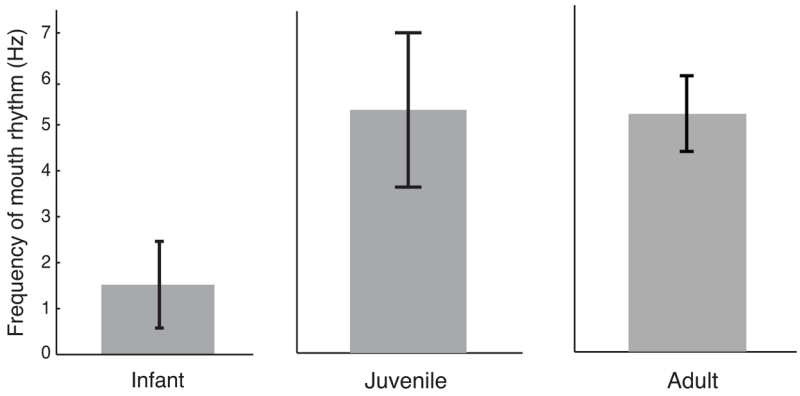

Fig. 4.
Rhythmic frequencies of lip-smacking across development. Figure reprinted from Morrill et al. 2012

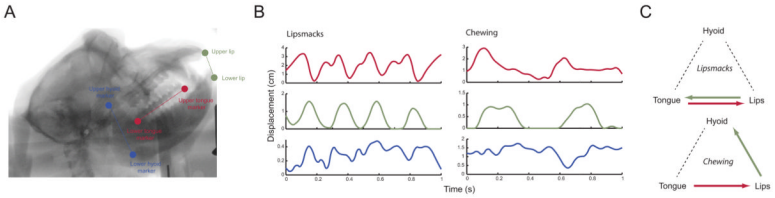

Fig. 5.

A. The anatomy of the macaque monkey vocal tract as imaged with cineradiography. The key vocal tract structures are labeled: the lips, tongue and hyoid. **B.** Time-displacement plot of the tongue, inter-lip distance, and hyoid for one exemplar each of lip-smacking and chewing. **C.** Arrow schematics show the direction of significant influence from each structure onto to the other two as measured by the partial directed coherence analysis of signals such as those in B. Figures reprinted from Ghazanfar et al. 2012