10

Talking Heads

We saw that songbirds and other vocal learning birds can serve as very interesting models for the acquisition of human speech. However, because not only their brains but also their peripheral phonatory systems have different evolutionary histories from those of humans, they provide only limited insight into the specific origin of speech. For this, we have to turn our attention to vocalizing mammals that at least share a cerebral cortex, homologous brainstem nuclei and vocal tract, vocal folds and larynx, as well as to human infants and children acquiring speech. Non-human primate calls, like those of other mammals, convey information to their conspecifics about the individual's social status, sex, age and identity. The transition from such stereotyped vocalizations to vocal learning and then to speech has become a major issue in the study of language evolution. In this chapter I will argue that speech is a complex behavior, in which historical continuity between monkeys, apes and humans can be found in the peripheral organs involved in speech, in their neural control, and in the cerebral networks involved in speech processing. The gradual evolution of these systems was a requisite for the acquisition of the phonological loop, and the origin of modern speech.

F. Aboitiz, *A Brain for Speech*,
DOI 10.1057/978-1-137-54060-7_10

Vocal Beasts

There are well-known examples of domestic dogs and other animals that can learn, often quickly, the meanings of many words and even phrases uttered by humans (Kaminski et al. 2004; Andics et al. 2016). However, the evolution of vocal learning is driven by modifications of the motor rather than the auditory system, which is more conserved across species, and relies on general-purpose mechanisms (see Chapter 9) (Feenders et al. 2008). While vocal learning is present in only a few species, the capacity for auditory learning is widespread in vertebrates.

Some species display voluntary use of vocalizations depending on the context, as some animals can learn to vocalize only in specific circumstances, and can modulate the intensity or duration of their calls in different settings. However, these vocalizations remain stereotyped in structure. This is seen in apes and monkeys, domestic animals and many other species (Hauser 1996). What is more interesting to us is the capacity to learn new sounds by imitation and to develop novel strings with different combinations of sounds. Although we seem to be alone among primates in our vocal abilities, other mammalian species have shown a sometimes impressive capacity to imitate, not only sounds made by conspecifics, but also physical phenomena and even the human voice. Vocal imitation is observed in seals, toothed whales and elephants, while the ability to generate new sequences of sounds generating a rudimentary syntax is much less common. Furthermore, the capacity for vocal learning is not strictly a gift given to some animals, but there are different levels of vocal learning abilities in different species. Christopher Petkov and Erich Jarvis categorized species on the basis of their vocal learning capacities, where animals like monkeys display a limited learning capacity, songbirds and parrots are relatively complex vocal learners and our species apparently exceeds all the others in the voluntary and fine control of phonation (Petkov and Jarvis 2012).

Echolocating species like bats and cetaceans are good vocal learners. These animals have developed several anatomical and biochemical hearing and vocal specializations that allow them to hear and emit a much wider range of sound frequencies (into the ultrasonic range) than can humans. Particularly, the protein prestin, which is expressed in external ciliary cells of the cochlea (they are only found in mammals), provides motility to the receptors, dramatically increasing their sensitivity. Notably, convergent mutations have been reported in this protein in echolocating bats and cetaceans that enhance their ultrasonic hearing range (Caspermeyer 2014). Furthermore, work in Nobuo Suga's laboratory has shown that the auditory cortex of some bats is proportionally quite large, and is subdivided into several highly specialized areas (Suga 1989). There is an area that processes frequency modulated signals (called FM-FM area), with neurons sensitive to the time delay of the echo, providing information about the distance to a target. Neurons in another area (called CF-CF area) process a constant frequency (around 30 Hz, the frequency of the echolocating call) and its harmonics, which is used to calculate changes in echo frequencies produced by the bat's velocity relative to a target (the Doppler effect). A third component is the DSCF area, which represents the frequency range of maximal sensitivity (an acoustic fovea), matching the echo call frequency. Bats in enclosed environments rely more on the FM signal, while bats in open environments use the CF signal more. It is not yet clear whether these auditory specializations are restricted to echolocation or if they also participate in vocal learning mechanisms. If this is so, it would be an exception to Erich Jarvis' claim that the evolution of vocal learning is mainly dependent on motor adaptations. Finally, there is significant asymmetry in the acoustic properties of the FM-FM and the DSCF areas of the bat auditory cortex, with the left hemisphere more sensitive to echo time delays, while the right specializes in frequency analysis (Washington and Tillinghast 2015).

Bats not only use their calls for echolocation but also sing intensely when landing on their roosts (and sometimes also when flying), producing songs that are sometimes as complex as those of highly specialized songbirds (Bradbury and Emmons 1974). As with bird songs, there is a tremendous variety of "bat songs" across species (Knörnschild et al. 2010; Morell 2014), some of which are simple and involve the rhythmical repetition of one variable note, while the songs of other species are hierarchically organized but also highly flexible, with structures that change according to circumstances, such as the appearance of males nearby. In species of the genus *Pipistrellus*, songs are formed by a sequence of phrases that signal the animal's species, the individual's identity, group information, and a landing site signal.

Like many songbirds, male bats are usually the ones that sing, both to signal territory and to attract mates. Some bats use the same song for both functions, but others have different songs for territorial disputes and courtship. Furthermore, bats do not sing when they are alone, and do not sing only in the mating season. Bats learn their songs by babbling and imitating, like songbirds and humans. The learning process may extend to the echolocation call, as young horseshoe bats adapt their echolocating call to that of their mother, and in many species pup calls have individual signatures that distinguish them from those of other pups in the colony.

Harbor seals, belugas and elephants have been reported to imitate human speech, while some whales display very long complex songs that change with time in synchrony with other members of the group (Ridgway et al. 2012; Stoeger et al. 2012; Reichmuth and Casey 2014; Janik 2009). As with bats, whale songs have repetitions of elements that can be organized into syntax-like structures similar to those of songbirds (Janik 2013, 2014; Janik and Sayigh 2013, King and Janik 2013). The nasal cavities of dolphins are rather complex, with two phonic lips and two air sacs that enable them to control the vibration of air before it comes out through the blowhole at the top of the head. The emitted sound is transmitted and radiated through the melon, a fatty deposit in the front of the head. Dolphins use vocalizations ("nasalizations" would apply better in this case) to localize prey and to explore their surroundings. As mentioned by Stephanie King and Peter McGregor cetacean vocalizations facilitate social bonding and group synchrony, which makes it perhaps more comparable to early human communication than birdsong (see Chapters 8 and 9) (King and McGregor 2016). Vincent Janik and collaborators observed that each individual dolphin produces a learned but individually specific signature call that allows them to recognize one other. However, this signature call is not categorically different from other calls shared by all members of the group, but is rather a variant of these. Recent research suggests that dolphin mothers start singing to their babies before they are born, apparently teaching them their signature whistle, a process that continues over the first weeks after birth (King et al. 2016)[1]. The signature call of

[1] http://www.livescience.com/55699-mother-dolphins-teach-babies-signature-whistle.html

dolphins serves to maintain group cohesion, as individuals separated from the group emit their signature whistles while the others respond with their own signatures until they come together again.

Although mice have traditionally been ignored in vocal learning studies, more recent studies by Julia Fischer and by Erich Jarvis and Gustavo Arriaga have focused on mouse ultrasonic vocalizations, which are remarkably song-like and produced by males to attract females (Fischer and Hammerschmidt 2011; Hammerschmidt et al. 2009, 2012; Arriaga et al. 2012; Arriaga and Jarvis 2013). Adult male mice produce sonic vocalizations by vibrating their vocal folds, and ultrasonic vocalizations by expiring air while maintaining the vocal folds tight and forcing the air to pass through a rigid slit. Male mice use and modulate ultrasonic vocalizations in different social contexts such as in the presence of other males. Deafening mice early in life affects the syllabic structure of the animal's calls, which suggests that mice song learning is partly a sensory-driven learning process, although it is not yet clear whether the songs develop through a babbling-like stage as in humans and songbirds. A laryngeal motor cortex-like region was recently described by Arriaga and Jarvis, which is active during vocalizations and makes a direct although faint projection into the nucleus ambiguus (Arriaga et al. 2012).

Noisy Primates

Although our primate cousins are highly vocal species, little vocal learning capacity has been observed in them, which is another argument for how unusual we are in our evolutionary family. But this is not to say that non-human primates have absolutely no voluntary control over their vocalizations. For example, the Indri, a social lemur from Madagascar, develops long cries that are synchronized along the group, but the song varies in structure according to social condition (Gamba et al. 2016). Monkeys and apes can choose when and what to vocalize depending on contextual cues, and there is evidence of voluntary control over the upper vocal tract, including lip musculature (Hage et al. 2013, 2016). Early studies showed that socially isolated and deafened monkeys display somewhat abnormal vocalizations (Newman and Symmes 1974; Talmage-Riggs et al. 1972; Egnor et al. 2006). South American tamarin monkeys are highly loquacious, and there are call differences among subspecies. Some individuals develop mixed calls, but these could be hybrids among subspecies (Bradley and McClung 2015). There are some indications that marmosets modify their calls and vocalizations depending on the presence of other individuals, but these consist of modulations of pitch and other acoustic features, and no evidence has yet been found of learning new vocalizations (Miller et al. 2003; Bezerra et al. 2009; Seyfarth and Cheney 2010). Nonetheless, Asif Ghazanfar and associates recently reported that the development of marmoset vocalizations, as in songbirds and humans, proceeds from an initial stage of highly variable sounds that become clustered in acoustical properties as individuals mature (Takahashi et al. 2015). But perhaps more interestingly, parental vocal feedback is a key variable in the maturation of vocalizations of marmosets, songbirds and humans, which indicates a common substrate for early vocal learning across species.

There is some evidence of vocal plasticity in old world monkeys, possibly because they have been heavily studied (Hage and Nieder 2013). The calls of rhesus monkeys display troop differences, and cross-fostered individuals develop the typical call of the group they belong to (Owren et al. 1993). Studies in the late twentieth century showed that these animals can be trained to modulate the amplitude and duration of their calls to match them with playbacks presented to them, but again these modifications are only in the length and strength of the expiration phase (Hauser 1996, 1998). Likewise, guenon monkeys have been observed to use specific combinations of calls in specific circumstances, like initiating group movements (Candiotti et al. 2012; Arnold and Zuberbühler 2012). In this context, David Reby and colleagues described a noticeable parallel between human and ape vocalizations, reporting that both species voluntarily modulate the fundamental frequency of speech, increasing or decreasing pitch in different contexts, as when one talks to a man, woman, or infant, or when tries to impose authority (Pisanski et al. 2016). These basic control mechanisms are shared with other mammals, especially primates, and are proposed to represent a starting point from which voluntary vocal control evolved in our species.

Some primates are characterized by their long complex calls that can be heard from very far away. One of these is the howler monkey in South America, which has been little studied, and the others are the gibbon and siamang of South Asia. Gibbons are relatively close to us, as we both belong to the hominoidea superfamily of Old World monkeys, commonly called "apes", which separated from other monkeys about 18–20 million years ago. Gibbons are arguably the most elaborate vocalizers among non-human primates (Clarke et al. 2006). They live in stable couples and both sexes sing to defend their territory, similar to tropical songbirds discussed in the previous chapter. Like some songbirds, male and female gibbons usually sing in coordinated duets where the female leads (Geissmann 2002). However, gibbon songs are highly stereotyped and show little geographic variation, and hybrids between two species develop a mixed call between the two parent species (Brockelman and Schilling 1984). Furthermore, phylogenetic trees have been established based on the divergence of song patterns of different species that match genetically based trees (Thinh et al. 2011). There is a role for learning in the gibbon song, which has mainly to do with coordinating vocalizations while making the male and female duet, although some evidence suggests that there is a maternal role in the maturation of the song's structure (Koda et al. 2013). It has been reported that gibbons can also modulate and change the organization of their songs in the presence of predators, but this needs to be confirmed. In summary, our primate lineage has been characterized by a limited vocal learning capacity. Even our closest relatives, chimpanzees and bonobos, have demonstrated little ability to imitate and combine vocal utterances, aside from the voluntary control of lip movements and facial gestures.

As they have limited respiratory control, many non-human primates, especially highly vocal ones like gibbons, some apes and howler monkeys, have developed laryngeal sacs, which are extensions of the vocal tract cavity principally in the laryngeal region (Hewitt et al. 2002). One interpretation is that the sacs serve as a reservoir of air that allows for producing longer and more rapid calls, or to provide an additional sound source without the risk of hyperventilating. There is some support for this hypothesis as vocalizations of primates with air sacs are free from body size constraints that apply to primates without air sacs, which limit respiratory frequency and call duration. According to another view, air sacs increase resonance to amplify vocalizations, or to provide the impression of being produced by a larger individual. It has been claimed that there is evidence of vestigial air sacs in humans, but this is controversial. As well, speech may not be as energy demanding as the calls of gibbons and other species. We will see below that early humans took another route to generate complex vocalizations, based on restructuring the laryngeal cavity.

Chimpanzees have been reported to elicit referential food calls that provide information about food quality to the group (Lalamn and Boesch 2013). The acoustic structure of such food calls was reported to adapt to that of a different group when an outsider joined it (Watson et al. 2015). However, these findings have been contested on the grounds that these calls and their modulation were likely driven by emotional arousal rather than directly signaling food quality (Wheeler and Fischer 2012; Fischer et al. 2015). Evidence for vocal modulation has also come from gorillas. After an extensive analysis of recorded videos, Markus Perlman and Nat Clark recently concluded that Koko, the gorilla raised by Francine Patterson (see Chapter 1), was able to modulate breathing-related vocalizations with tongue and lip movements (Perlman and Clark 2015). Furthermore, this was done in the context of manual actions and gestures. More recently, a study reported food-associated “songs” of gorillas, but it is still unclear whether these are actually learned (Luef et al. 2016). Finally, orangutans have been reported to imitate the human voice[2] (Lameira et al. 2015, 2016; Wich et al. 2009). Although not our immediate ancestor like the chimpanzee, this species can provide a useful model for acquisition of an open-ended vocal repertoire (see below).

[2] https://www.youtube.com/watch?v=0zr2eunVDxw

Neanderthal Throats

To begin with the ancestry of human speech, it is important first to depict the structure of our phonatory organs. These can be divided into a deep component that generates the vocal sounds, consisting of the larynx and the vocal folds. Vocal folds are two membranous infoldings that make up a vibrating groove in the upper larynx that generates a variety of acoustically complex sounds, depending on their vibration rate. The folds are controlled by the nucleus ambiguus of the brainstem, via the vagus nerve. In turn, the nucleus ambiguus receives input from the vocal motor cortex, providing voluntary control of the vocal folds. Some authors have equated cortical control of vocal fold motoneurons to the capacity for vocal learning, but we will see that there is more to tell about this. The sound generated by the vocal folds resonates in the oral cavity until it is expelled and radiates outwardly at the lips. The other component of the speech apparatus is provided by superior or supralaryngeal organs including the oral and nasal cavities, which provide a sound filter that modulates the air column; and the tongue and the lips, which control airflow and are critical for the production of most consonants.

In the late 1960s and early 1970s, Philip Lieberman observed that the human vocal tract was longer, and the larynx in a lower position than in other primates (Fig. 10.1) (Lieberman 1968, 1979, 1984). This results in an additional cavity, the pharyngeal, located between the larynx and the oral cavity that increases resonance in the air column. According to Lieberman, this makes it possible to produce a diversity of vowel types, but has the drawback of increasing the risk of choking while ingesting food, as the larynx remains open while swallowing. This impedes us from breathing and swallowing at the same time. Interestingly, human babies are born with a high-positioned larynx, which enables them to swallow and breathe simultaneously. With age, the position of the larynx descends to its adult position. Thus, our larynx diverges from that of other primates, but in the infant it is similar to that of non-human primates, presumably to avoid choking in the young. This apparently contrasts with the widespread notion, championed by the late Stephen Jay Gould, and mentioned earlier in this book, that we are juvenilized

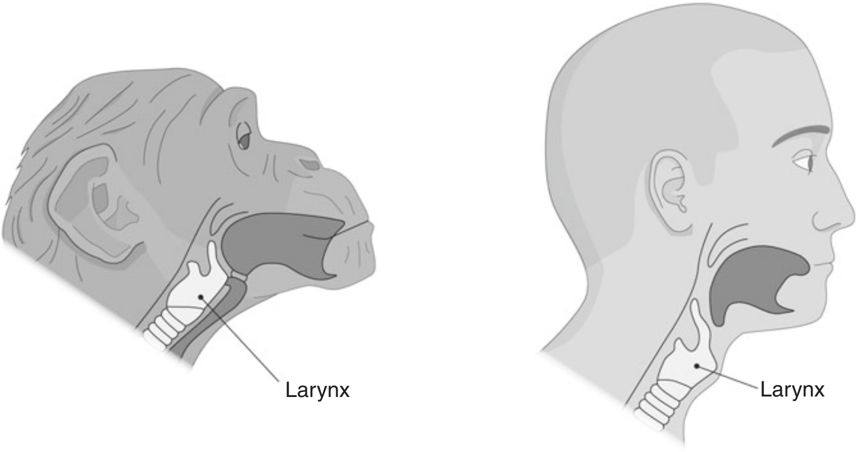

Fig. 10.1 The position of the larynx in the chimpanzee and in the human

apes that have retained infant-like features into adulthood (see also Chapter 11). Nonetheless, another possible explanation is that the shortening of the human face and jaw, a juvenilized character as we discussed above, forced the movement of the larynx in order to maintain the length of the vocal tract to permit adequate breathing and swallowing (Coquerelle et al. 2013).

Lieberman and his collaborators went further in their analyses into the structure and position of the hyoid bone in Neanderthals. Lieberman concluded that the larynx of this species was in an upward position relative to that of humans; hence they were limited in the vowels they could produce. This sparked intense debate about Neanderthal speech, which was not really evidence-based as the vocal tract and larynx are largely formed of soft tissue that does not fossilize, and inferring the position of the hyoid bone under these conditions is very questionable. Furthermore, biomechanical analysis of the Neanderthal hyoid suggests that Neanderthals were able to pronounce vowels as well as we do (D'Anastasio et al. 2013). Finally, the descended larynx is not uniquely human, being present at least in deer. Tecumseh Fitch noted that while at rest, the deer larynx has a position similar to that of the human larynx, while when deer roar the larynx descends, even reaching the sternum (Fitch and Reby 2001). According to Fitch, the descended larynx contributes to decreasing resonant frequencies and exaggerates perceived body size, a feature of sexual selection. This is because larger animals with larger vocal tracts tend to generate lower frequency vocalizations than smaller animals, all else being equal. This interpretation has also been invoked for the lower voice tone acquired by men compared to women after puberty. In any case, whatever the reason for the origin of the descent of the larynx and its status in Neanderthals, our ancestors may have taken advantage of it to produce a more diverse vocal repertoire.

Read My Lips

"Read my lips: no new taxes", said US presidential candidate George H. W. Bush in 1988, a promise that although probably key to his winning the election, remained unfulfilled during his term, and was an important reason for his losing the next election to Bill Clinton. The deaf and some hearing people are usually able to understand speech just by looking at lip movements. In fact, speech is a lot more than just the larynx and vocal folds. The word "language" itself derives from the Latin "lingua", which means "tongue". Most research on vocal communication in primates and other mammals has focused on the larynx and vocal folds, and dismissed the crucial role of the upper vocal tract. While we have gathered some evidence on the neural processes involved in controlling the larynx, comparative studies on the generation of lip and tongue movements are only beginning. The human tongue and lips are highly movable and allow us to modulate the air column to produce different types of sounds that are essential for speech and for phoneme production. The lips are innervated by the facial nerve and nucleus, like other muscles for facial expression, and the tongue is innervated by the hypoglossal cranial nucleus. Both structures are under voluntary control in monkeys and in humans, although non-human primates seem to have finer control of lip than of tongue movements. A recent study showed that the hypoglossal nucleus in monkeys is directly innervated by the motor and premotor cortices in addition to other regions like the insula and the anterior cingulate gyrus (Morecraft et al. 2014).

While the vocal folds are mostly involved in vowel production, the lips and tongue participate in both vowel and consonant production. Furthermore, consonants are made largely by rapid movements of the tongue and lips, most of them not requiring vocal fold vibrations to be produced. Consonants outnumber vowels in practically all human languages, there being overall some 600 while there are only about 200 vowels in all known languages. In this line, Adriano Lameira and collaborators called attention to supralaryngeal control in the evolution of speech, arguing that while the separation between vocal learners and vocal non-learners is rather clear in primates, a continuum can be observed between humans and other primates in the control of voiceless calls involving the lips, tongue and jaws (Lameira et al. 2014). There are a variety of such calls in apes, including "clicks", "smacks" (I referred to lip smacks in Chapter 8), "kissing sounds" and "whistles". In contrast to their voiced calls, the voiceless calls of apes are apparently influenced by social learning and fine-tuned with experience. There is also evidence of imitation of voiceless calls like lip vibrations by chimpanzees. Lameira and collaborators mention the case of Viki, the vocally trained chimp, who learned to say words like "mama", "papa" or "cup", by using only the supralaryngeal tract (see Chapter 1) (Hayes and Hayes 1951). But orangutans seem to be vocally more gifted than chimps, being able to mimic not only the human voice (see above), but also voiceless sounds like whistles (Lameira et al. 2013, 2015). Dialect-like vocal variability has been observed in different orangutan populations, some using "raspberry calls" during nest building, while others use lip smacks, and still others are silent for the same behavior; although it still needs to be shown that these differences are the product of learning. Lameira and colleagues claim that a first stage in the evolution of speech was represented by control of the supralaryngeal tract, which was then followed by voluntary modulation of the vocal folds (Lameira et al. 2014). The tongue and lips are also fundamental for the production of vowels. The vertical and horizontal positions of the tongue largely determine the type of vowel that is being produced, as their combinations define two axes that determine what is called the vowel diagram: closed vs. open vowels (e.g. /i/and /u/vs. /ae/, respectively), and front vs. back vowels (e.g. /i/ and /ae/vs. /u/and /o/). Thus, the production of both vowels and consonants depends perhaps more on the upper rather than the lower vocal tract.

Finally, in a very recent study, Tecumseh Fitch and Asif Ghazanfar analyzed with X-ray videos the jaw movements of a macaque while it ate, drank, yawned, vocalized and made lip smacking movements, and compared these movements with the oral movements involved in human speech. Notably, they observed that all oral movements required for speech are within the macaque's repertoire. Furthermore, the researchers were even able to generate a computerized simulation of the monkey's vocal tract, uttering sentences like "will you marry me?". The result was that the simulation was perfectly able to generate these movements, producing an acoustically intelligible sentence. To the authors, this implies that the vocal tract of macaques is already equipped for speech. The difference, the authors say, must rely on the neuronal control of these movements (Price 2016). An additional study reported that monkeys emit vowel-like sounds in their daily behavior (Boë et al. 2017). Despite this evidence, it is yet possible that further modifications of the upper vocal tract, like lowering the larynx, may have contributed to optimize vocal production in our recent ancestors.

The Origin of Rhythm

The rhythmic movements of voice and lips are an essential component of speech, and are marked by the duration of vowels, and the variability of consonant intervals. Speech rhythm is determined by changes in intensity, duration and pitch, in a complex hierarchy that ranges from phonemes to higher order linguistic units, and is considered to be fundamental in infant language learning. Notably, the rhythmic characteristics of a language determine some of its fundamental grammatical structure, as in object-verb languages and verb-object languages. Furthermore, speech rhythm is perceived multimodally, using both acoustic and visual cues, from observing lip movements (Peña et al. 2016). In humans, the perception of rhythm has been associated with activation of dorsal premotor areas, the basal ganglia, and auditory regions, showing some overlap with language networks, particularly as auditory-prefrontal connectivity is apparently mediated by the parietal lobe.

The analysis of lip movements of monkeys has provided a new and interesting perspective that may provide a plausible scenario of speech origins. Tecumseh Fitch, Asif Ghazanfar and collaborators found that lip smacking movements in monkeys are dissociated from throat movements and have a frequency of close to 5 cycles per second, which is similar to the frequency of lip movements during human speech and much more rapid than chewing (Ghazanfar et al. 2012). Like human lip movements, monkey lip smacking gradually changes during development from variable and slow movements to rapid and stereotyped movements (Morrill et al. 2012). Jaw, lip, tongue and hyoid bone (to which laryngeal muscles are attached) movements are coordinated during speech and chewing. However, these movements are more closely tuned for chewing than either speech or lip-smacking. David Poeppel and collaborators have shown that during speech perception, the human auditory cortex displays oscillations at about the same frequency as lip smacking movements (see Chapter 2) (Chait et al. 2015; Ding et al. 2016).

Our faces also move concomitantly with speech, making both speech-induced and indirectly related emotional expressions that are processed multimodally in the ventrolateral prefrontal cortex (including Broca's area). In human speech, facial movements (particularly those of the mouth) are finely coordinated at specific rhythms with the emitted sounds. However, monkey vocalizations dissociate from gestures, and rhythmicity takes place only in the acoustic dimension. Ghazanfar and collaborators proposed that vocalizations in the human lineage synchronized with lip-smacking, which in a first instance would have involved babbling-like expressions that evolved into speech-like behavior (Ghazanfar and Takahashi 2014a, b). Notably, primates use lip-smacking as an approaching signal or during close contact like grooming and face-to-face interactions. An instance of lip-vocal synchronization has been observed in male geladas (a large baboon-like monkey), which make a sound called a "wobble" to approach females. In another study, Morgan Gustison and colleagues found that as in human speech, gelada vocalizations include shorter segments as the vocalization they make is more complex, which follows Menzertah's law that longer words tend to have shorter syllables (Gustison et al. 2016). This may make the gelada a particularly good model to study speech origins.

Since rhythmic behavior may be a key element in the evolution of speech, researchers have investigated whether non-human animals are able to perceive rhythmic sequences. Monkeys and apes display very limited capacity for rhythmic beat, while vocal learning birds like parrots can synchronize to rhythmic beats, but not as well as humans do. Considering this evidence, Anhiruddh Patel proposed the "vocal learning and rhythmic synchronization hypothesis" that vocal learning is required for the capacity to synchronize motor outputs with a musical beat (Patel et al. 2009). Likewise, Ghazanfar proposed that the perception of rhythm partly depends on synchronization between the upper and lower vocal tracts (Ghazanfar and Takahashi 2014a, b). However, Peter Cook and colleagues, and others more recently showed that the California sea lion, whose vocal plasticity is supposedly below that of typical vocal learners, is able to follow acoustic rhythms by head bobbing (Cook et al. 2013, Ravignani et al. 2016, Rouse et al. 2016). More studies are needed to determine the vocal imitation capacities of these animals. In any case, if the ability to learn rhythmic beats is only present in vocal learning animals, then the capacity to perform rhythmic movements with the hands, as in tool making, might be a derivative of vocal learning capacity rather than the other way around, hand skills inducing vocal rhythmic capacity, as the gestural theory of language origins may suggest.

Ghazanfar and colleagues also addressed another key aspect of speech, which is the production of conversational vocal exchanges between people, with distinct gaps of silence between each turn (Takahashi et al. 2013, 2015, 2016). The capacity for reciprocal conversation is a universal character of speech, and is seen even in babbling children. Conversations may not have been produced originally to convey meaning, but rather as a bonding mechanism. We saw that gibbons engage in such vocal turn-taking behavior. In addition, Takahashi, Ghazanfar and collaborators have analyzed in more detail the turn-taking behavior of the marmoset monkey, whose vocalizations are more variable than those of gibbons. Marmosets are highly social and cooperative, showing biparental or multiparental care of their young. These monkeys communicate vocally, with few manual gestures, and engage in vocal exchanges where both individuals take turns. Marmoset conversations may take up to 30 minutes, with a turn-taking frequency that is not unlike that of human conversations. Another conversing animal seems to be dolphins, which can match each other's vocalizations in the wild, alternating their respective emissions.

The Melodic Ape

Music is acquired in a similar way as is speech. It is learned quite early in life, as studies suggest that 1- or 2-year-old babies already have a capacity to follow music by vocalizing and moving their limbs. As with speech learning, babies memorize melodies in terms of sequences that are more likely to occur than others, which then generate expectations of the notes that will follow in a given melody (Patel 2008). The brain regions involved in musical perception include the auditory cortex, premotor cortex, intraparietal sulcus, and other emotion-related regions like the anterior insula and the anterior cingulate cortex, the latter also involved in the control of innate vocalizations in both humans and monkeys (Patel 2008). Music has an intricate and recursive structure that has often been compared to that of language. Notably, important researchers in the neuroscience of music like Robert Zatorre and Anhiruddh Patel emphasize that the capacity to store sequences of notes in working memory is critical for learning and anticipating melodies (see Chapter 6) (Patel 2008; Grimault et al. 2009; Peretz and Zatorre 2005; Zatorre 2003; Zatorre and Salimpoor 2013). This is subserved by strong loops connecting frontal and temporal cortices, using both the dorsal (conveying time and sequence information) and ventral auditory pathways (conveying signals like pitch and emotion).

Darwin proposed that speech could have first evolved as primitive melodies that are controlled by the vocal folds rather than the lips and tongue, which seem to be more related to rhythm. This is partly supported by comparative evidence showing that vocal learners are usually melodic, and that melodic-like vocalizations can be found in many species including our relative the gibbon. However, as Timothy Gentner's group recently argued, the acoustically relevant cues of these "songs" may be more speech-like than melodic (see Chapter 9) (Bregman et al. 2016).

Another defender of a musical origin of speech is Tecumseh Fitch, who proposes the existence of a "musical protolanguage", the central aspects of which were prosodic and phonological, and from which a hierarchic syntax-like organization emerged (Fitch 2009, 2010). Fitch argues that as with songbirds this melodic protolanguage characteristically lacked a rhythmic component. However, if we accept that superior vocal tract-based consonants are more related to rhythm, and that these may have been under voluntary control before the control of exhalative vocalizations, it seems more likely that this musical protolanguage contained both melodies and rhythmic components. From this initial condition, vocal communication may have diverged into songs on the one hand, and semantically based speech on the other.

There are some disorders of music perception and production, one of which is called amusia. Studies by Patel have revealed that congenital amusia implies speech impairments, particularly in the capacity to perceive intonation and emotion, and in the capacity to maintain auditory patterns in working memory (Liu et al. 2015). Moreover, Maija Hausen and her collaborators assessed the perception of music and speech prosody in healthy adults, evidencing an association between the two capacities but not with visual perception (Hausen et al. 2013). Another similarity between prosodic and musical processing was reported by Anastasia Glushko and collaborators, who evidenced that the closure positive shift, a specific event-related potential associated with prosodic phrase boundaries, is also observed at the onset of musical phrase boundaries (Glushko et al. 2016). Like prosody, harmony perception in untrained individuals tends to be right-lateralized (rhythm is left-lateralized), although trained musicians use their left hemisphere to process musical information (Springer and Deutsch 1981; Patel 2008). Thus, (proto-)music and (proto-)speech may be separated concomitant with enhancement of a rudimentary level of brain lateralization, music to the right and speech to the left. Subsequently, with the appearance of more complex musical aspects that require syntactic processing in trained musicians, the left hemisphere becomes dominant again.

From Meaning to Grammar

It was probably semantics that separated speech and song. But how did semantics appear? How did social-bonding vocalizations begin to represent events in the world around us? There are examples of semantic vocal communication in mammals like vervet monkeys, baboons and suricates. Some years ago, Robert Seyfarth, Dorothy Cheney and Peter Marler observed that vervet monkeys make different alarm calls when they spot a leopard, eagle or snake, each eliciting distinct escape behaviors (Seyfarth et al. 1980). For example, after an eagle alarm call, animals on the ground run to a bush, while after a snake alarm call they disperse. This selectivity is learned, as vervet infants often use predator alarms in response to harmless species. When they hear false alarms, adults usually look up but do not respond, but when the alarm is correct they repeat it. This provides feedback to the young that permits them to gradually refine their calls (Seyfarth and Cheney 2003a, b). However, alarm calls are triggered by a reflex pattern and are not necessarily goal-directed behavior in which there is a pursued outcome. On the other hand, symbolic reference depends on learned associative relations between different kinds of stimuli, say visual and auditory. According to Steffen Hage and Andreas Nieder, a precursor of semantic associations can be observed in neurons of the monkey ventrolateral prefrontal cortex, where single neurons associate arbitrary visual signs with numerical stimuli that are presented contingently (Diester and Nieder 2007; Hage and Nieder 2016). This mechanism probably depends on interactions with hippocampal and medial temporal lobe structures.

As I said in Chapter 8, early meanings based on vocal mimicry and gestural pantomimes may have been the first instances of a primitive semantic system. Moreover, a factor that may have been relevant for the evolution of conventionalized meaning is the capacity to share an attentional state between two individuals. Converging attention may be directed to an external object within sight, or more abstractly, to internal meaning like the concept of an absent object, where both subjects generate similar representations of the referred object (Garcia 2014). Thus, the capacity to direct the other's attention is a critical step for the acquisition of meanings. Children use at least two mechanisms to direct the attention of adults to specific objects. One is finger-pointing and gaze direction, and the other is vocal behavior. Vocalizations and other gestures are more generic and do not convey spatial information. In the monkey, vocalizations are associated with an innate circuit that includes the cingulate gyrus in the frontal cortex (described in the next section). This region signals the occurrence of contextually incongruent or unpredicted events, and triggers activation of executive brain systems in the dorsolateral prefrontal cortex that engage in resolving the conflict or incongruence. Thus, the cingulate cortex signals behaviorally relevant events that are transmitted to others by vocalizations (Paus 2001; Roelofs and Hagoort 2002). One possibility is that vocalizations and non-pointing gestures served to attract the other's attention to the respective individual, and then gaze direction and pointing served to direct attention to a given position in space. These steps, I believe, might represent a missing link between predominantly emotional communication and a primitive referential system based on vocal mimicry and pantomimic sounds.

In a further step, imitation of animals or physical events, be they vocal, gestural, or both, may have represented an early instance of symbolic reference (see Chapter 8). But how did we go from simple meanings like those conveyed in vocal imitation and pantomime to more complex meanings involving phrases depicting actions? Hage and Nieder argue that complex behaviors like sequence planning, behavioral sequences and strategy changes are encoded by monkey lateral prefrontal cortex neurons, which may be considered a precursor of syntactic structure (Hage and Nieder 2016, Fujii and Graybiel 2003, Wallis et al. 2001). Likewise, fMRI studies suggest that complex, nested motor plans including speech could be processed by Broca's area (Koechlin and Jubault 2006). On the other hand, the insula and other speech-related areas may rather control the precise timing of complex motor acts.

More into behavior, Bickerton proposes that words began to be combined in different ways to achieve increasingly complex meanings, first in random order (as in Creole languages; see the Chapter 1), but then using short words like prepositions and other elements as links between them (Bickerton 2009). For Bickerton, the lexical properties of words determine the linguistic elements that can be bound to them, and syntax emerges from these binding rules. Others like Otto Jespersen, and Tecumseh Fitch after him have argued that propositional utterances originally consisted of entire sung phrases, which had a kind of holistic or contextual meaning, perhaps more like vervet monkey calls. In a subsequent stage, this continuous message was decomposed into several mobile "chunks", or primitive words that combined with others, which is similar to what Michael Arbib proposed (see Chapter 8) (Jespersen 1922; Fitch 2009, 2010; Arbib 2012). This is compatible with the notion that vocalizations were originally used for play and other apparently non-utilitarian behaviors rather than for directly transmitting relevant information about the environment. Similarly, Simon Townsend and collaborators have argued that syntactically ordered messages appeared early in human vocal evolution, and that a phonologically organized system of words was a late acquisition (Collier et al. 2014). In any of these cases, the learning, processing and combination of words probably put a much heavier load on working memory capacity, and it is tempting to propose that the origin of syntax was associated with a significant amplification of the phonological loop.

My own perspective is that the two kinds of signaling had different uses. The primitive speech of our ancestors may have contained both long structured vocal strings signaling identity, group membership and other social signals; and short primitive word-like utterances and gestures used to call the attention of others and to direct their attention to specific objects or events. Playful behavior and learned emotional signals like vocalizations and gestures may have been important to generate cohesive behavior and generate dyads, not only between mother and child, but also among allies within a group. In addition, the capacity to use word-like utterances like onomatopoeias and conventionalized vocalizations to address simple meanings generated primitive semantics. In this context, Damián Blasi and collaborators found a strong statistical relationship between speech sounds and word meanings, as in /t/ for "tongue" and /n/ for nose, across about two-thirds of the world's languages (Blasi et al. 2016), which is in line with the onomatopoeia hypothesis of speech origins. If I had to guess, I would bet that early *Homo,* and even Australopithecines communicated extensively through learned vocalizations, which developed from a condition similar to the babbling of human infants. This type of vocalization developed a primitive syntactic organization, as in songbirds. From this behavior, a "lexical component" emerged, signaling events that may have been relevant for group behavior, such as recruiting individuals to obtain a distant food source. Another possibility for the emergence of a lexical component is tool-making (see Chapters 4 and 8). These two components, lexical and syntactical, coalesced in later stages, possibly in very recent times, to produce the first forms of speech as we know it. This may have taken place concomitant with the origin of modern humans, and the associated Cultural Revolution that took place at that time. In this line, Vitor Nóbrega and Shigeru Miyagawa proposed that there are two key formal elements in language structure, the expressive and the lexical (Nóbrega and Miyagawa 2015). The expressive system is related to intentionality and is hierarchically organized but limited (finite) in its possible structures, while the lexical system conveys semantic content but does not admit structural organization. There are examples of expressive communication among non-human animals such songbirds, and of lexical communication as in vervet monkeys, but the two never occur together. The two communication domains fused only in human language. We may never know when this happened in human evolution, giving rise to modern speech. It may have been gradual, starting concomitantly with the incipient phonological loop, perhaps already in Australopithecines, but it may have been fully expressed much later, as human culture became increasingly complex. This perspective differs from the notions of protospeech and protosign of Derek Bickerton and Mike Arbib, respectively, in that early human communication took place in two different domains, not just conveying signals about external events, but also a rudimentary phonology in which preverbal strings were used for socializing.

Down from the Cortex

The above is a plausible, but admittedly very speculative scenario of speech origins. Now I will turn to harder evidence of the neural mechanisms involved in vocalizations. The rhythmic organization of vocalizations, together with other orofacial movements like breathing, licking, chewing, and swallowing, depend on brainstem sensorimotor circuits that control the precise coordination of many muscles involved in generating and regulating subtly different movements of the respiratory muscles, nose, lips, tongue and throat. Feeding behaviors like eating, drinking and swallowing require the coordinated action of some 26 muscles and five brainstem motor nuclei to be correctly executed. All these behaviors have to be strictly timed with breathing, which is essential to staying alive. These processes are coordinated by small brainstem circuits called central pattern generators, which in some cases are synchronized by upstream central rhythmic generators. Jaw movements are controlled in the brainstem by the trigeminal motor nucleus; tongue movements by the hypoglossal nucleus, lip and nose movements by the facial nucleus, and finally swallowing and vocalization by the nucleus ambiguus. The close neural relationship in mammals between the control of swallowing and vocalization suggests to many that primitive vocalizations originated from modulations of ingestive behaviors, which is supported by the findings of mirror neurons involved in swallowing that I mentioned in Chapter 8, and may relate to the recently reported "food songs" of gorillas (see above). Likewise, the production of speech requires elaborate neural control of the phonatory system at different levels. For example, we need close regulation of thoracic musculature during exhalation when we speak because air has to be expelled much more slowly and in bursts than during normal respiration. Furthermore, there have to be very rapid inspirations as we run out of air while speaking. This pattern contrasts with the inhalation-exhalation patterns observed during vocalization by non-human primates, who alternate inspirations and expirations much more regularly during long vocalizations, putting a limit on the duration and rate of their calls. Asif Ghazanfar and others have argued that the thoracic vertebral canal, housing the thoracic spinal cord that innervates breathing muscles, is larger in modern humans and Neanderthals than in non-human primates and other fossil hominids (Ghazanfar and Rendell 2008). However, whether larger size indicates increased motor control has not been demonstrated.

Some vocalizations made by non-human primates and other mammals during play are reminiscent of human laughter. Robert Provine has studied laughter in non-human primates and has observed that with chimpanzees, such vocalizations are accompanied by a “play face” that allows us to recognize this as a form of laughter (Provine 2013, 2016). While human laughter sounds like “ha-ha-ha”, chimp laughter sounds like “ah-ah-ah”, and sounds more like panting. This is because humans can parse the exhalation into a sequence of several vowel-like sounds, while chimps produce one laughing sound per respiratory cycle. This is proposed to result from a tight coupling of the respiratory cycle to locomotion that is observed in four-legged animals, and remains in most apes even if they walk bipedally. According to Provine, full bipedality in humans released this constraint, and provided the necessary flexibility of the respiratory system for the emergence of speech. Interestingly, Provine also notes that there are plenty of laughter during normal speech, which normally intercalates with it between phrases, not interfering with syntactical structure. Provine calls this the punctuation effect, which serves to provide emphasis to speech, and probably co-evolved with speech in a similar way to hand gestures and other behaviors. Finally in this line, I have always been intrigued by the variety of laughs different people have. One can quickly tell whether a friend is in a place or not just by hearing his or her laughter. Perhaps laughter partly evolved as an individual signature like what is seen in dolphins and other species, which favored group cohesion.

Descending control for speech originates in different brain centers and establishes the patterns of the activity of brainstem and spinal motor neurons in organized sequences. There are two basic circuits that control vocal behavior in humans. The first and more ancient is shared with monkeys and other mammals, and participates in reflex and emotionally triggered vocalizations like laughing and crying, as well as in monkey vocalizations. Recently reviewed by Gert Holstege and Hari Subramanian, this circuit encompasses the anterior cingulate cortex, orbitofrontal cortex, insula and amygdala, with projections to a brainstem region called the periaqueductal gray (Fig. 10.2). From there, projections reach the reticular formation and are then directed to the ambiguus nucleus (more specifically for vocalizations, the nucleus retroambiguus). This nucleus controls the soft palate, pharynx, larynx and respiratory muscles (Holstege and Subramanian 2016). In turn, these muscles control the air pressure in the respiratory tract, which is needed for proper vocalization. Activation of this circuit may partly explain the unimpaired speech of stutterers when they sing or swear. Perhaps related to these functions is Marco Catani's frontal aslant tract (see Chapter 2), running down from the dorsomedial prefrontal cortex to Broca's area, which is involved in motivational aspects of speech (Catani et al. 2013).

The second descending circuit is associated with Broca's region and related cortical areas that connect with the premotor and motor cortices and the basal ganglia, thalamus and the cerebellum. Originating from the laryngeal cortex in the primary motor cortex, axons reach the brainstem reticular formation, and directly innervate the ambiguus nucleus containing the motor neurons that control the laryngeal muscles for vocalization. The cortical- ambiguus projection is specific to humans among primates, and is thought to participate in learned vocalizations including speech (see Chapter 8), and has been compared to the descending projection to vocal brainstem nuclei in songbirds (Chapter 9). Electrophysiological stimulation of the human larynx area (usually the left) of the ventral motor cortex triggers vocalizations and oral movements, while activation of regions related to Broca's area can elicit more complex, speech-like behavior (see Chapter 2).

In the monkey primary motor cortex there is also a larynx representation that Kristina Simonyan and Uwe Jürgens have exhaustively characterized (Simonyan and Jürgens 2003; Simonyan et al. 2016). Axons from the monkey laryngeal motor cortex reach brainstem reticular neurons, which in turn connect with ambiguus motor neurons. However, some axons are found in the periphery of the nucleus ambiguus, which implies that the purported differences from humans in motor innervation of the nucleus ambiguus are of degree rather than kind. Furthermore, the group of Leonardo Fogassi reported the existence of ventral premotor neurons in the macaque that activate specifically with voluntary vocalizations, and a ventral premotor region that produces simple mouth movements and contraction of laryngeal muscles but not vocalizations when stimulated, while lesions in this region have no effect on spontaneous vocalizations (Coudé et al. 2011). Likewise, Steffen Hage and Andreas Nieder evidenced call-related neurons in the ventrolateral prefrontal cortex that predict preparation for voluntary vocalizations (Hage and Nieder 2013). In addition, Simonyan and collaborators recently reported a substantial increase from

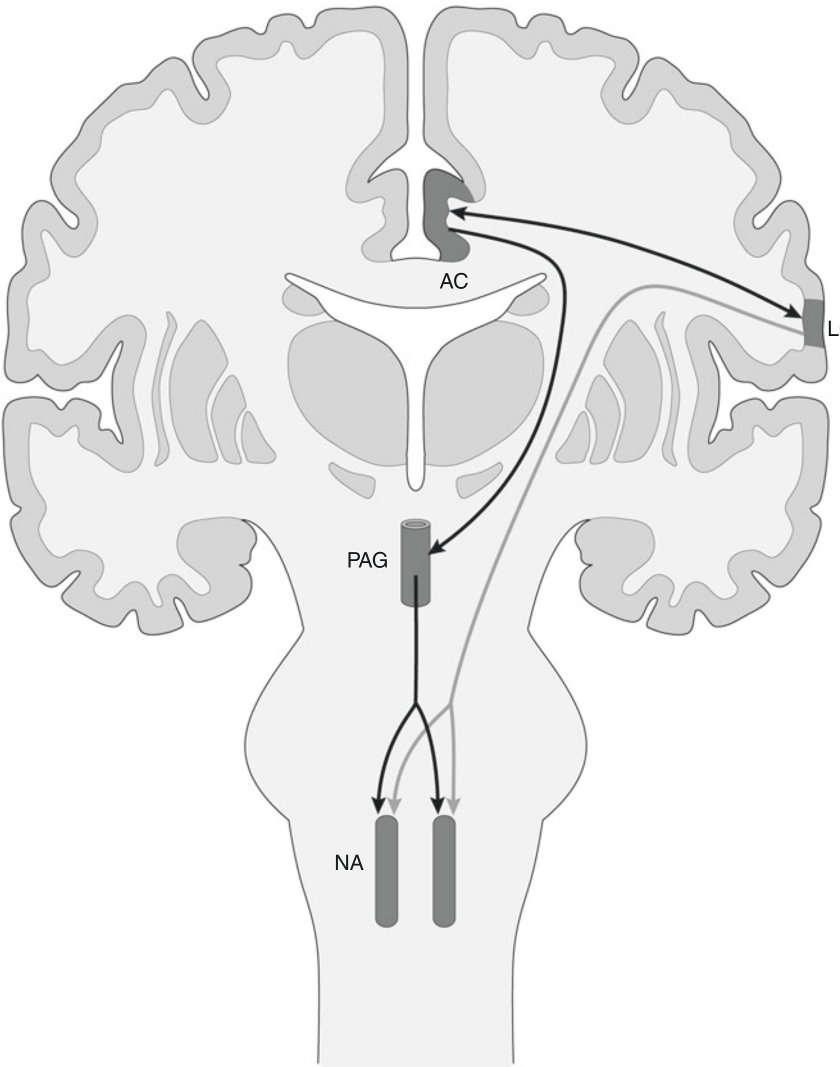

Fig. 10.2 The cortical pathways for vocal control in the human. Black arrows show the pathway that is shared with other primates and mammals. Gray arrows show a tract that is more developed in our species. AC, anterior cingulate; LC, laryngeal cortex; PAG, periaqueductal gray; NA, nucleus accumbens

monkey to human in the connectivity of the laryngeal motor cortex with somatosensory and inferior parietal cortex, which indicates greater sensory and cognitive control of this region (see Chapter 7) (Kumar et al. 2016).

Hage and Nieder have recently summarized this evidence, describing two neural circuits controlling speech (Hage and Nieder 2016). The first is a volitional motor network dependent on the prefrontal cortex, centered in Broca's area, and a phylogenetically older vocal motor network that depends on limbic areas (involved in the initiation of vocalizations) and brainstem systems (controlling the motor output). Hage and Nieder argue that the volitional network is present in monkeys, but its link with the vocal motor network is weak, only allowing to control the initiation of stereotyped vocal output. In human evolution, the volitional network has progressively gained control over the vocal motor network, mainly through inhibitory interactions. This is coupled with two other major innovations, the reinforcement of direct descending projections from the motor cortex to the brainstem, and the amplification of auditory projections to the frontal executive network (contributing to the phonological loop). In summary, rudimentary motor cortical control of the orofacial region already existed in non-human primates, from which voluntary control of speech may have emerged at some point.

Look Who's Talking

Another approach to the study of speech origins is infant vocal and speech development. Kimborough Oller and collaborators have found notable flexibility in early infants' affective vocalizations (squeals, vowel-like sounds, and growls), denoting positive, neutral or negative emotions depending on the context (Oller et al. 2013). These vocalizations contrast with the innately specified cry and laughter, which always denote negative and positive effects, respectively, and can be compared to the vocalizations of non-human primates. Flexibility of the former calls may have been important for subsequent acquisition of vocal learning in early humans, contributing to the emotional meaning of speech.

Consistent with the notion of the origin of the phonological loop as a key innovation in speech origins, the anatomical development of the language network fits the trajectory of speech acquisition in infants. Michael Skeide and Angela Friederici recently summarized the chronology of speech development in infants, correlating it with the maturation of the dorsal and ventral auditory pathways (Skeide and Friederici 2016). Firstly, bottom-up mechanisms in the ventral auditory stream develop in the first year of life, generating the capacity to segment phonological information into distinct word forms, and after 12 months lexical items and lexico-semantic information is already being processed in the superior temporal lobe. By the time the baby is 2 years of age, morphosyntactic processing and grammatical categorization have matured, mostly involving the ventrolateral prefrontal cortex. Finally, at around 4 years of age, the child starts developing top down processing mechanisms through the dorsal pathway, and the neural systems involved in syntactic and semantic processing begin to diverge into distinct but overlapping networks. This process ends when children reach the age of ten, when adult-like syntactic and semantic networks are evident. Friederici's group and others have recently observed that the maturation of the arcuate fasciculus correlates with both the maturation of hemodynamic responses during sentence comprehension, and with the subject's behavioral performance in the task (Skeide et al. 2016; Yeatman et al. 2011). The group led by Ghislaine Dehaene-Lambertz showed that the ventral language pathway matures earlier, while the dorsal pathway (including the arcuate fasciculus) is slower to mature but catches up in the first months after birth, which they suggest relates to the development of combinatorial speech processing (Dubois et al. 2016). Likewise, Pascale Tremblay and collaborators have recently found that the length of the frontal aslant tract in the left hemisphere predicts the receptive language development of 5-to-8-year-old children, a period characterized by improvements in language and communication (Broce et al. 2015).

In addition, an intense research agenda has focused on the mechanisms of speech perception in infants. Nonetheless, these studies have failed to show a language-specific mechanism for the early processing of speech. Take the example of categorical perception described in the Chapters 1 and 8. In early language learning, infants need to weigh the fact that different speakers generate different acoustic patterns for the same linguistic sound, for which they cluster each speaker's sounds into discrete categories, which are normalized for the general population, such as the sound of the letter /a/ is classified as the same regardless of the speaker. However, this capacity has been observed in other species as well. Furthermore, when infants are exposed to language, they are said to commit their brains to the sound patterns that are specific to the particular language they use, while blocking the capacity to distinguish the sound patterns of other languages (Kuhl 2004). Commitment to one language is believed to occur during a critical or sensitive period, where from the 800 or so sounds that are present in all languages, the child becomes sensitive to some 40 elements that are characteristic of the mother's language. Children who acquire an additional language after this critical period develop the typical foreign accent most of us have in our second languages. Again, this process is consistent with general mechanisms of brain development. Giorgio Innocenti has suggested that the processes of axonal pruning and connectional rearrangement that take place during normal development may be particularly relevant during critical periods, including those instances of sensory deprivation during early language acquisition. In this case, the consequences may range from minor ones like the lack of exposure to phonemic boundaries in some languages, to more severe consequences among socially deprived individuals. Innocenti speculates that the process of terminal retraction may be stunted in these cases (Innocenti 2007). A recent report by Jacques Mehler, Marina Nespor and collaborators concluded that there are innate preferences of infants for syllable types. Using a non-invasive technique called near infrared spectroscopy, which can measure brain oxygen supply on the surface of the infant's skull, these authors found that syllable structures that are common in many languages elicited a specific response in language-related regions of the left hemisphere of 2-5-day old infants, as opposed to uncommon syllable structures (Gómez et al. 2014). However, this assumes that newborns have had no exposure to speech before they were born, which may not be totally correct. Babies may start hearing their mother's voice when they are still in the womb.

Once infants recognize the sounds of a language, they begin to break it down into words. Separating the verbal string into its constituent words is a difficult task, as any one knows who has been exposed to an unknown language. The groups led by Jenny Saffran, Patricia Kuhl and Jacques Mehler have made outstanding contributions to this problem, evidencing that infants perform a statistical analysis of the frequencies of syllable transitions during speech, and detect unlikely transitions as candidates for word separations (Saffran et al. 1996; Kuhl 2004; Peña et al. 2002). This capacity requires maintaining the syllable string in the memory. Children can detect word candidates after only two minutes of exposure to continuous syllable strings when they hear an artificial language for the first time consisting of syllable sequence patterns. This process may also contribute to extract structural regularities of the speech string, which can be generalized to novel inputs, contributing to the acquisition of certain grammatical rules. Not only infants, but also children and adults can perform well in tasks requiring statistical learning of syllables. Furthermore, Lucca Bonatti, Marcela Peña, and collaborators have disclosed different roles for consonants and vowels in language learning (Bonatti et al. 2007). As in working memory studies, consonants have been associated with lexical structure, while vowels are linked to grammatical order (see Chapter 6). Manuel Carreiras has further shown that vowels put processing demands on prosody, while as expected, consonants are related to lexical and semantic analysis (Carreiras and Price 2008). The arcuate fasciculus is still immature in infants and Broca's and motor areas are still incapable of programming accurate articulatory mechanics, which makes unlikely that infants are using adult-like verbal working memory circuits to understand speech (Skeide et al. 2016). Nonetheless, the auditory system seems to be sufficiently mature to perform these tasks, and probably contains the critical elements of sensory processing that will be later used in working memory networks. Note however that statistical learning is not exclusive to speech learning and can also be observed in sequences of tones and melodies. It is not exclusively human either as it has been observed in non-human primates and even rats (Toro and Trobalón 2005).

Many regular associations in language are distantly related to each other, with intervening elements located between them. These are the long distance dependencies discussed in Chapter 6, a feature that has been considered a central aspect of language structure. Elissa Newport and Richard Aslin showed that adults had difficulty learning syllable transitions when a pair of critical syllables that are normally together in words are separated by an intermediate syllable (e.g. /ba/-/di/-/te/; /ba/-/ku/-/te/; /ba/-/to/-/te/). However, the same subjects had no problem learning pairs of non-adjacent letters, be they two vowels separated by a variable consonant, or two consonants separated by a variable vowel (Newport and Aslin 2004). It is possible that tracking non-adjacent syllables puts a computationally intractable load on the memory system. In a second study, Newport and Aslin joined the ethologist Marc Hauser and trained tamarin monkeys in these tasks, who were able to learn non-adjacent relationships between vowels but not between consonants (Newport et al. 2004). Like human adults, monkeys were unable to detect distant regularities between syllables, but performed well in the case of letters. Notably, human infants were unable to learn any of these non-adjacent dependencies made by syllables or letters. Tracking non-adjacent dependencies in music is also possible depending on the context. For example, Ansgar Endress showed that adult humans have less difficulty doing these tasks using tonal melodies (as opposed to random melodies), indicating the role of higher-level processes in this type of learning (Endress 2010). Summarizing this evidence, the known perceptual strategies that infants use for recognizing speech, like categorical perception or statistical learning, are not uniquely human or speech-specific, which again reinforces the notion that the perceptual mechanisms involved in speech recognition are domain general processes underlying neural plasticity in many systems.

Infants use not only the complicated strategies described above to recognize and separate words, but also other social signals, like prosody, semantics and pragmatics. For example, mothers and adults in all cultures speak to their infants in what is called baby talk, or motherese, in which adults modulate some aspects of prosody like slowing rhythm, exaggerating pitch differences and stretching vowel sounds, to which infants have been found to benefit in discriminating speech sounds. Infants are not only aroused by their parents' intonation patterns, but also use prosodic cues to extract lexical and grammatical information. Furthermore, it is known that actual face to face and even physical contact with adults are highly important for language learning, as well as for song learning in songbirds. In addition, the pragmatics of language, which settles the behavioral context in which this takes place, is a critical aspect of language development, as it consists of the association of words and sentences with current and previous situations, and with the mother and child's ongoing behavior and intentions. All of this contextual information is required to make reference to external events and associate specific strings with different behavioral situations (Skeide and Friederici 2016). The relative contribution of statistical processing and behavioral cues for learning words remains to be determined, but I would predict that the latter make at least a robust contribution to this process.

Gene Tracks

The last, but in no way the least aspect of speech origins that I will discuss in this chapter is genetics. In Chapter 9, I mentioned the notable parallelism in patterns of gene expression between language-related areas in humans, and regions involved in vocal learning in songbirds (Pfenning et al. 2014). Another approach has been the study of inherited speech disorders. As with other human traits, there are many conditions that affect the development of speech and language, which have been collectively termed as specific language impairment to distinguish them from language disorders that are secondary to other conditions like autism or hearing impairments (see Chapter 6). However, this clinical category is far from homogeneous, there being dysfunctions in expressive language, in both expressive and receptive language, phonological disorders, stuttering and non-specified communication disorders. Furthermore, clinicians usually separate speech disorders like stuttering and phonological disorders from language disorders that are more profound and affect morphology (formation of words), grammar, and in some cases semantics and pragmatics. Despite the attempts to separate and categorize these alleged subgroups, there is a noticeable overlap among these conditions, and few children fit neatly into any of these subgroups (van der Lely and Pinker 2014).

In many cases, genetic factors contribute to susceptibility of language and speech disorders, as these conditions usually run in families, and have shown a strong heritability: monozygotic twins show a stronger concordance for language disorders than dizygotic twins. However, pinpointing genes specifically affecting the development of speech and language is difficult. The first gene to be associated with language, and possibly the one we now know most about is FOXP2, first identified in the analysis what is called the KE family. Some of the members of this family have impaired control of orofacial musculature, resulting in interrupted speech and the incapacity to articulate words fluently and intelligibly, a condition termed verbal dyspraxia. This condition resembles the deficits observed in acquired Broca's aphasia, showing deficits in grammatical competence, especially with past tenses in irregular verbs, and in word formation. Notably, the affected subjects have no deficits in manual praxis and are usually taught sign language to supplement their communication difficulties. It is however not clear if learning highly skilled manual operations like playing musical instruments would be deficitary in these subjects.

In 1995, Fanareh Vargha-Khadem and collaborators published an influential article showing that the inherited disorder of the KE family is not language-specific, including not only deficits in articulated speech and syntactic operations, but also in general orofacial functions and intellectual capacity (verbal and non-verbal IQ) (Vargha-Khadem et al. 1995, 2005). Affected members of the KE family vary in the severity of this condition, some having relatively mild deficits while others show profound intellectual and communicative impairment. Three years later, Vargha-Khadem and collaborators located an alteration in a specific region in chromosome 7 strongly associated with the condition (Fischer et al. 1998). In a subsequent paper in 2001, the gene in chromosome 7 was identified as coding for a regulatory gene called FOXP2 (Lai et al. 2001).

Neuroanatomical observations of the affected family members point to a significant reduction of cerebellar regions and the caudate nucleus (a component of the basal ganglia), and underactivation of the basal ganglia during speech production. Broca's area and related regions connected to the basal ganglia have less gray matter in affected individuals than in normal subjects, and show lower than normal activation during speech. Conversely, in posterior regions, Wernike's area, the superior temporal sulcus and the putamen showed an increased level of gray matter compared to the normal condition (Vargha-Khadem et al. 1998). In normal postmortem subjects, FOXP2 was found to be active in several brain regions, including the cerebral cortex, limbic regions, and especially the basal ganglia and the thalamus (Vargha-Khadem et al. 2005; Fisher 2009). All together, these findings suggest that FOXP2 is important for the development of motor neural circuits involving the cerebral cortex, the basal ganglia and the cerebellum, which participate in learning an planning skilled motor movements. These movements are principally orofacial but could involve manual skills as well. Further research has found another gene, CNTNAP2 that codes for a protein involved in cell adhesion and neuronal recognition, to which the FOXP2 protein binds (Newbury et al. 2010). Variants of this gene have been associated with cases of specific language impairment. In addition, human mutations in FOXP1, a gene closely related to FOXP2, produce language disorders and other symptoms like developmental delay, dropping eyelids and hand and foot contractions. Other speech-related candidate genes have appeared recently, but there is still much research to be done.

Notably, FOXP2 is found in most studied non-human species, many of which are vocal non-learners, and in these the gene expression profile is similar to that in humans. Moreover, the gene structure is highly conserved and is identical among primates like macaques, gorillas and chimpanzees (Enard et al. 2002). Human FOXP2 differs in two amino acids from that of other primates, while the mouse FOXP2 differs from primate FOXP2 in only one amino acid. Recall that mice are vocal learners and develop ultrasonic "songs". Mutations characterizing human FOXP2 are claimed to date from about 200,000 years ago. Notably, DNA sequences from our Neanderthal cousins suggest that they had already acquired these two mutations (Krause et al. 2007). However, the common ancestor of modern humans and Neanderthals is believed to have existed some 300,000 years ago (Johansson 2014). Clearly more studies are needed to settle the chronology of these events in human evolution.

Mice are an excellent animal model for many genetic conditions, as so-called knock-out mutants can be produced in which a specific gene is suppressed, for example, by inserting a non-functional copy of the gene that replaces the normal one in the single cell embryo (Enard et al. 2002, 2009; Hammerschmidt et al. 2015). Several mutants for FOXP2 have been produced, which show a general developmental delay and usually die 3 or 4 weeks after birth, possibly because of respiratory impairment. Since FOXP2 mutants live for a short time, research has focused on the largely innate ultrasonic vocalizations of young pups, which they usually emit when separated from their mother. The evidence from different mutants indicates that inactivation of FOXP2, or associated genes like SRPX2, results in the absence or reduction of isolation calls and other mild behavioral deficits, but in more stringent conditions they do emit these vocalizations, with a generally conserved acoustic structure. Reviewing this evidence, Simon Fisher and Constance Scharff concluded that the evidence does not yet support a parallel between the development of mouse vocalizations and that of human speech, nor in the role of FOXP2 in vocalizations of the two species (Fisher 2009). However, new findings using heterozygous FOXP2 mutant mice, who live until adulthood, have found that in these animals the songs are altered in syllable structure and rhythmicity, and develop abnormally in relation to wild type mice (Castelucci 2016; Chabout et al. 2016). Other interesting animals in which to assess FOXP2 functions are songbirds, which express this gene in the corpus striatum, cerebellum and other brain regions. FOXP2 is upregulated in the striatal area X during song learning and is higher when adult males sing to females than when they sing alone (White et al. 2006). In another study, the group of Constance Scharff showed that after interfering with FOXP2 gene expression by means of a directed viral infection in area X of the zebrafinch, young birds were impaired in their song learning abilities, copying some notes but not others from the exposed songs (Haesler et al. 2007). Furthermore, even imitated notes were not accurate, which was interpreted as a deficit in sensorimotor coordination, FOXP2 mediating the proper synaptic activity of two main neurotransmitters in the basal ganglia: glutamate and dopamine.

It has been proposed that FOXP2 participates in the development of the neural networks that make up the frontal cortex, corpus striatum and cerebellum, involved in sensorimotor integration and particularly in motor skill learning (Schreiweis et al. 2014; Groszer et al. 2008). This evidence in part has prompted authors like Constance Scharff and Johan Bolhuis to suggest that FOXP2 was recruited in several instances (songbirds, humans and possibly bats) to assist vocal learning capacities (Bolhuis et al. 2010; Scharff and Petri 2011). This may be another case of "deep homology" (in my opinion, a better term is genetic homology), like the one I described in Chapter 9, where the gene Pax6 was recruited for the development of eyes in widely different animal lineages, presumably because it originally served some role in the specification of phototransducing receptors.

Beyond Genetics

I have argued throughout this book that the acquisition of the phonological loop was a key evolutionary innovation, in which an auditory-motor interface provided by the development of the dorsal pathway supported vocal learning. These anatomical features are likely to have been associated with increasing brain size, showing a gradation from early *Homo*, with a brain volume of some 500 cc., to the large 1,500 cc. brains of modern humans that appear about 300 thousand years ago. Nonetheless, since we achieved our large brains, a very long time passed before the critical transition some 40,000 to 50,000 years ago that marked the emergence of more sophisticated cultures. Richard Wrangham and others attribute this cultural explosion to the process of self-domestication, in which we adapted ourselves to live in community by inhibiting aggressive behavior and other traits (I will go further on this in the next chapter) (Wrangham 2003, Hare et al. 2012). As I said above, modern speech combining the lexico-semantic and the phonological-syntactic components could have appeared slowly, from an initial stage in which phonological sequences were used for social bonding and lexical-semantic items were used for contingent behavioral coordination. These two domains may have been progressively mixed, until a stage at which it triggered major cultural innovations. Some authors, like Michael Arbib, consider that the acquisition of modern speech was very recent and a largely cultural event, with few genetic changes, in which individuals learned new behaviors, first through imitation of others and subsequently by instruction (Chapter 8). This may be partly correct, as we are born to learn from others, an evolutionary tendency that probably began before we acquired speech. Nevertheless, were there genetic changes associated with this transition that were favored by natural selection (see the next chapter). Paradoxically, apart from FOXP2 and a couple other genes, there has not been a great deal of success in finding a genetic blueprint for language evolution. Of course one possibility is that we still have not found the relevant genes for this development and further research into the genetic foundations of language is certainly worth pursuing.

Another emerging possibility is the epigenetic action of proteins like histones and other chemicals like methyl groups and the small RNAs that surround the DNA molecule and participate in gene expression regulation. Epigenetic processes are normally involved in cell differentiation during development, in which many genes are repressed while others are activated, generating different cellular phenotypes. There has been an explosion of studies analyzing the effects of epigenetic modifications, some of which have been contentiously reported to be transmitted across generations. Michael Skinner is one of the defenders of this process as an important evolutionary mechanism (Skinner 2014). Despite all the criticism and controversy his work has received, Skinner claims that this may be one of the biggest scientific paradigm shifts of the century. Epigenetic changes in development have been linked to several clinical conditions, particularly stress response and developmental disorders like autism and attention deficit hyperactivity disorder.

Epigenetic processes may have also been involved in human brain evolution. Working with Svante Pääbo and a team of other researchers, Liran Carmel mapped the methylation patterns of modern humans and compared these with genomes extracted from ancient human, Neanderthal and Denisovan fossils (Gokhman et al. 2014). They used a sophisticated computational approach to infer the original methylation pattern in the highly distorted DNA of ancient humans. Although this is an indirect measure, it has provided intriguing results that still need to be confirmed by other methods. The genetic difference between modern and ancient humans is minimal (fewer than 100 genes), but they found important differences in DNA methylation patterns, particularly in disease-related genes, and about one third of these were in genes associated with neuropsychiatric disorders. On the other hand, another recent study shows that regions in the genome that underwent most accelerated evolution in the human lineage are associated with risk of cognitive and social disorders like autism (Doan et al. 2016). Thus, it is possible that both, rapid genetic evolution and epigenetic modifications have contributed to the origin of the human mind and its diseases.

This evidence points to the intriguing possibility that epigenetic mechanisms had a profound effect on the evolution of language-related circuits. Is the acquisition of the phonological loop related to such changes? This remains an enigmatic question that fortunately can be addressed by future studies. Evidence indicates that epigenetic mechanisms were involved in the domestication of the silkworm and the chicken, as reported by the group of Per Jensen (Jensen 2014). If the proposal of Wrangham is correct and our species experienced a domestication-like process in the last 50,000 years, it may be that epigenetics played a role in the evolution of our sociality, and perhaps in the acquisition of critical traits like the phonological loop (Wrangham 2003). In the last chapter of this book, I will refer to the contextual and social circumstances in which speech may have emerged, as it is not an isolated achievement but is interwoven with several other behavioral innovations.

References and Notes

Arbib MA (2012) How the Brain Got Language. The Mirror System Hypothesis. Oxford University Press, Oxford

Andics A, Gábor A, Gácsi M, Faragó T, Szabó D, Miklósi Á (2016) Neural mechanisms for lexical processing in dogs. Science 353:1030–1032

Arriaga G, Jarvis ED (2013) Mouse vocal communication system: are ultrasounds learned or innate? Brain Lang 124:96–116

Arriaga G, Zhou EP, Jarvis ED (2012) Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds. PLoS One 7:e46610

Arnold K, Zuberbühler K (2012) Call combinations in monkeys: compositional or idiomatic expressions? Brain Lang 120:303–309

Bezerra BM, Souto Ada S, de Oliveira MA, Halsey LG (2009) Vocalisations of wild common marmosets are influenced by diurnal and ontogenetic factors. Primates 50:231–237

Bickerton D (2009) Adam's Tongue. How Humans Made Language, How Language Made Humans. Hill and Wang, New York

Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH (2016) Sound-meaning association biases across thousands of languages. Proc Natl Acad Sci U S A 113:10818–10823

Boë LJ, Berthommier F, Legou T, Captier G, Kemp C, Sawallis TR, Becker Y, Rey A, Fagot J (2017) Evidence of a Vocalic Proto-System in the Baboon (Papio papio) Suggests Pre-Hominin Speech Precursors. PLoS One 12(1): e0169321

Bolhuis JJ, Okanoya K, Scharff C (2010) Twitter evolution: converging mechanisms in birdsong and human speech. Nat Rev Neurosci 11:747–759

Bonatti LL, Peña M, Nespor M, Mehler J (2007) On consonants, vowels, chickens, and eggs. Psychol Sci 8:924–925

Bradbury JW, Emmons LH (1974) Social organization of some trinidad bats I. Emballonuridae. Z Tierpsychol 36:137–183

Bradley CE, McClung MR (2015) Vocal divergence and discrimination of long calls in tamarins: A comparison of allopatric populations of Saguinus fuscicollis nigrifrons and S. f. lagonotus. Am J Primatol 77:679–687

Bregman MR, Patel AD, Gentner TQ (2016) Songbirds use spectral shape, not pitch, for sound pattern recognition. Proc Natl Acad Sci U S A 113:1666–1671

Broce I, Bernal B, Altman N, Tremblay P, Dick AS (2015) Fiber tracking of the frontal aslant tract and subcomponents of the arcuate fasciculus in 5-8-year-olds: Relation to speech and language function. Brain Lang 149:66–76

Brockelman WY, Schilling D (1984) Inheritance of stereotyped gibbon calls. Nature 312:634–636

Candiotti A, Zuberbühler K, Lemasson A (2012) Context-related call combinations in female Diana monkeys. Anim Cogn 15:327–339

Carreiras M, Price CJ (2008) Brain activation for consonants and vowels. Cereb Cortex 18:1727–1735

Caspermeyer J (2014) For bats and dolphins, hearing gene prestin adapted for echolocation. Mol Biol Evol 31:2552

Castellucci GA, McGinley MJ, McCormick DA (2016) Knockout of Foxp2 disrupts vocal development in mice. Sci Rep 6:23305.

Catani M, Mesulam MM, Jakobsen E, Malik F, Martersteck A, Wieneke C, Thompson CK, Thiebaut de Schotten M, Dell'Acqua F, Weintraub S, Rogalski E (2013) A novel frontal pathway underlies verbal fluency in primary progressive aphasia. Brain 136:2619–2628

Chabout J, Sarkar A, Patel SR, Radden T, Dunson DB, Fisher SE, Jarvis ED (2016) A FOXP2 mutation implicated in human speech deficits alters sequencing vocalizations in adult male mice. Front Behav Neurosci 10:197

Chait M, Greenberg S, Arai T, Simon JZ, Poeppel D (2015) Multi-time resolution analysis of speech: evidence from psychophysics. Front Neurosci 9:214

Clarke E, Reichard UH, Zuberbühler K (2006) The syntax and meaning of wild gibbon songs. PLoS One 1:e73

Collier K, Bickel B, van Schaik CP, Manser MB, Townsend SW (2014) Language evolution: syntax before phonology? Proc Biol Sci 281:20140263

Cook P, Rouse A, Wilson M, Reichmuth C (2013) A California sea lion (Zalophus californianus) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. J Comp Psychol 127:412–427

Coquerelle M, Prados-Frutos JC, Rojo R, Mitteroecker P, Bastir M (2013) Short faces, big tongues: developmental origin of the human chin. PLoS One 8:e81287

Coudé G, Ferrari PF, Rodà F, Maranesi M, Borelli E, Veroni V, Monti F, Rozzi S, Fogassi L (2011) Neurons controlling voluntary vocalization in the macaque ventral premotor cortex. PLoS One 6:e26822

D'Anastasio R, Wroe S, Tuniz C, Mancini L, Cesana DT, Dreossi D, Ravichandiran M, Attard M, Parr WC, Agur A, Capasso L (2013) Micro-biomechanics of the Kebara 2 hyoid and its implications for speech in Neanderthals. PLoS One 8:e82261

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164

Diester I. Nieder A (2007) Semantic associations between signs and numerical categories in the prefrontal cortex. PLoS Biol 5:e294

Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S, Gascon GG; Homozygosity Mapping Consortium for Autism., Nieto M, Walsh CA (2016) Mutations in Human

Accelerated Regions Disrupt Cognition and Social Behavior. Cell 167:341–354

Dubois J, Poupon C, Thirion B, Simonnet H, Kulikova S, Leroy F, Hertz-Pannier L, Dehaene-Lambertz G (2016) Exploring the Early Organization and Maturation of Linguistic Pathways in the Human Infant Brain. Cereb Cortex 26:2283–2298

Egnor SE, Iguina CG, Hauser MD (2006) Perturbation of auditory feedback causes systematic perturbation in vocal structure in adult cotton-top tamarins. J Exp Biol 209:3652–3663

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872

Enard W, Gehre S, Hammerschmidt K, Hölter SM, Blass T, Somel M, Brückner MK, Schreiweis C, Winter C, Sohr R, Becker L, Wiebe V, Nickel B, Giger T, Müller U, Groszer M, Adler T, Aguilar A, Bolle I, Calzada-Wack J, Dalke C, Ehrhardt N, Favor J, Fuchs H, Gailus-Durner V, Hans W, Hölzlwimmer G, Javaheri A, Kalaydjiev S, Kallnik M, Kling E, Kunder S, Mossbrugger I, Naton B, Racz I, Rathkolb B, Rozman J, Schrewe A, Busch DH, Graw J, Ivandic B, Klingenspor M, Klopstock T, Ollert M, Quintanilla-Martinez L, Schulz H, Wolf E, Wurst W, Zimmer A, Fisher SE, Morgenstern R, Arendt T, de Angelis MH, Fischer J, Schwarz J, Pääbo S (2009) A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137:961–971

Endress AD (2010) Learning melodies from non-adjacent tones. Acta Psychol 135:182–190

Feenders G, Liedvogel M, Rivas M, Zapka M, Horita H, Hara E, Wada K, Mouritsen H, Jarvis ED (2008) Molecular mapping of movement-associated areas in the avian brain: a motor theory for vocal learning origin. PLoS One 3:e1768

Fischer J, Hammerschmidt K (2011) Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication. Genes Brain Behav 10:17–27

Fischer J, Wheeler BC, Higham JP (2015) Is there any evidence for vocal learning in chimpanzee food calls? Curr Biol 25:R1028–1029

Fisher SE, Scharff C (2009) FOXP2 as a molecular window into speech and language. Trends Genet 25:166–177

Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME (1998) Localisation of a gene implicated in a severe speech and language disorder. Nat Genet 18:168–170

Fitch WT (2009) Musical protolanguage: Darwin's theory of language evolution revisited. http://languagelog.ldc.upenn.edu/nll/?p=1136

Fitch WT (2010) The Evolution of Language. Cambridge University Press, Cambridge.

Fitch WT, Reby D (2001) The descended larynx is not uniquely human. Proc Biol Sci 268:1669–1675

Fujii N, Graybiel AM (2003) Representation of action sequence boundaries by macaque prefrontal cortical neurons. Science 301:1246–1249

Gamba M, Torti V, Estienne V, Randrianarison RM, Valente D, Rovara P, Bonadonna G, Friard O, Giacoma C (2016) The Indris Have Got Rhythm! Timing and Pitch Variation of a Primate Song Examined between Sexes and Age Classes. Front Neurosci 10:249

García RR, Zamorano F, Aboitiz F (2014) From imitation to meaning: circuit plasticity and the acquisition of a conventionalized semantics. Front Hum Neurosci 8:605

Geissmann T (2002) Duet-splitting and the evolution of gibbon songs. Biol Rev Camb Philos Soc 77:57–76

Ghazanfar AA, Rendall D (2008) Evolution of human vocal production. Curr Biol 18:R457–R460

Ghazanfar AA, Takahashi DY (2014a) The evolution of speech: vision, rhythm, cooperation. Trends Cogn Sci 18:543–553

Ghazanfar AA, Takahashi DY (2014b) Facial expressions and the evolution of the speech rhythm. J Cogn Neurosci 26:1196–1207

Ghazanfar AA, Takahashi DY, Mathur N, Fitch WT (2012) Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Curr Biol 22:1176–1182

Glushko A, Steinhauer K, DePriest J, Koelsch S (2016) Neurophysiological Correlates of Musical and Prosodic Phrasing: Shared Processing Mechanisms and Effects of Musical Expertise. PLoS One 11:e0155300

Gokhman D, Lavi E, Prüfer K, Fraga MF, Riancho JA, Kelso J, Pääbo S, Meshorer E, Carmel L (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. Science 344:523–527

Gómez DM, Berent I, Benavides-Varela S, Bion RA, Cattarossi L, Nespor M, Mehler J (2014) Language universals at birth. Proc Natl Acad Sci U S A 111:5837–5841

Grimault S, Lefebvre C, Vachon F, Peretz I, Zatorre R, Robitaille N, Jolicoeur P (2009) Load-dependent brain activity related to acoustic short-term memory for pitch: magnetoencephalography and fMRI. Ann N Y Acad Sci 1169:273–277

Groszer M, Keays DA, Deacon RM, de Bono JP, Prasad-Mulcare S, Gaub S, Baum MG, French CA, Nicod J, Coventry JA, Enard W, Fray M, Brown SD, Nolan PM, Pääbo S, Channon KM, Costa RM, Eilers J, Ehret G, Rawlins JN, Fisher SE (2008) Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. Curr Biol 18:354–362

Gustison ML, Semple S, Ferrer-I-Cancho R, Bergman TJ (2016) Gelada vocal sequences follow Menzerath's linguistic law. Proc Natl Acad Sci U S A 113: E2750–E2758

Haesler S, Rochefort C, Georgi B, Licznerski P, Osten P, Scharff C (2007) Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. PLoS Biol 5:e321

Hage SR, Nieder A (2013) Single neurons in monkey prefrontal cortex encode volitional initiation of vocalizations. Nat Commun 4:2409

Hage SR, Nieder A (2016) Dual Neural Network Model for the Evolution of Speech and Language. Trends Neurosci 39:813–829

Hage SR, Gavrilov N, Nieder A (2013) Cognitive control of distinct vocalizations in rhesus monkeys. J Cogn Neurosci 25:1692–1701

Hage SR, Gavrilov N, Nieder A (2016) Developmental changes of cognitive vocal control in monkeys. J Exp Biol 219:1744–1749

Hammerschmidt K, Radyushkin K, Ehrenreich H, Fischer J (2009) Female mice respond to male ultrasonic "songs" with approach behaviour. Biol Lett 5:589–592

Hammerschmidt K, Reisinger E, Westekemper K, Ehrenreich L, Strenzke N, Fischer J (2012) Mice do not require auditory input for the normal development of their ultrasonic vocalizations. BMC Neurosci 13:40

Hammerschmidt K, Schreiweis C, Minge C, Pääbo S, Fischer J, Enard W (2015) A humanized version of Foxp2 does not affect ultrasonic vocalization in adult mice. Genes Brain Behav 14:583–590

Hare B, Wobber V, Wrangham R (2012) The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim Behav 83:573–585

Hausen M, Torppa R, Salmela VR, Vainio M, Särkämö T (2013) Music and speech prosody: a common rhythm. Front Psychol 4:566

Hauser M (1996) The Evolution of Communication. MIT Press, Cambridge

Hauser MD (1998) Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. Anim Behav 55:1647–1658

Hayes KJ, Hayes C (1951) The intellectual development of a home-raised chimpanzee. Proc Am Philos Soc 95:105–109

Hewitt G, MacLarnon A, Jones KE (2002) The functions of laryngeal air sacs in primates: a new hypothesis. Folia Primatol 73:70–94

Holstege G, Subramanian HH (2016) Two different motor systems are needed to generate human speech. J Comp Neurol 524:1558–1577

Innocenti GM (2007) Subcortical regulation of cortical development: some effects of early, selective deprivations. Prog Brain Res 164:23–37

Janik VM (2009) Whale song. Curr Biol 19:R109–R111

Janik VM (2013) Cognitive skills in bottlenose dolphin communication. Trends Cogn Sci 17:157–159

Janik VM (2014) Cetacean vocal learning and communication. Curr Opin Neurobiol 28:60–65

Janik VM, Sayigh LS (2013) Communication in bottlenose dolphins: 50 years of signature whistle research. J Comp Physiol A Neuroethol Sens Neural Behav Physiol 199:479–489

Jensen P. (2014) Behavior genetics and the domestication of animals. Annu Rev Anim Biosci 2:85–104

Jespersen O (1922) Language: Its Nature, Development and Origin. W W Norton & Company, New York

Johansson S (2014) Neanderthals did speak, but FOXP2 doesn't prove it. Behav Brain Sci 37:558–559

Kaminski J, Call J, Fischer J (2004) Word learning in a domestic dog: Evidence for "fast mapping". Science 304:1682–1683

King SL, Janik VM (2013) Bottlenose dolphins can use learned vocal labels to address each other. Proc Natl Acad Sci U S A 110:13216–13221

King SL, McGregor PK (2016) Vocal matching: the what, the why and the how. Biol Lett 12:20160666

King SL, Guarino E, Keaton L, Erb L, Jaakkola K (2016) Maternal signature whistle use aids mother-calf reunions in a bottlenose dolphin, Tursiops truncatus. Behav Processes 126:64–70

Knörnschild M, Nagy M, Metz M, Mayer F, von Helversen O (2010) Complex vocal imitation during ontogeny in a bat. Biol Lett 6:156–159

Koda H, Lemasson A, Oyakawa C, Rizaldi, Pamungkas J, Masataka N (2013) Possible role of mother-daughter vocal interactions on the development of species-specific song in gibbons. PLoS One 8:e71432

Koechlin E, Jubault T (2006) Broca's area and the hierar-chical organization of human behavior. Neuron 50:963–974

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Pääbo S (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17:1908–1912

Kuhl PK (2004) Early language acquisition: cracking the speech code. Nat Rev Neurosci 5:831–843

Kumar V, Croxson PL, Simonyan K (2016) Structural organization of the laryngeal motor cortical network and its implication for evolution of speech production. J Neurosci 36:4170–4181

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A fork-head-domain gene is mutated in a severe speech and language disorder. Nature 413:519–523

Lalamn A, Boesch C (2013) Do wild chimpanzees have functionally referential food calls? Folia Primatol 84:290

Lameira AR, Hardus ME, Kowalsky B, de Vries H, Spruijt BM, Sterck EH, Shumaker RW, Wich SA (2013) Orangutan (Pongo spp.) whistling and implications for the emergence of an open-ended call repertoire: a replication and extension. J Acoust Soc Am 134:2326–2335

Lameira AR, Maddieson I, Zuberbühler K (2014) Primate feedstock for the evolution of consonants. Trends Cogn Sci 18:60–62

Lameira AR, Hardus ME, Bartlett AM, Shumaker RW, Wich SA, Menken SB (2015) Speech-like rhythm in a voiced and voiceless orangutan call. PLoS One 10:e116136

Lameira AR, Hardus ME, Mielke A, Wich SA, Shumaker RW (2016) Vocal fold control beyond the species-specific repertoire in an orangutan. Sci Rep 6:30315

Lieberman P (1968) Primate vocalizations and human linguistic ability. J Acoust Soc Am 44:1574–1584

Lieberman P (1979) Hominid evolution, supralaryngeal vocal tract physiology, and the fossil evidence for reconstructions. Brain Lang 7:101–126

Lieberman P (1984) The Biology and Evolution of Language. Harvard Univeristy Press, Harvard

Liu F, Jiang C, Wang B, Xu Y, Patel AD (2015) A music perception disorder (congenital amusia) influences speech comprehension. Neuropsychologia 66:111–118

Luef EM, Breuer T, Pika S (2016) Food-associated calling in Gorillas (Gorilla g. gorilla) in the Wild. PLoS One 11:e0144197

Miller CT, Flusberg S, Hauser MD (2003) Interruptibility of long call production in tamarins: implications for vocal control. J Exp Biol 206:2629–2639

Morell V. (2014) When the bat sings. Science 344:1334–1337

Morecraft RJ, Stilwell-Morecraft KS, Solon-Cline KM, Ge J, Darling WG (2014) Cortical innervation of the hypoglossal nucleus in the non-human primate (Macaca mulatta). J Comp Neurol 522:3456–3484

Morrill RJ, Paukner A, Ferrari PF, Ghazanfar AA (2012) Monkey lipsmacking develops like the human speech rhythm. Dev Sci 15:557–568

Newbury DF, Fisher SE, Monaco AP (2010) Recent advances in the genetics of language impairment. Genome Med 2:6

Newman JD, Symmes D (1974) Vocal pathology in socially deprived monkeys. Dev Psychobiol 17:351–358

Newport EL, Aslin RN (2004) Learning at a distance I. Statistical learning of non-adjacent dependencies. Cogn Psychol 48:127–162

Newport EL, Hauser MD, Spaepen G, Aslin RN (2004) Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. Cogn Psychol 49:85–117

Nóbrega VA, Miyagawa S (2015) The precedence of syntax in the rapid emergence of human language in evolution as defined by the integration hypothesis. Front Psychol 6:271

Oller DK, Buder EH, Ramsdell HL, Warlaumont AS, Chorna L, Bakeman R (2013) Functional flexibility of infant vocalization and the emergence of language. Proc Natl Acad Sci U S A 110:6318–6323

Owren MJ, Dieter JA, Seyfarth RM, Cheney DL (1993) Vocalizations of rhesus (Macaca mulatta) and Japanese (M. fuscata) macaques cross-fostered between species show evidence of only limited modification. Dev Psychobiol 26:389–406

Patel AD (2008) Music, Language and the Brain. Oxford University Press, Oxford

Patel AD, Iversen JR, Bregman MR, Schulz I (2009) Studying synchronization to a musical beat in nonhuman animals. Ann N Y Acad Sci 1169:459–469

Paus T (2001) Primate anterior cingulate cortex: where motor control, drive and cognition interface. Nat Rev Neurosci 2:417–424

Peña M, Bonatti LL, Nespor M, Mehler J (2002) Signal-driven computations in speech processing. Science 298:604–607

Peña M, Langus A, Gutiérrez C, Huepe-Artigas D, Nespor M (2016) Rhythm on your lips. Front. Psychol 7:1708

Peretz I, Zatorre RJ (2005) Brain organization for music processing. Annu Rev Psychol 56:89–114

Perlman M, Clark N (2015) Learned vocal and breathing behavior in an enculturated gorilla. Anim Cogn 18:1165–1179

Petkov CI, Jarvis ED (2012) Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. Front Evol Neurosci 4:12

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, Mouncastle J, Moseley MA, Thompson JW, Soderblom EJ, Iriki A, Kato M, Gilbert MT, Zhang G, Bakken T, Bongaarts A, Bernard A, Lein E, Mello CV, Hartemink AJ, Jarvis ED (2014) Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346:1256846

Pisanski K, Cartei V, McGettigan C, Raine J, Reby D (2016) Voice modulation: a window into the origins of human vocal control? Trends Cogn Sci 20:304–318

Price M (2016) Why monkeys can't talk—and what they would sound like if they could. Science Magazine online, Dec. 9. http://www.sciencemag.org/news/2016/12/why-monkeys-can-t-talk-and-what-they-would-sound-if-they-could

Provine RR (2013) Laughing, grooming, and pub science. Trends Cogn Sci 17:9–10

Provine RR (2016) Laughter as a scientific problem: an adventure in sidewalk neuroscience. J Comp Neurol 524:1532–1539

Ravignani A, Fitch WT, Hanke FD, Heinrich T, Hurgitsch B, Kotz SA, Scharff C, Stoeger AS, de Boer B (2016) What pinnipeds have to say about human speech, music, and the evolution of rhythm. Front Neurosci 10:274

Reichmuth C, Casey C (2014) Vocal learning in seals, sea lions, and walruses. Curr Opin Neurobiol 28:66–71

Ridgway S, Carder D, Jeffries M, Todd M (2012) Spontaneous human speech mimicry by a cetacean. Curr Biol 22:R860–R861

Roelofs A, Hagoort P (2002) Control of language use: cognitive modeling of the hemodynamics of Stroop task performance. Cogn Brain Res 15:85–97

Rouse AA, Cook PF, Large EW, Reichmuth C (2016) Beat keeping in a Sea Lion as coupled oscillation: implications for comparative understanding of human rhythm. Front Neurosci 10:257

Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. Science 274:1926–1928

Scharff C, Petri J (2011) Evo-devo, deep homology and FoxP2: implications for the evolution of speech and language. Philos Trans R Soc Lond B Biol Sci 366:2124–2140

Schreiweis C, Bornschein U, Burguière E, Kerimoglu C, Schreiter S, Dannemann M, Goyal S, Rea E, French CA, Puliyadi R, Groszer M, Fisher SE, Mundry R, Winter C, Hevers W, Pääbo S, Enard W, Graybiel AM (2014) Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. Proc Natl Acad Sci U S A 111:14253–14258

Seyfarth RM, Cheney DL (2003a) Signalers and receivers in animal communication. Annu Rev Psychol 54:145–173

Seyfarth RM, Cheney DL (2003b) Meaning and emotion in animal vocalizations. Ann N Y Acad Sci 1000:32–55

Seyfarth RM, Cheney DL (2010) Production, usage, and comprehension in animal vocalizations. Brain Lang 115:92–100

Seyfarth RM, Cheney DL, Marler P (1980) Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. Science 210:801–803

Simonyan K, Jürgens U (2003) Efferent subcortical projections of the laryngeal motorcortex in the rhesus monkey. Brain Res 974:43–59

Simonyan K, Ackermann H, Chang EF, Greenlee JD (2016) New developments in understanding the complexity of human speech production. J Neurosci 36:11440–11448

Skeide MA, Friederici AD (2016) The ontogeny of the cortical language network. Nat Rev Neurosci 17:323–332

Skeide MA, Brauer J, Friederici AD (2016) Brain functional and structural predictors of language performance. Cereb Cortex 26:2127–2139

Skinner MK (2014) Environment, epigenetics and reproduction. Mol Cell Endocrinol 398:1–3

Springer SP, Deutsch G (1981) Left Brain, Right Brain. WH Freeman & Company, New York

Stoeger AS, Mietchen D, Oh S, de Silva S, Herbst CT, Kwon S, Fitch WT (2012) An Asian elephant imitates human speech. Curr Biol 22:2144–2148

Suga N (1989) Principles of auditory information-processing derived from neuroethology. J Exp Biol 146:277–286

Takahashi DY, Narayanan DZ, Ghazanfar AA (2013) Coupled oscillator dynamics of vocal turn-taking in monkeys. Curr Biol 23:2162–2168

Takahashi DY, Fenley AR, Teramoto Y, Narayanan DZ, Borjon JI, Holmes P, Ghazanfar AA (2015) The developmental dynamics of marmoset monkey vocal production. Science 349:734–738

Takahashi DY, Fenley AR, Ghazanfar AA (2016) Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys. Philos Trans R Soc Lond B Biol Sci 371:1693

Talmage-Riggs G, Winter P, Ploog D, Mayer W (1972) Effect of deafening on the vocal behavior of the squirrel monkey (Saimiri sciureus). Folia Primatol 17:404–420

Thinh VN, Hallam C, Roos C, Hammerschmidt K (2011) Concordance between vocal and genetic diversity in crested gibbons. BMC Evol Biol 11:36

Toro JM, Trobalón JB (2005) Statistical computations over a speech stream in a rodent. Percept Psychophys 67:867–875

van der Lely HK, Pinker S (2014) The biological basis of language: insight from developmental grammatical impairments. Trends Cogn Sci 18:586–595

Vargha-Khadem F, Watkins K, Alcock K, Fletcher P, Passingham R (1995) Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. Proc Natl Acad Sci U S A 92:930–933

Vargha-Khadem F, Watkins KE, Price CJ, Ashburner J, Alcock KJ, Connelly A, Frackowiak RS, Friston KJ, Pembrey ME, Mishkin M, Gadian DG, Passingham RE (1998) Neural basis of an inherited speech and language disorder. Proc Natl Acad Sci U S A 95:12695–12700

Vargha-Khadem F, Gadian DG, Copp A, Mishkin M (2005) FOXP2 and the neuroanatomy of speech and language. Nat Rev Neurosci 6:131–138

Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. Nature 411:953–956

Washington SD, Tillinghast JS (2015) Conjugating time and frequency: hemispheric specialization, acoustic uncertainty, and the mustached bat. Front Neurosci 9:143

Watson SK, Townsend SW, Schel AM, Wilke C, Wallace EK, Cheng L, West V, Slocombe KE (2015) Vocal learning in the functionally referential food grunts of chimpanzees. Curr Biol 25:495–499

Wheeler BC, Fischer J (2012) Functionally referential signals: a promising paradigm whose time has passed. Evol Anthropol 21:195–205

White SA, Fisher SE, Geschwind DH, Scharff C, Holy TE (2006) Singing mice, songbirds, and more: models for FOXP2 function and dysfunction in human speech and language. J Neurosci 26:10376–10379

Wich SA, Swartz KB, Hardus ME, Lameira AR, Stromberg E, Shumaker RW (2009) A case of spontaneous acquisition of a human sound by an orangutan. Primates 50:56–64

Wrangham R (2003) The Evolution of Cooking. In: Brockman J (ed), The New Humanists: Science at the Edge. Sterling Publishing, New Dehli, p 108–109

Yeatman JD, Dougherty RF, Rykhlevskaia E, Sherbondy AJ, Deutsch GK, Wandell BA, Ben-Shachar M (2011) Anatomical properties of the arcuate fasciculus predict phonological and reading skills in children. J Cogn Neurosci 23:3304–3317

Zatorre RJ (2003) Music and the brain. Ann N Y Acad Sci 999:4–14

Zatorre RJ, Salimpoor VN (2013) From perception to pleasure: music and its neural substrates. Proc Natl Acad Sci U S A 110 (Suppl 2):10430–10437