ANNUAL REVIEWS

Annual Review of Linguistics

The Biology and Evolution of Speech: A Comparative Analysis

W. Tecumseh Fitch

Department of Cognitive Biology, University of Vienna, Vienna 1090, Austria

ANNUAL REVIEWS **Further**

Click here to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

Annu. Rev. Linguist. 2018. 4:255–79

The *Annual Review of Linguistics* is online at linguist.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-011817-045748

Copyright © 2018 by W. Tecumseh Fitch. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 (CC-BY) International License, which permits unrestricted use, distribution, and reproduction in any medium and any derivative work is made available under the same, similar, or a compatible license. See credit lines of images or other third-party material in this article for license information.

OPEN ACCESS CC BY-SA

Keywords

human evolution, speech perception, speech production, evolution of language, monosynaptic connections, paleo-DNA

Abstract

I analyze the biological underpinnings of human speech from a comparative perspective. By first identifying mechanisms that are evolutionarily derived relative to other primates, we obtain members of the faculty of language, derived components (FLD). Understanding when and why these evolved is central to understanding the evolution of speech. There is little evidence for human-specific mechanisms in auditory perception, and the hypothesis that speech perception is "special" is poorly supported by comparative data. Regarding speech production, human peripheral vocal anatomy includes several derived characteristics (permanently descended larynx, loss of air sacs), but their importance has been overestimated. In contrast, the central neural mechanisms underlying speech production involve crucial derived characteristics (direct monosynaptic connections from motor cortex to laryngeal motor neurons, derived intracortical dorsal circuitry between auditory and motor regions). Paleo-DNA from fossil hominins provides an exciting new opportunity to determine when these derived speech production mechanisms arose during evolution.

1. INTRODUCTION

Faculty of language, broad sense (FLB): the set of all mechanisms involved in acquiring, processing, producing, and/or perceiving language

Faculty of language, derived components (FLD): the subset of FLB mechanisms acquired subsequent to our evolutionary divergence from chimpanzees

Faculty of language, narrow sense (FLN): the subset of FLB mechanisms that are both unique to humans, and specific to language within human cognition

Speech, as the preferred output modality for human language, is an unusual feature of our species that depends upon a complex but well-understood set of mechanisms, including vocal/motor, auditory/perceptual, and central neural mechanisms. The capacity for speech clearly differentiates humans from other primates, indicating that some of these mechanisms have diverged, in recent human evolution, from those of our prelinguistic ancestors. However, the capacity to vocally imitate sounds is not uniquely human and is shared with a surprisingly diverse group of organisms, including many bird species, most marine mammals, elephants, and some bats. This means that this capacity has evolved, convergently, many times in vertebrate evolution. These facts conspire to make the evolution of speech better suited for comparative evolutionary analyses than most other components of human language.

Although scientific attention to the evolution of speech predates Darwin, the last two decades have witnessed fundamental progress, on the basis of both an improved understanding of the vocal and auditory periphery (particularly in nonhuman animals) and advances in our understanding of central processing (due largely to various noninvasive imaging methods). These advances place the evolution of speech on a firm empirical foundation, and have led to a slow reorientation of attention away from the periphery toward central neural control mechanisms. This reorientation implies, unfortunately, that cues to peripheral anatomy gleaned from fossils tell us much less than traditionally hoped about the evolution of speech. However, the silver lining is that DNA recovered from fossil hominins can potentially be used to determine the evolutionary timing of novel human capacities including neural circuitry. Over the coming decades this development promises to revolutionize our understanding not only of speech but also of the biology and evolution of language more generally.

In this review, I analyze the evolution of speech from a comparative and phylogenetic perspective, using comparisons with other living animals to draw inferences about the abilities of extinct organisms. My focus is on characteristics where the data are solid enough to warrant clear conclusions. The picture painted by these comparisons may be surprising to many readers, as it indicates that most of the mechanisms involved in human speech—the mechanisms comprising part of the faculty of language in a broad sense (FLB)—have very deep evolutionary roots. This suggests that the evolution of speech in humans required only a few changes, most of them neural, resulting in a short list of derived characteristics that should be the focus of future evolutionary research.

I term these few crucial changes the "faculty of language, derived components," or FLD. This term differs from the faculty of language in the narrow sense (FLN), defined previously by my colleagues and me (Fitch et al. 2005, Hauser et al. 2002), which excluded traits shared with any other species (e.g., with birds—thus excluding vocal learning) or with other cognitive domains (e.g., music). I now consider these criteria too stringent regarding speech, and propose that the more biologically meaningful set consists of those language-relevant mechanisms that are derived relative to our nonlinguistic primate ancestors, irrespective of whether these are language specific or not: the FLD.

My focus in this review is on the biological evolution of the human capacity for speech, not the cultural history of linguistic word forms investigated by comparative and historical linguists. The comparative method as practiced in modern biology has many similarities with the methods of comparative linguistics familiar to most linguist readers, but it also differs in important ways. By adopting the cladistic terminology of modern phylogenetics, I try to keep these differences clear.

1.1. The Comparative Method in Biology

Comparative biological analysis starts with a breakdown of a complex trait (such as vision or language) into multiple subcomponent mechanisms or traits, which are then arrayed on a phylogenetic tree. To avoid circularity, the phylogenetic tree is derived from independent data, today using genomic data. When this approach is taken with cognitive traits it can be termed "cognitive phylogenetics" (Fitch et al. 2010).

Examination of patterns of similarity and difference relative to this phylogenetic tree allows us to distinguish examples of homology, in which a trait is shared by multiple species due to inheritance from a common ancestor, from analogy, in which the trait was derived independently in different clades (akin to distinguishing cognates from accidental similarity or borrowing in historical linguistics). Homologies can be used, like cognates, to reconstruct ancestral states: The more broadly shared a homologous trait is, the further in the past the ancestral form must have existed. Sets of homologous traits thus provide the equivalent of a time machine allowing us to reconstruct an evolutionary sequence of ancestral forms.

Analogies (convergently evolved traits) play a different role. When there are multiple instances of convergent evolution, analogies provide independent samples to test evolutionary hypotheses, with each instance providing an independent data point. Throughout this review I emphasize the importance of distinguishing derived from shared characteristics, and homology from analogy, when analyzing comparative data.

1.2. Shared and Derived Traits

The two types of homologous traits most relevant in the context of this review are shared ancestral states (plesiomorphies) and derived traits that typify a species or group of species (autapomorphies) (**Figure 1*a***). A shared ancestral or "primitive" trait is termed a plesiomorphy, and a derived trait, in general, is termed an apomorphy. A less useful term, homoplasy, denotes all nonhomologous traits, lumping together traits evolved via convergence, parallelism, and reversals; I avoid it and use the more precise terms analogy and convergence.

Of particular interest are those derived homologous traits shared by an entire clade, termed synapomorphies. These indicate the presence of that derived trait in the last common ancestor (LCA) of that clade. For example, lactation, hair, and three middle ear bones are synapomorphies of living mammals, allowing us to infer that these traits were all present in the ancestral mammal. Also crucial are autapomorphies: derived characters specific to a species or clade, for example, fully bipedal locomotion in humans relative to other apes. It is this class of derived human traits relevant to speech—speech-related autapomorphies—that are my focus here, because these make up crucial components of the FLD that evolved since our divergence from other apes.

Note that the interpretation of these cladistics terms—plesiomorphy, synapomorphy, and autapomorphy—is context dependent, and depends upon the specific clades under discussion. Thus, lactation is plesiomorphic for humans as mammals (a shared ancestral trait) but synapomorphic for mammals as vertebrates (a derived trait in this broader context).

A group of species related by common descent is called a clade, a generic term that makes no commitments about what traditional phylogenetic level (e.g., genus, family, order, or phylum) the grouping represents. The names of particular clades thus constitute further important terminology. Many of these clade names, such as vertebrate, mammal, or bird, will be familiar; others, like amniote or tetrapod, may not. Given my focus on speech, I anthropocentrically discuss only those clades most relevant to human phylogeny (**Figure 1*b***), starting with the most inclusive clades and becoming increasingly specific.

Homologous traits: characteristics shared in a set of species due to inheritance from the common ancestor of those species

Analogous traits: characteristics acquired independently in two lineages by convergent evolution

Plesiomorphy: a characteristic or mechanism retaining the ancestral "primitive" state, typically shared by most members of a clade

Autapomorphy: an apomorphic trait uniquely defining a particular clade within some specific context (e.g., in humans relative to other apes)

Apomorphy: a characteristic or mechanism displaying a derived state, specific to some clade with a larger plesiomorphic context

Synapomorphy: an apomorphic trait shared throughout a clade, which defines that clade relative to a larger phylogenetic context

Last common ancestor (LCA): the most recent ancestor shared by two specific clades (e.g., apes and humans, or Neanderthals and modern humans)

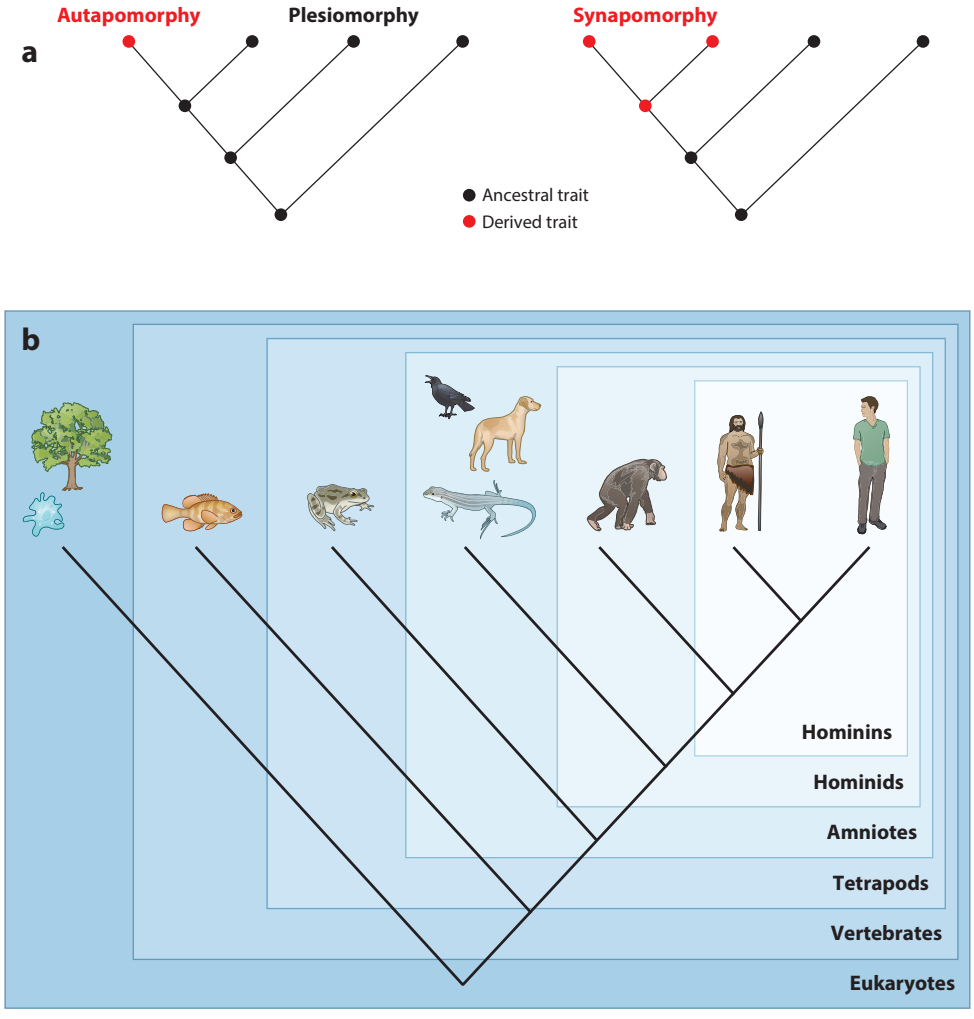


Figure 1

Human cladistics in a nutshell. (*a*) Graphic depiction of key cladistic terms. Plesiomorphies are ancestral or "primitive" traits, shared by most members of a clade, whereas autapomorphies are derived features that are unique to a particular species or clade (*left*). Synapomorphies are derived traits that are shared by a number of living species; their existence supports the inference that the trait was present in the last common ancestor of those species (*right*). (*b*) Illustration of some clades crucial for understanding human evolution (some important but well-known clades, including mammals and primates, are omitted to save space).

1.3. Clades Relevant to the Evolution of Speech

The broadest relevant clade for us is the eukaryotes, which all have cells with a nucleus and mitochondria. Multicellular eukaryote species include plants, animals, and fungi. Most of the basic intracellular biochemical machinery of our bodies, from the genetic code to sugar metabolism or cell division, is shared with this vast clade. Among the animals, a large and important clade is the Bilateria, a broad grouping that includes us vertebrates along with arthropods (e.g., crabs, flies) and mollusks (e.g., squid, snails). The last few decades have revealed surprisingly deep shared traits

among bilaterians; for example, the same genes often serve the same developmental functions in humans and fruit flies.

My main comparative focus here is on the vertebrates, in particular the mostly terrestrial vertebrates termed tetrapods (mammals, birds, amphibians, and the diverse group traditionally called reptiles, more correctly termed nonavian reptiles). Humans share a significant proportion of our basic machinery of hearing and vocal production with other tetrapods (e.g., frogs or alligators). Within tetrapods, the amniotes are the vertebrates that thoroughly conquered land by evolving eggs that can develop on dry land. Amniotes include mammals, birds, and "reptiles." Mammals are distinguished from other amniotes by fur and lactation; a six-layered neocortex; three middle ear bones; and, for all mammals except the egg-laying platypuses and echidnas, internal fetal development.

Within mammals, humans are, of course, primates. Within the primates, we form with our closest relatives, the lesser apes (gibbons and siamangs) and the great apes (chimpanzees, bonobos, gorillas, and orangutans), the hominoids. Traditionally, taxonomists split the apes into two groups, the hominids (humans and our extinct ancestors and relatives such as Neanderthals) and the pongids (all other apes). But the revelation that we are more closely related to chimpanzees than chimpanzees are to orangutans rendered the term "pongid" phylogenetically incorrect, and most paleoanthropologists today use "hominin" to denote humans and our extinct non-ape relatives (**Figure 1*b***), and "hominid" to include the great apes.

Hominid: in current systematic parlance, the clade including humans and the other great apes (previously termed "hominoid")

Hominin: in current systematic parlance, the clade comprising humans and all of their extinct non-ape ancestors and relatives (previously "hominid")

Exaptation: the evolutionary process by which mechanisms change their function, before they are adaptively tuned to their new function

Deep homology: a situation in which a trait, despite being phenotypically convergent, is based on homologous genetic and/or developmental causes

2. THE AUDITORY PERIPHERY

The capacity to sense airborne pressure signals—hearing—has a truly ancient pedigree that can be reconstructed using the comparative method and nicely illustrates several recurring themes in human evolution. One is exaptation: the evolutionary reuse of existing mechanisms for new purposes (e.g., transforming gills into jawbones or ear ossicles). Another is deep homology, in which homologous genetic mechanisms are used to build convergently evolved structures (Fitch 2009a, Shubin et al. 2009).

At the core of vertebrate audition are the sensory hair cells in the inner ear, specialized to convert sonic energy into neural signals. These are homologous to mechanosensory cells that are found in all animals with a nervous system, including cnidarians like jellyfish (Fritzsch et al. 2006). Because the same cell type is used for hearing in insects, via the antennae, hearing via hair cells is shared with a vast array of bilaterian species. In all vertebrates, hair cells are also found within the fluid-filled cavities of the vestibular system, which senses head orientation and rotation, and in most aquatic vertebrates hair cells also detect water movements in the lateral line system on the body surface. Specialized hearing organs have been convergently derived from the vestibular system in many fish species and all tetrapods (Manley 2000).

Among amniotes, the fluid-filled cochlea is the primary organ for sensing airborne sound and develops as an outpouching of the vestibular organ (Fritzsch et al. 2006). Compared with other amniotes, mammals are hearing specialists, sensitive to a much wider frequency range extending well above the 10-kHz limit typical of birds and reptiles. This limit is surpassed via the elongation of the mammalian cochlea, which allows more hair cells to be packed within this organ. To accommodate its increased length, the mammalian cochlea coils into a snail shape different from the bean-shaped cochlea of other amniotes. The length of the basilar membrane is the main determinant of hearing sensitivity and range in mammals, and is greatest in echolocating bats (Manley 2000).

The mammalian cochlea is also unusual in possessing two types of hair cells: the standard mechanosensory inner hair cells, shared with birds or lizards, and a novel class termed outer hair cells, which act as tiny motors to actively tune and amplify vibrations of the basilar membrane

within the cochlea. This mammalian synapomorphy plays an important role in auditory sensitivity and acuity.

Transcription factor: a gene whose protein product binds to DNA to regulate the expression of other genes

The gene regulatory network underlying differentiation of the vertebrate auditory periphery centers on *atonal1*, a transcription factor first discovered in fruit flies. Transcription factors generate a protein that binds to DNA elsewhere in the genome, thus regulating the expression of other genes. Transcription factors thus act as "molecular switches" that activate or suppress expression of other genes. The vertebrate homolog of *atonal1* is called atonal homolog 1, or *Atoh1* (Fritzsch et al. 2006). *Atoh1* is the master control gene for audition, and its expression leads to differentiation of hair cells, cochlea, and much of the brainstem circuitry involved in relaying auditory signals to cortex. The *atonal1* gene plays a similar role in fly hearing, and indeed hearing in deaf mutant mice can be restored by inserting the homologous fruit fly gene (Wang et al. 2002). Because hearing evolved independently in insects and humans, this is an excellent example of deep homology.

Hair cells convert vibrations in the fluid-filled cochlea to a neural signal transmitted by sensory neurons in the cochlea, via the auditory nerve, to the cochlear nucleus in the brainstem. These auditory signals then pass through a series of further specialized brainstem and thalamic nuclei up to the auditory cortex. *Atoh1* also plays a developmental role in all of these auditory nuclei, and this basic layout of the entire auditory system, from cochlea to cortex, is plesiomorphic to all mammals.

In contrast to the inner ear, the middle and outer ears have independent evolutionary histories in different tetrapod clades, and both the tympanum (eardrum) and middle ear ossicles evolved independently in mammals and other tetrapods (Lombard & Bolt 1979). The emergence of tetrapods onto land posed the physical problem of converting weak pressure signals in air into fluid movement in the cochlea: The impedance mismatch between air and liquid means that almost all energy is reflected away from, rather than into, the fluid-filled inner ear. To compensate, most tetrapods possess a light membrane, the tympanum, which connects via one or more bones to the cochlea and mechanically amplifies airborne vibrations. Intriguingly, this system appears to have evolved independently at least three times in tetrapods—in frogs, in mammals, and in birds and other reptiles—but in all cases, components of the jaw have been miniaturized and repurposed as ear ossicles (Tucker 2017). This provides an unusual example of convergent exaptation of the same precursor for the same novel function. In nonmammals a single bone called the columella serves this purpose, whereas in mammals three different jawbones were enlisted (the malleus, incus, and stapes).

Although three ossicles are synapomorphic in mammals, middle ear bone morphology varies considerably among species, and there are fine differences between human ossicles and those of chimpanzees or fossil hominins, which have led to attempts to use ossicle anatomy to reconstruct the hearing abilities of extinct hominins (e.g., Martínez et al. 2013). Based on apparent differences in chimpanzee and human audition, the argument is that human hearing has become specialized for speech sensitivity—a putative auditory autapomorphy. There are two bases for skepticism about these attempts, however. First, as mentioned above, the primary determinant of frequency sensitivity range is the length of the basilar membrane and the cochlea, not middle ear bone morphology (Hemilä et al. 1995, Manley 2000). Second, existing data on chimpanzee hearing are sparse and contradictory. Because some chimpanzees show a W-shaped audiogram, it has been argued that the chevron shape of the normal human audiogram is somehow special, requiring special ear anatomy (Elder 1934, Kojima 1990). But, in fact, most mammals show chevron-shaped audiograms, shifted up or down in frequency (Heffner 2004), and it is the chimpanzee audiogram that is peculiar among mammals, not that of humans. Although it remains possible that this represents a derived trait in chimpanzees, noise-induced hearing loss in humans also leads to a W-shaped audiogram, with a divot around 3 kHz, strikingly similar to that observed in some chimpanzees. This putative specialization may thus simply reflect noise-induced hearing loss in some subjects (the reverberant concrete housing environment of many captive chimpanzees,

combined with their frequent loud vocalizations, produces a truly deafening environment). More data, particularly from young chimpanzees housed in more natural and acoustically friendly conditions, would be needed to resolve this issue.

In summary, the comparative data demonstrate that the human auditory periphery has an ancient pedigree, sharing its sensory elements with jellyfish, genetic determinants with flies, inner ear sense organ with all other tetrapods, and virtually all of its complex sensory and brainstem hardware with other mammals. There is nothing about the human ear that is strikingly different from that of other primates and no convincing evidence of any autapomorphic human characteristics in the auditory periphery. Thus, our peripheral hearing apparatus was in place, in our primate ancestors, in essentially modern form long before we evolved the capacity for speech. Our search for human auditory autapomorphies must therefore turn to central processing mechanisms and the cortex.

3. A BEHAVIORAL INTERLUDE: HOW SPECIAL IS SPEECH PERCEPTION?

A long tradition in psychology holds that speech perception is "special"—both distinct from other auditory perception ("modular") and biologically specialized in our species. This belief dates back to early attempts at Haskins Laboratories to create reading machines for the blind, where phonemes were represented by mechanical sounds like buzzes or bells instead of speech sounds. These attempts failed utterly—such sounds merge into an indiscriminable auditory soup when played at rates as fast as those of speech. This early work led to the conclusion that the speech signal, rather than a sequential stream of phoneme-sized signals, involves a complex encoding of motor gestures (e.g., opening the lips or tapping the tongue against the palate) into the acoustic signal (Liberman & Mattingly 1985). Humans are able to somehow decode this complex auditory signal back into motor commands, as demonstrated by our ability to correctly repeat novel words on first attempt. This capacity to solve this complex decoding problem is central to both speech perception and production.

The creation of electronic speech synthesis at Haskins overcame some of the problems encountered in the earlier attempts, and also opened the door to controlled experiments on speech perception (Liberman 1957). One of the first results was the discovery of categorical perception (CP) of speech—the finding that the discriminability of pairs of speech sounds chosen from a continuum varies nonlinearly, with best discrimination observed across category boundaries, rather than varying as a linear function of physical similarity. This result suggested that our perceptual system is specialized to fit the details of speech production, and led to the hypothesis that "speech is special"—both modular to language and human specific (Liberman 1957, Liberman & Mattingly 1985).

Both aspects of this claim were soon questioned. First, Cutting & Rosner (1974) showed that certain musical sounds were also perceived categorically, although these results were later called into question (Cutting 1982, Rosen & Howell 1981). More convincing were demonstrations by Kuhl & Miller (1975) that animals including rhesus macaques and chinchillas also perceive human speech sounds categorically, and parallel findings that a variety of species perceive their own conspecifics' vocalizations categorically as well (Fischer 2006, Nelson & Marler 1989). Moreover, CP is not universal in speech; although stop consonants are perceived quite categorically, vowel perception is more continuous (Kuhl et al. 1992)—an effect also observed in other animals (Kluender et al. 1998). It now seems clear that CP is not special to speech, and more likely reflects a style of processing available in both auditory and visual domains and to a wide range of species (Harnad 1987). Combined with parallel research showing robust perception of vowels and consonants in a wide variety of birds and mammals (Baru 1975, Dooling 1992, Hienz et al. 2004), the comparative

data suggest instead that the perceptual apparatus underlying speech is simply the human version of a general voice-perception system shared with other species (see Section 4, below).

The demise of the special status of CP did not, however, eliminate enthusiasm for the "speech is special" hypothesis, and new phenomena arose as supporting evidence. One was lateralization—it has long been known that there is a left-hemisphere bias for both production and perception of speech, initially thought to be uniquely human (Geschwind 1970). Early research demonstrating lateralization in birdsong production called this assumption into question (Nottebohm 1971), and since then abundant data have clarified that perceptual lateralization is a typical feature of vertebrates, from fish to birds to primates (Rogers & Andrew 2002, Vallortigara & Rogers 2005). Although a precise characterization of this functional asymmetry remains elusive, one compelling hypothesis is for a right-hemisphere (left-eye) bias in perceiving conspecifics that would constitute a vertebrate plesiomorphy. Unfortunately, the comparative data are clearest in the visual domain, and hemispheric asymmetry in the auditory domain (typically measured via head-turn responses) remains less clear and may vary between species (Fischer et al. 2009, Gil-da-Costa & Hauser 2006). Nonetheless, the hypothesis that perceptual lateralization is unique to humans is no longer tenable.

Four other oft-cited potentially "special" speech-perceptual phenomena are sine-wave speech, duplex perception, the McGurk effect, and vocal tract normalization. Sine-wave speech replaces complex, multiharmonic formants with single sine waves varying in time—akin to substituting an ink-sketch caricature for a color photo. Most adults perceive these sounds as bizarre but intelligible speech (Remez et al. 1981), suggesting another human-specific perceptual specialization. However, recent perceptual experiments with a chimpanzee exposed intensively to human speech showed that she, too, immediately perceived sine-wave analogs as intelligible words, without training (Heimbauer et al. 2011). This finding suggests that, at least given appropriate developmental exposure, the capacity to parse sine-wave speech is shared with other apes.

Duplex perception is a laboratory phenomenon in which the formant transitions corresponding to a particular stop consonant are played to one ear while the vocalic "base" is played to the other. Under these circumstances, subjects report hearing both a fused normal stop consonant and a high-frequency "chirp" corresponding to the transition (Liberman et al. 1981). Like CP, this phenomenon was initially proposed as an indicator of an independent phonetic module, but later research showed that duplex perception occurs both for musical stimuli (chords) and for environmental sounds (slamming doors), suggesting that this, too, reflects more general auditory processing (Fowler & Rosenblum 1990, Pastore et al. 1983). To my knowledge, no experiments investigating duplex perception in animals have been conducted.

The McGurk effect (McGurk & MacDonald 1976) is a fascinating interaction between auditory and visual perception. When a video image of a mouth saying /g/ is played synchronously with playback of the sound /b/, what is perceived is /d/—a sound intermediate in articulation. I know of no study investigating this specific effect in animals, but recent research on audiovisual interactions in macaques suggests that monkeys also spontaneously link the auditory and visual components of conspecific calls, preferentially looking at video displays whose mouth shape matches a played call (Ghazanfar et al. 2005). Thus, integration of audiovocal information in vocal perception is not unique to humans.

A final phenomenon once thought to represent a human autapomorphy is vocal tract normalization—the temporary adjustment of speech-perceptual processes to reflect the vocal characteristics of the current speaker. This is crucial because the different vocal tract dimensions of men, women, and children mean that the "same" speech sounds have quite different absolute formant and fundamental frequencies depending on the speaker. Again, however, abundant comparative data suggest that normalization is not uniquely human. All mammals tested to date perceive formants in conspecific calls, and in many cases use them as cues to the body

size of the caller (Charlton et al. 2012, Fitch & Fritz 2006, Ghazanfar et al. 2007, Reby et al. 2005), suggesting that formant-based normalization is plesiomorphic to mammals. Furthermore, recent data indicate that zebra finches trained on a subtle vowel discrimination with men's voices can generalize without training to a woman's vowels (Ohms et al. 2009). Again, human speech perception seems to reflect broadly shared amniote perceptual mechanisms for conspecific voice perception rather than specializations for speech.

In summary, 50 years of comparative speech research reveals a broadly shared set of perceptual mechanisms that, although potentially evolved for conspecific voice perception in amniotes, are in no sense unique to human speech. Some phenomena such as categorical perception may represent domain-general processing mechanisms, whereas others such as vocal tract normalization may be specific to vocal sounds. Certain perceptual phenomena deemed special to speech, such as duplex perception, remain uninvestigated in other species (Repp 1982), and some rather subtle differences in speech perception have been found between humans and other primates (Sinnott & Adams 1987, Sinnott & Williamson 1999). But it seems clear that most of the phenomena initially suggested as human-specific mechanisms turn out, upon comparative investigation, to be broadly shared among primates, mammals, or amniotes.

4. CENTRAL CORTICAL PROCESSING OF VOCALIZATIONS

I now turn to the central cortical processing of vocal signals. Here, our general knowledge for humans greatly outstrips that for nonhuman animals, except at the level of direct neuronal recording, where nonhuman research dominates. This discrepancy makes firm conclusions difficult and highlights the need for more work comparing different species using the same methods (Andics et al. 2016, Petkov et al. 2008).

In general, the basic layout of auditory cortex in the temporal lobe is the same across mammals (where the best-studied species are cats and various bat and monkey species; Clarey et al. 1992, Rauschecker & Tian 2000, Suga 1990). In primates, the primary auditory cortex is locally surrounded by concentric belt and para-belt regions, which then project elsewhere in cortex via two routes. The ventral route, through the anterior temporal lobe and thence to the prefrontal cortex, appears to play a role in sound identification. The functional role of the dorsal route, which travels back through the posterior temporal cortex and then up through the parietal cortex, was originally considered a "where" pathway (Tian et al. 2001), by analogy to the "what/where" pathway in vision (Goodale & Milner 1992), because the same sound played back from different locations activates different cells. However, in humans this dorsal pathway appears closely linked to motor representations for speech production—more of a "how" function—and some data suggest a similar role in self-monitoring in monkeys (Eliades & Wang 2008). This dorsal route is discussed further below, in connection with vocal control.

Although the auditory regions discussed above are responsive to all sounds, there is abundant evidence for voice-specific neural circuitry in multiple species (Belin 2006) in roughly comparable cortical areas. Here, a distinction between voice-specific and speech-specific regions is important, because speech represents only one class of human vocal sounds, and speech contains abundant nonlinguistic information concerning speaker sex, age, size, and identity. Belin et al. (2000) contrasted blocks of human vocalizations including speech, laughs, cries, and so forth with blocks containing nonvocal sounds. They found consistent preferential activation of the anterior superior temporal sulcus (STS), bilaterally but with a bias toward the right hemisphere (Belin et al. 2000). A later study showed that these activations were specific to human vocalizations, and not elicited by cats' or other species' vocalizations (Fecteau et al. 2004), suggesting that voice-specific circuitry is also species specific.

Similar circuitry exists in macaques. Perrodin et al. (2011) used functional magnetic resonance imaging (fMRI) to isolate voice-specific regions in rhesus macaques, and then used electrodes to record from individual neurons in this region. They found a considerable proportion of cells that responded much more to macaque voices than to environmental sounds or nonconspecific vocalizations, localized mainly in the anterior superior temporal plane, just above the STS. Similar studies in other primate species also reveal cells and regions apparently specialized for conspecific vocalizations (Poremba et al. 2004, Wang & Kadia 2001). It thus seems likely that human voice-specific regions have direct homologs in other primates, and had already evolved in our LCA with other primates.

Voice-selective regions, in both humans and other primates, appear to be biased toward the right hemisphere, whereas speech phonetic processing exhibits a left-hemisphere bias (Zatorre et al. 1992). In animals we typically have little understanding of what would constitute "phonetic," meaning-changing cues in different vocalization types. One exception is a classic series of studies examining call variants in Japanese macaques, *Macaca fuscata*. Two types of "smooth" coos are distinguished by whether their peak frequency is early or late in the call, and Japanese macaques are more sensitive to this relatively fine distinction than are other macaque species (Zoloth et al. 1979) and show a right-ear (left-hemisphere) advantage for it that is absent in other species (Petersen et al. 1984). To verify that this ear advantage truly reflects a cerebral asymmetry, Heffner & Heffner (1986) made unilateral lesions to the right versus left auditory cortex, finding that only after left lesions was the early/late distinction impaired. Thus, in this case of a relatively fine "phonetic" distinction, macaque vocal perception also appears to be left-biased as for humans with speech.

In summary, the comparative neural data available regarding central auditory processing of vocalizations paint a similar picture to the behavioral and peripheral data. The human auditory system shares its fundamental structure and functional organization with that of other mammals. Regarding voice-specific processing, we have solid data only for primates, but again, voice-selective regions exist, suggesting that such regions constitute a primate synapomorphy at least. One recent study on dogs suggests that the existence of voice-selective regions may extend more broadly among mammals (Andics et al. 2014).

Despite a long tradition of argument holding that speech perception is special to humans, putative specializations have been repeatedly rejected by comparative data, and there is little empirical evidence supporting this hypothesis (despite continuing rear-guard defense; e.g., Trout 2003). Instead, the data lead to the conclusion that the primate auditory system had already evolved to a "speech-ready" level of sophistication long before spoken language evolved in our species, and that modern human speech perception relies on processing circuitry that in most relevant essentials is shared with other primates, or with mammals more widely.

5. PERIPHERAL COMPONENTS OF SPEECH PRODUCTION

Despite excellent speech-perceptual abilities (Heimbauer et al. 2011, Savage-Rumbaugh et al. 1993) and good manual imitation abilities, apes are unable to learn to produce novel vocalizations based on an auditory model, and no nonhuman primate has ever learned to speak (Kellogg 1968, Yerkes & Yerkes 1929). Even with intensive training, a human-reared chimpanzee was able to produce only three poorly intelligible approximations of human words (Hayes 1951). This contrasts sharply with the vocal learning capacity of humans (and some other nonprimate species). What are the key determinants of this evident human autapomorphic ability?

Although humans possess several derived traits relevant to vocal production, the basic anatomy and physiology of the human larynx and vocal tract are shared with other mammals. It has become clear in the last two decades that the essential principles of the source-filter theory of speech

production apply to vertebrate vocal production as well (Fant 1960, Taylor & Reby 2010). Vocalizations represent a "source" signal, typically generated in the larynx, that is subsequently filtered by multiple vocal tract formants. Furthermore, the myoelastic aerodynamic theory of the laryngeal source (Titze 2006, van den Berg 1958) applies to laryngeal mechanics in most other mammals (e.g., Herbst et al. 2012) and also to the avian syrinx (Elemans et al. 2014), despite this avian source organ representing a novel synapomorphy in birds. Thus, the essential physical principles of speech production familiar to phoneticians apply equally well to most other tetrapods (cf. Suthers et al. 2016).

Beginning with the source, the larynx represents a tetrapod synapomorphy, with both its basic anatomy and neural control shared among terrestrial vertebrates. However, the mammalian larynx is unique in possessing a thyroid cartilage and a thyroarytenoid muscle contained within the vocal folds. These mammalian synapomorphies give mammals a greater degree of control over vocal fold dynamics than other tetrapods. Similarly, the possession of a fleshy tongue, lips, and a velum is a mammalian synapomorphy. Also shared among mammals are the details of innervation and the location of the brainstem motor nuclei controlling the larynx and articulators (Wall & Smith 2001). Thus, the ground plan of the human vocal tract was already laid out, in detail, in the LCA of living placental mammals some 70 Mya, and we should consider any potential human autapomorphies relative to this shared context (cf. Fitch 2010).

Source: the initial sound-producing organ, in which (silent) air flow is converted into sound; typically the larynx in mammals, including humans. The source determines the pitch of a sound

Filter: the air in the passageways downstream of the source (in the throat, mouth, and nose), which vibrates at preferred frequencies, termed formants, that are emphasized in the final output signal

The best-known human autapomorphy is our lowered larynx. In most mammals, and human newborns, the larynx is positioned in the back of the mouth directly beneath the velar port, enabling them to raise the larynx into the nasal cavity, creating a sealed respiratory pathway from nostrils to lungs. In the first year of life, the human larynx begins to retract from this plesiomorphic position, and by adulthood we can no longer create such a seal (Sasaki et al. 1977). Thus, whereas newborn mammals can breathe nasally while swallowing milk orally, human adults cannot breathe and swallow simultaneously without choking. It is often stated that adult humans are unique in this respect, but this appears to be a myth: Many vertebrates, including fish, birds, and mammals, can choke to death on food (e.g., Rao et al. 1984, Skead 1980), and mammals must typically separate the larynx from the velum when swallowing a bolus of solid food (Thexton & Crompton 1998), putting them in danger of choking. I know of no empirical data indicating that humans suffer from a higher risk of choking than any other meat-eating species (cf. Clegg & Aiello 2000).

Furthermore, X-ray videos show that the larynx is typically lowered away from the velum during vocal production in mammals, including primates, ungulates, and carnivores (Fitch 2000b, Fitch et al. 2016). These data indicate that the capacity to dynamically reposition the larynx lower in the vocal tract is a mammalian plesiomorphy.

Nonetheless, the permanently descended larynx of adult humans is a derived feature of our species, and the accompanying reconfiguration of vocal tract geometry has long been thought to provide a greatly expanded phonetic range relative to other primates (Lieberman et al. 1969, 1972). These early estimates were inferred from dead animals, before the possibility of dynamic vocal reconfiguration was appreciated. To address this shortcoming, my colleagues and I recently made X-ray videos of long-tailed macaques performing a variety of vocal tract maneuvers including vocalization, facial displays, chewing, and swallowing (Fitch et al. 2016). We used these radiographic images to create a vocal tract model, strictly limited to empirically observed vocal tract configurations, and then calculated the corresponding formant frequencies. The potential vowel space of this new model, although slightly smaller than that of an adult human, was eight times larger than previously estimated. Perceptual experiments showed that this extended vocal range could generate at least five clearly distinguishable vowels, and intelligible speech. Thus, previous estimates of nonhuman primate vocal potential were drastic underestimates, because they ignored the dynamic flexibility of the mammalian vocal tract. This observation refutes the pervasive idea

that the lowered human larynx is a necessary precondition for distinct speech, and clearly contradicts the widespread belief that larynx position explains the inability of chimpanzees or other apes to speak (e.g., Crystal 2003, Harley 2014).

Another relatively recent discovery is that a permanently descended larynx is not uniquely human. Although first discovered in two deer species (Fitch & Reby 2001), a permanently descended larynx has now been observed in various other mammals, including lions and other *Panthera* cats, koalas, and some ungulates (Charlton et al. 2011, Frey & Riede 2003, Weissengruber et al. 2002). This repeated convergent evolution of a descended larynx suggests some consistent selective force, and none of these other mammals produce speechlike variation in formant patterns. The current leading hypothesis is that a descended larynx, by elongating the vocal tract and lowering all formant frequencies, enables callers to mimic the formants of a larger animal and thus to acoustically exaggerate their size (Charlton & Reby 2016, Fitch & Reby 2001). Playback experiments, using resynthesized vocalizations, have verified this "size exaggeration" hypothesis in several species (Charlton et al. 2012, Reby et al. 2005). Also consistent with this hypothesis is the finding that human males, at puberty, undergo a secondary descent of the larynx (Fitch & Giedd 1999). Although this does not increase phonetic range or intelligibility (de Boer 2010), it lowers formants, which provide a key cue to body size (Pisanski et al. 2016). Thus, size exaggeration provides a clear alternative hypothesis to phonetic range expansion as an adaptive function of laryngeal descent. Finally, imaging studies of developing chimpanzees show that they too undergo a small but significant descent of the larynx and hyoid during the first years of life, so early laryngeal descent itself is not autapomorphic to humans (Nishimura et al. 2006).

In summary, the relevance of a descended larynx to human speech has been greatly overemphasized, and an unmodified primate or mammal vocal tract would be perfectly adequate to produce intelligible spoken language. Although the early and marked descent of the human larynx remains a derived human trait, relative to other primates, and undoubtedly has an influence on our species' vocal range, it is not the Rubicon it is often suggested to be.

A second derived trait of the human vocal tract has received much less attention—our loss of laryngeal air sacs (Fitch 2000a). All nonhuman great apes have large air sacs opening out of the larynx and extending down into the chest (Hewitt et al. 2002). This great ape synapomorphy indicates that our LCA with chimpanzees possessed laryngeal air sacs, which were lost in subsequent hominin evolution—a human autapomorphy. Indeed, the recent discovery of a bullate hyoid in *Australopithecus*, closely resembling that of chimpanzees, strongly suggests that early hominins retained air sacs (**Figure 2**). Later fossil hyoids from, for example, Neanderthals closely resemble those of modern humans, suggesting that the loss of air sacs is synapomorphic for the genus *Homo* (cf. Fitch 2010).

Understanding why we lost air sacs requires a clear understanding of their function in other apes, which remains elusive. Although multiple hypotheses are tenable (cf. Hewitt et al. 2002), both modeling (de Boer 2009) and empirical (Harris et al. 2006) analyses suggest that one function of air sacs is to lower formant frequencies, paralleling the acoustic effect of a descended larynx. Air sacs could thus also have a size-exaggerating vocal function (de Boer 2009). It is therefore plausible that the loss of air sacs in *Homo* was acoustically compensated for by a lowered larynx. Because air sacs can harbor life-threatening infections (Guilloud & McClure 1969), their loss in the *Homo* lineage may have been a preadaptation for, or response to, the emigration of our genus from warmer African regions into the temperate Old World.

In summary, the peripheral vocal anatomy of our species, although following a basic ground plan shared with other mammals, is autapomorphic relative to other primates in two respects: We have lost laryngeal air sacs and gained a permanently descended larynx. Although both traits affect vocal acoustics, neither is responsible for the inability of nonhuman primates to produce novel,

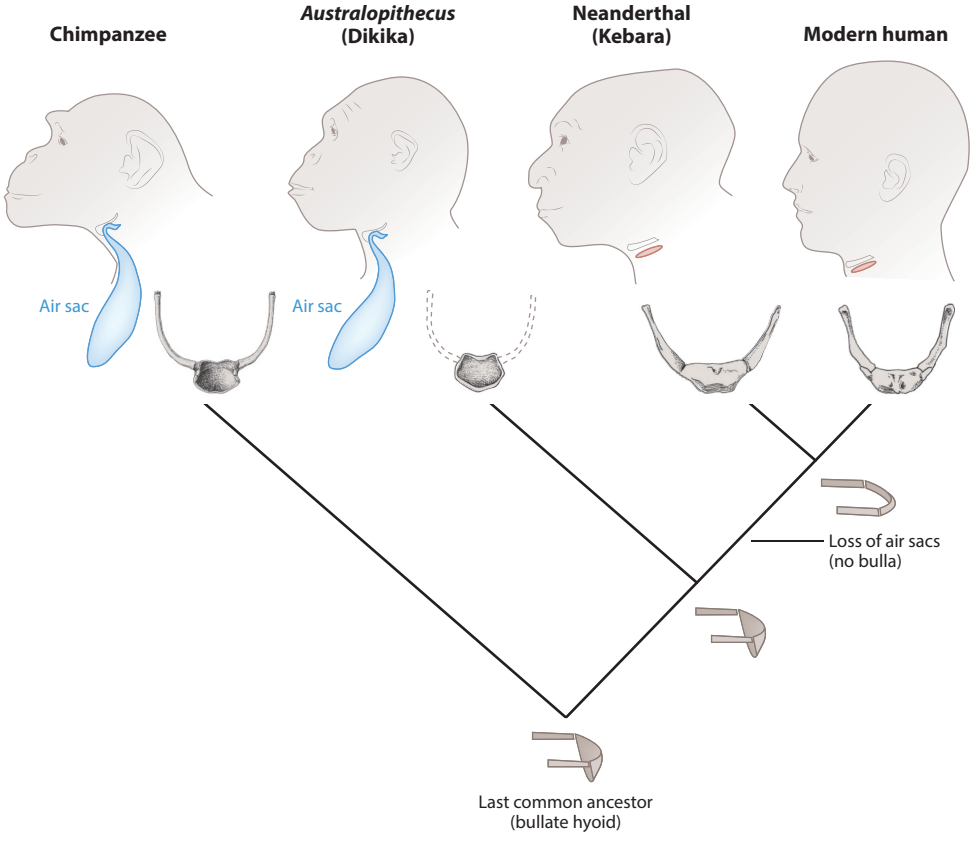


Figure 2

Loss of laryngeal air sacs as inferred from fossil hyoid bones. Air sacs are balloon-like outpouchings extending from the laryngeal vestibule out into the neck and chest region. Chimpanzees have large laryngeal air sacs that invaginate into the swollen, concave "bulla" of the basihyoid bone (illustrated below each species). The Dikika australopithecine's hyoid bone also has such a bulla, implying that these hominids still possessed air sacs. The bulla is lost in known hyoid bones from the genus *Homo*, including the Kebara Neanderthal, implying that the common ancestor of Neanderthals and modern humans had already lost air sacs.

learned vocalizations. To understand this core difference, we must turn to the neural control of vocalization.

6. CENTRAL CONTROL OF VOCAL PRODUCTION

In contrast to the vocal periphery, where the two well-attested human autapomorphies are of limited relevance to the human capacity to produce learned speech sounds, there are at least two autapomorphies regarding the neural control of this apparatus that are of central significance: (*a*) direct (monosynaptic) connections between cortex and vocal motor neurons and (*b*) extensive novel, dorsal intracortical connections between motor and auditory cortices. I discuss each in turn, arguing that together they underpin the rich capacity for vocal production learning characterizing humans but not other primates (Janik & Slater 2000).

Direct connections: monosynaptic connections between the motor cortex and the final motor neurons in the brainstem or spinal cord, believed to subserve voluntary motor control

The primary motor neurons that control the larynx and vocal articulators are located in the brainstem, and send their axons to the muscles in these organs via cranial nerves shared with

other mammals—the shared "brainstem chassis" for vocal motor output (Fitch 2010, Jürgens 2002). In all mammals, unlearned species-typical vocalizations ("innate calls") are produced by a subcortical neural network that feeds these motor neurons, centered on the periaqueductal gray (PAG) (Jürgens 1994). Stimulation of PAG neurons elicits natural-sounding innate calls in many species, and humans born without cortex can still produce laughter, screams, and cries (the innate calls of our own species). This core vocal circuitry receives only indirect input from the cortex, especially the anterior cingulate and motor cortex. These indirect connections are presumably responsible for the ability of all mammals to voluntarily inhibit or produce innate vocalizations. Consequently, and contrary to widespread belief, all primates tested can be trained to voluntarily produce or withhold vocalizations, albeit with difficulty (Adret 1992, Hage et al. 2016, Larson et al. 1973)—a basic form of vocal learning (termed "call usage learning"; Janik & Slater 2000), supported by this plesiomorphic neural circuitry.

What makes humans unusual is our capacity to produce novel, learned vocalizations beyond the innate call repertoire. The current dominant hypothesis explaining this ability is that humans, uniquely among primates, have an additional category of neural connections: direct monosynaptic connections between neurons in the motor cortex and the primary motor neurons controlling laryngeal musculature (Jürgens 2002, Ploog 1988). This class of direct cortico-motor connections is similar to those that monosynaptically connect the cortex to the motor neurons controlling the fingers in apes and some monkey species, providing them with voluntary control over individual finger movements (Lemon & Griffiths 2005). In addition, apes and some monkeys have direct connections to the facial motor neurons controlling the lips and hypoglossal neurons controlling the tongue (Jürgens & Alipour 2002, Simonyan 2014). Consistent with this, apes have good voluntary control and learning abilities over their lips and tongues (Marshall et al. 1999, Reynolds Losin et al. 2008, Wich et al. 2009).

Thus, it is specifically direct connections to the laryngeal motor neurons controlling phonation, in the posterior nucleus ambiguus of the medulla, that appear to be missing in other primates (Jürgens 2002). The absence of these connections in nonhuman primates is demonstrated by direct tracing studies in many primate species (Simonyan & Jürgens 2003). Their absence in great apes is less clear, because apes are rarely subjected to such invasive neural tracing procedures. The single tracing study on chimpanzees (Kuypers 1958b) reported direct connections to the anterior nucleus ambiguus (from whence the cricothyroid muscle, involved in pitch control, is innervated) but not the posterior portion controlling the phonatory muscles (cricoarytenoids). Thus, the key neural connections mediating human voluntary laryngeal control appear to be missing in chimpanzees.

Experimental tracing data are not available for humans, where we must rely on spontaneously occurring lesions, and the resulting axonal deterioration, for evidence, but two such patient-based studies are consistent with the existence of direct monosynaptic connections in humans (Iwatsubo et al. 1990, Kuypers 1958a). Also consistent are the short delay times (~10 ms) from cortical stimulation to laryngeal muscular contraction in humans, which are comparable to those characterizing known monosynaptic motor targets like the tongue (Rödel et al. 2003, 2004).

Finally, support for this "direct connections" hypothesis comes from convergent evolution in those nonhuman species that are capable of vocal production learning, which include numerous clades of birds and mammals (cf. Fitch & Jarvis 2013, Janik & Slater 2000). In both parrots and songbirds, which are vocal-learning birds, direct connections to the syringeal motor neurons have been documented that are not present in non-vocal-learning birds (Striedter 1994, Wild 1997). Similar studies are possible, but not yet completed, in vocal-learning mammals such as seals or bats (Knörnschild 2014, Reichmuth & Casey 2014, Vernes 2017).

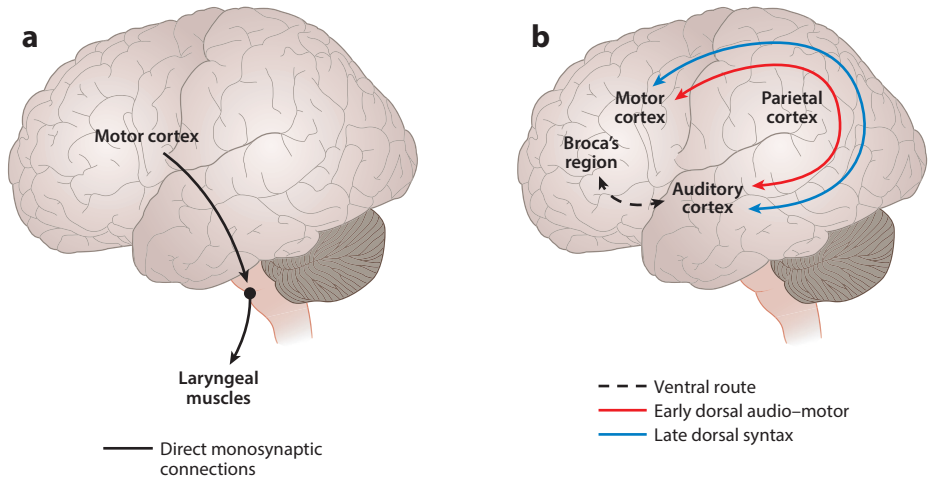


Figure 3

Neural autapomorphies associated with vocal motor control. (*a*) Direct connections. Humans appear to be unique among primates in possessing direct, monosynaptic connections from motor cortex to the laryngeal motor neurons in the medulla (*black circle*) that control the laryngeal muscles controlling phonation. (*b*) Intracortical connections. Humans share with other primates a ventral pathway connecting auditory with premotor cortex (*dashed line*); but humans have two additional dorsal pathways that are uniquely developed in our species. The inner pathway is mature early and is hypothesized to play a role in audio–motor mapping and vocal production learning; the other outer pathway matures later and is hypothesized to play a key role in syntax processing.

In summary, a considerable amount of comparative data is consistent with the direct connections hypothesis. Direct monosynaptic connections from the motor cortex onto the primary motor neurons controlling phonation, absent in other primates, thus constitute a key human autapomorphy for vocal control that evolved in hominins sometime after our divergence from chimpanzees (**Figure 3*a***).

A newly proposed potential neural autapomorphy, based on recent fMRI and comparative data, is that humans have two laryngeal motor control regions rather than the single region found in monkeys (Belyk & Brown 2017). The ventral region in humans appears to be homologous to that of monkeys, but a dorsal region, from which direct cortico-motor connections in humans stem (Simonyan & Horwitz 2011), may be unique to our species.

Of course, enhanced voluntary motor control alone is not sufficient for vocal production learning; the ability to map auditory representations onto a motor output is also required. This ability appears to require dorsal intracortical connections from the auditory cortex via the parietal to motor cortex that are uniquely well developed in humans—an autapomorphic dorsal circuit (Friederici 2017, Rilling et al. 2008). Macaques, chimpanzees, and humans all possess a robust ventral intracortical "what" pathway connecting the auditory to the prefrontal cortex, so this is a primate synapomorphy. These connections are believed to support auditory categorization (Rauschecker & Scott 2009). However, humans have an additional uniquely well-developed dorsal "how" pathway, involved in speech planning and audio–vocal matching, which is only weakly developed in other primates, where it seems to support spatial localization or "where" processing (Rauschecker & Scott 2009).

The autapomorphic human pathway at the heart of the dorsal circuit, the arcuate fasciculus, presumably supports the audio–vocal mapping process required to learn words, and matures

early in development (Bräuer et al. 2011). Intriguingly, humans also have a second parallel dorsal pathway, which matures much later (Friederici 2017), connecting the temporal and parietal cortex with Broca's area, the most expanded cortical region known in humans (Schenker et al. 2010). Although Broca's area appears to support morphology and syntax more than speech, it is possible that it arose in evolution via an exaptive duplication of the vocal–motor pathway, which would thus constitute a preadaptation for syntax (cf. Carstairs-McCarthy 1999, Fitch 2011). Both components of the arcuate fasciculus thus constitute neural autapomorphies in our species, but only the early-developing inner pathway appears to be directly relevant to speech and vocal learning.

Broca's area: a greatly expanded region of the human lateral prefrontal cortex (Brodmann's areas 44 and 45), playing a central role in language processing

7. FOSSIL CUES TO SPEECH?

There have been many attempts over the years to draw inferences about speech and/or language from fossils, but most rest on a weak comparative foundation (Fitch 2009b). The most clearly cognitively relevant data come from braincase volumes, which demonstrate that brain sizes in australopithecines, who already walked upright, were still within the ape range (Holloway et al. 2004). A large increase occurred with *Homo erectus*, and brain size continued to increase in our genus until reaching a maximum in Neanderthals and some archaic *Homo* specimens. But any inferences that can be derived about language abilities from brain size are tenuous because even microcephalic humans can have normal language; language depends more on specific circuitry than overall brain size (Lenneberg 1967).

Another potential fossil cue is the size of the thoracic spinal canal, which contains motor neurons controlling intercostal breathing (but not diaphragmatic breathing). This canal is enlarged in humans relative to apes (MacLarnon & Hewitt 2004), potentially reflecting superior breath control in our species (although I know of no behavioral evidence for this superiority). The canal remained small in australopithecines and *Homo ergaster* but had enlarged to modern size in Neanderthals, which is hypothesized to reflect an increase in breath control important for speech (MacLarnon & Hewitt 2004). It may also be relevant to singing, which requires finer breath control than speech (cf. Fitch 2009b).

Other proposed fossil cues have fared less well. Basicranial angle was once proposed to provide a cue to larynx height (George 1978), but this correlation does not exist in developing humans (Lieberman et al. 2001) or other mammals with descended larynges. Hypoglossal canal volume, potentially an indicator of tongue control (Kay et al. 1998), turns out not to reflect the size of the hypoglossal nerve within it, and overlaps considerably between apes and humans (Jungers et al. 2003). Hyoid bone anatomy, as discussed above, was chimpanzee-like in *Australopithecus* but already modern in Neanderthals, providing an indication of when air sacs were lost, but little more (Alemseged et al. 2006, Arensburg et al. 1990). Nor do fossilized ear ossicles, also discussed above, provide convincing evidence of speech-related auditory changes, despite previous claims (Martínez et al. 2004). Thus, despite decades of research and debate, fossil cues remain quite inconclusive, although when taken together they suggest that the genus *Homo* contained the first hominins in which potentially speech-related changes occurred.

8. GENETIC UNDERPINNINGS OF SPEECH

I end with a brief discussion of the genetic basis for speech (for more detail, see Fisher 2017, Fisher & Vernes 2015, Pääbo 2014). Nothing is currently known about the physiological or genetic determinants of laryngeal descent or air sac loss, so I focus on the genetic determinants of neural mechanisms, where exciting progress has been made in the last two decades.

By far the best-known gene involved in human vocal control is *FOXP2*, a transcription factor (see Section 2) that is found in all vertebrates but exists in modified form in humans. The gene was discovered thanks to a large British family in which this gene is damaged; family members with a damaged *FOXP2* allele show severe developmental speech control deficits as their core symptom (Vargha-Khadem et al. 2005). Detailed research shows that the problem is quite specific to oromotor sequencing control, including speech and sequences of nonspeech oral actions (contra early erroneous claims, based on cursory examination, of a "grammatical" deficit; Gopnik 1990). Affected family members appear to have normal singing ability (and thus preserved laryngeal control; Alcock et al. 2000) and other motor abilities (e.g., hand control for signed language; Watkins et al. 2002a). They also have more-subtle perceptual deficits that may be secondary to their production problems. Genetic analysis of this family led to the discovery and sequencing of the *FOXP2* gene (Fisher et al. 1998, Lai et al. 2001).

Forkhead box P2 (*FOXP2*): a transcription factor thought to play a key role in improved oral motor sequencing capabilities in humans

Once the human gene was isolated, it was rapidly sequenced in other species. It quickly became evident that the protein product of *FOXP2* is generally highly conserved among mammals, but that the human version differs from that of chimpanzees in two amino acids (Enard et al. 2002). With the exception of clinical cases, this derived human version is shared by all humans around the world: It is a genetic human autapomorphy showing precisely the comparative pattern expected of a gene involved in human speech and language.

The full panoply of modern genetic methods is now being utilized to discover the detailed function of *FOXP2*. The first steps were to damage or suppress the gene's expression in animal models and to insert the humanized version of the gene into genetically engineered mice (Enard et al. 2009, French et al. 2007, Groszer et al. 2008). These studies demonstrated subtle differences in motor learning, but little or no effect on vocalizations—not surprisingly, because mice are not vocal production learners (some subtle potential effects on newborn pup calls do not extend later in development; Hammerschmidt et al. 2015), although more recent work suggests that *Foxp2* knockout may also have an effect on motor development in mice (Castellucci et al 2016). However, investigations of vocal learning songbirds showed that *FOXP2* and its close relative *FOXP1* do play an important role in song learning in zebra finches (Haesler et al. 2007, Wohlgemuth et al. 2014)—another example of deep homology in which a convergently evolved trait (vocal production learning in birds and humans) depends on a homologous developmental gene, *FOXP2* (Fitch 2009a, Scharff & Petri 2011).

Brain imaging studies in humans suggest that *FOXP2* damage has both a general disruptive effect on the cortical machinery of speech production and more specific effects within the basal ganglia (Watkins et al. 2002b). Comparisons of humanized and wild-type mice allow a precise dissection of these effects: The human *FOXP2* allele leads to increased complexity in the dendritic arbors of medium spiny neurons in the striatum in the basal ganglia (Enard et al. 2009). These changes appear to affect the balance between hippocampal- and striatal-based motor learning, leading to earlier "proceduralization" of motor knowledge in humanized mice (Schreiweis et al. 2014). Ideally, similar genetic experiments will be carried out in mammalian vocal learners (e.g., bats or seals) to investigate what this might mean for vocal learning in particular. Intriguingly, bats show uniquely extensive interspecific variation in the *FOXP2* gene (Li et al. 2007), but the neural and behavioral significance of this variation remains unknown, providing an important rationale for investigations of vocal learning in bats (Vernes 2017).

Perhaps the most exciting development concerning *FOXP2* came when the gene was sequenced in Neanderthals, using DNA recovered from fossils (Krause et al. 2007). This research showed that Neanderthals shared the derived version of the human allele, producing an autapomorphic protein identical to our own. Given the importance of *FOXP2* in vocal development and sequencing, this finding strongly suggests that the common ancestor of humans and Neanderthals (along

with Denisovans, another extinct hominin) had already evolved complex vocal learning abilities comparable to our own, presumably supporting speech in this extinct ancestor who lived about 500,000 years ago (Pääbo 2014). However, in a recent twist, high-coverage sequencing revealed a human-specific difference in a noncoding regulatory region of the *FOXP2* gene, absent in Neanderthals, suggesting that there was further, later refinement of *FOXP2* expression unique to our species (Maricic et al. 2013). In any case, the new possibility of testing hypotheses about the timing of particular evolutionary changes, using paleo-DNA, is an exciting and promising new tool in our empirical arsenal for understanding language evolution (cf. Fitch 2017).

Unfortunately, *FOXP2* remains the only well-understood gene of the many genetic changes that presumably underlie the phenotypic changes underlying human speech and language. For example, the genetic underpinnings of either the direct cortico-motor or intracortical connections involved in human vocal control are virtually unknown. However, detailed comparisons of neural gene-expression patterns in songbirds and humans suggest multiple intriguing parallels (cf. Pfenning et al. 2014). The corresponding connections in birds involve axon-guidance genes in the SLIT/ROBO family, which has also been implicated in human dyslexia (Paracchini et al. 2007, Wang et al. 2015). Regarding intracortical connections, studies of songbird brain development reveal a central role for cadherin expression during the establishment of analogous connections between song-related brain regions, which may potentially be relevant to human direct connections. Finally, a recent hypothesis that new song nuclei are generated via gene duplication and divergence provides an attractive, if still speculative, hypothesis for the process that generated the two laryngeal motor regions found in humans (Belyk & Brown 2017). Crucially, as the genetic underpinnings of speech-related traits become better understood, we can rapidly check, for each candidate gene, whether novel alleles were shared with Neanderthals (whose entire genome is already sequenced) and thus derive increasingly strong inferences regarding Neanderthal speech abilities. Thus, genetic data perhaps provide the most promising and exciting empirical pathway for future research on the biology and evolution of speech.

9. SUMMARY

Comparative data acquired over the past two decades provide a sharper picture of both the shared plesiomorphic foundations of speech and key autapomorphic differences of humans relative to our nearest primate relatives. These human autapomorphies constitute key ingredients of the FLD—derived features that must have evolved in the last six million years since our divergence from chimpanzees. Overall, auditory perception appears to be based on mechanisms shared with other species, and the traditional hypothesis that speech perception is special to humans is not well supported. Rather, it seems that the basic circuitry required for auditory speech perception was already present in our mammalian ancestors long before speech evolved.

In contrast, regarding speech production, several clear autapomorphies (derived traits distinguishing humans from other apes) have been identified:

1. the loss of laryngeal air sacs;
2. the marked descent, during childhood, of the larynx and tongue base;
3. direct neural connections between the motor cortex and the motor neurons involved in phonation;
4. a massive expansion of dorsal intracortical connections spanning the auditory and motor cortices; and
5. a unique, derived form of *FOXP2*, shared by all nonclinical humans, that differs from that of chimpanzees.

A sixth potential autapomorphy remains preliminary: a novel cortical region devoted to laryngeal motor control (potentially associated with item 3). Similarly, the vast expansion of Broca's area, which may be tied to item 4, might be considered an independent human autapomorphy (although perhaps more concerned with syntax than speech; Friederici 2017).

Since I first reviewed the evolution of speech (Fitch 2000a), there have been several areas of major progress. The value of an explicitly comparative approach, based on a wide range of species, has been increasingly recognized. The hypothesis that the principal phenotypic changes required for the evolution of speech were neural, rather than anatomical components of the periphery, is becoming more widely accepted.

More specifically, the loss of air sacs can be dated, on the basis of fossil hyoid bones, roughly to the transition between *Australopithecus* and the genus *Homo*. The origin of our derived *FOXP2* allele can be dated, using paleo-DNA, to at latest our common ancestor with Neanderthals, roughly half a million years ago. Direct monosynaptic cortico-motor neurons have evolved, convergently, in songbirds and parrots, supporting the hypothesis that such connections are required for complex vocal learning. A descended larynx has evolved convergently in multiple mammalian species, demonstrating that this change can serve functions having nothing to do with speech; the leading hypothesis concerning the adaptive role of a descended larynx in mammals (including adult male humans) is that it exaggerates acoustic indicators of body size. Genes involved in audition and the *FOXP2* gene involved in vocal control provide examples of deep homology, where homologous genes have been harnessed to support the development of convergently evolved traits. Finally, the modern ability to recover DNA from fossils, and to combine these data with genetic studies in living humans to test hypotheses about when particular abilities or traits evolved, has already been demonstrated for *FOXP2*, and we can anticipate further exciting progress in this direction in the coming years.

In conclusion, in recent decades the evolution of speech has become the best-understood component of language evolution. Although speech is only one possible output modality for language (sign and writing are prominent alternatives), it provides a model for how the comparative approach could fuel progress in more central components of human language, including syntax, semantics, and pragmatics (cf. Fitch 2017).

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Preparation of this review was supported by an Austrian Science Fund (FWF) DK Grant "Cognition & Communication" (W1262-B29).

LITERATURE CITED

Adret P. 1992. Vocal learning induced with operant techniques: an overview. *Neth. J. Zool.* 43:125–42

Alcock KJ, Passingham RE, Watkins KE, Vargha-Khadem F. 2000. Pitch and timing abilities in inherited speech and language impairment. *Brain Lang.* 75:34–46

Alemseged Z, Spoor F, Kimbel WH, Bobe R, Geraads D, et al. 2006. A juvenile early hominin skeleton from Dikika, Ethiopia. *Nature* 443:296–301

Andics A, Gábor A, Gácsi M, Faragó T, Szabó D, Miklósi Á. 2016. Neural mechanisms for lexical processing in dogs. *Science* 353:1030–32

Andics A, Gácsi M, Faragó T, Kis A, Miklósi Á. 2014. Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Curr. Biol.* 24:574–78
Arensburg B, Schepartz LA, Tillier AM, Vandermeersch B, Rak Y. 1990. A reappraisal of the anatomical basis for speech in middle Paleolithic hominids. *Am. J. Phys. Anthropol.* 83:137–46
Baru AV. 1975. Discrimination of synthesized vowels [a] and [i] with varying parameters (fundamental frequency, intensity, duration and number of formants) in dog. In *Auditory Analysis and Perception of Speech*, ed. G Fant, MAA Tatham, pp. 91–101. New York: Academic
Belin P. 2006. Voice processing in human and non-human primates. *Philos. Trans. R. Soc. Lond.* 361:2091–107
Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. *Nature* 403:309–12
Belyk M, Brown S. 2017. The origins of the vocal brain in humans. *Neurosci. Biobehav. Rev.* 77:177–93
Bräuer J, Anwander A, Friederici AD. 2011. Neuroanatomical prerequisites for language functions in the maturing brain. *Cereb. Cortex* 21:459–66
Carstairs-McCarthy A. 1999. *The Origins of Complex Language*. Oxford, UK: Oxford Univ. Press
Castellucci GA, McGinley MJ, McCormick DA. 2016. Knockout of *Foxp2* disrupts vocal development in mice. *Sci. Rep.* 6:23305
Charlton BD, Ellis WAH, Larkin R, Fitch WT. 2012. Perception of size-related information in male koalas (*Phascolarctos cinereus*). *Anim. Cogn.* 15:999–1006
Charlton BD, Ellis WAH, McKinnon AJ, Cowin GJ, Brumm J, et al. 2011. Cues to body size in the formant spacing of male koala (*Phascolarctos cinereus*) bellows: honesty in an exaggerated trait. *J. Exp. Biol.* 214:3414–22
Charlton BD, Reby D. 2016. The evolution of acoustic size exaggeration in terrestrial mammals. *Nat. Commun.* 7:e12739
Clarey JC, Barone P, Imig TJ. 1992. Physiology of thalamus and cortex. In *The Mammalian Auditory Pathway: Neurophysiology*, ed. AN Popper, RR Fay, pp. 232–334. Berlin: Springer
Clegg M, Aiello LC. 2000. Paying the price for speech? An analysis of mortality statistics for choking on food. *Am. J. Phys. Anthropol.* 126(Suppl. 30):9482–83
Crystal D. 2003. *The Cambridge Encyclopedia of Language*. Cambridge, UK: Cambridge Univ. Press
Cutting JE. 1982. Plucks and bows are categorically perceived, sometimes. *Percept. Psychophys.* 31:462–76
Cutting JE, Rosner BS. 1974. Category boundaries in speech and music. *Percept. Psychophys.* 16:564–70
de Boer B. 2009. Acoustic analysis of primate air sacs and their effect on vocalization. *J. Acoust. Soc. Am.* 126:3329–43
de Boer B. 2010. Investigating the acoustic effect of the descended larynx with articulatory models. *J. Phon.* 38:679–86
Dooling RJ. 1992. Perception of speech sounds by birds. *Adv. Biosci.* 83:407–13
Elder JH. 1934. Auditory acuity of the chimpanzee. *J. Comp. Physiol. Psychol.* 17:157–83
Elemans CPH, Rasmussen JH, Herbst CT, Düring DN, Zollinger SA, et al. 2014. Universal mechanisms of sound production and control in birds and mammals. *Nat. Commun.* 6:8978
Eliades SJ, Wang X. 2008. Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* 453:1102–6
Enard W, Gehre S, Hammerschmidt K, Holter SM, Blass T, et al. 2009. A humanized version of *Foxp2* affects cortico-basal ganglia circuits in mice. *Cell* 137:961–71
Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, et al. 2002. Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* 418:869–72
Fant G. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton
Fecteau S, Armony JL, Joanette Y, Belin P. 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. *NeuroImage* 23:840–48
Fischer J. 2006. Categorical perception in animals. In *Encyclopedia of Language and Linguistics*, ed. K Brown, pp. 248–51. Oxford, UK: Elsevier. 2nd ed.
Fischer J, Teufel C, Drolet M, Patzelt A, Rübsamen R, et al. 2009. Orienting asymmetries and lateralized processing of sounds in humans. *BMC Neurosci.* 10:1–9
Fisher SE. 2017. Evolution of language: lessons from the genome. *Psychon. Bull. Rev.* 24:34–40

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.* 18:168–70
Fisher SE, Vernes SC. 2015. Genetics and the language sciences. *Annu. Rev. Linguist.* 1:289–310
Fitch WT. 2000a. The evolution of speech: a comparative review. *Trends Cogn. Sci.* 4:258–67
Fitch WT. 2000b. The phonetic potential of nonhuman vocal tracts: comparative cineradiographic observations of vocalizing animals. *Phonetica* 57:205–18
Fitch WT. 2009a. The biology and evolution of language: "deep homology" and the evolution of innovation. In *The Cognitive Neurosciences IV*, ed. MS Gazzaniga, pp. 873–83. Cambridge, MA: MIT Press
Fitch WT. 2009b. Fossil cues to the evolution of speech. In *The Cradle of Language*, ed. RP Botha, C Knight, pp. 112–34. Oxford, UK: Oxford Univ. Press
Fitch WT. 2010. *The Evolution of Language*. Cambridge, UK: Cambridge Univ. Press
Fitch WT. 2011. The evolution of syntax: an exaptationist perspective. *Front. Evol. Neurosci.* 3:1–12
Fitch WT. 2017. Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.* 24:3–33
Fitch WT, Fritz JB. 2006. Rhesus macaques spontaneously perceive formants in conspecific vocalizations. *J. Acoust. Soc. Am.* 120:2132–41
Fitch WT, Giedd J. 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *J. Acoust. Soc. Am.* 106:1511–22
Fitch WT, Hauser MD, Chomsky N. 2005. The evolution of the language faculty: clarifications and implications. *Cognition* 97:179–210
Fitch WT, Huber L, Bugnyar T. 2010. Social cognition and the evolution of language: constructing cognitive phylogenies. *Neuron* 65:795–814
Fitch WT, Jarvis ED. 2013. Birdsong and other animal models for human speech, song, and vocal learning. In *Language, Music, and the Brain: A Mysterious Relationship*, ed. MA Arbib, pp. 499–539. Cambridge, MA: MIT Press
Fitch WT, Mathur N, de Boer B, Ghazanfar AA. 2016. Monkey vocal tracts are speech-ready. *Sci. Adv.* 2:e1600723
Fitch WT, Reby D. 2001. The descended larynx is not uniquely human. *Proc. R. Soc. Lond. B* 268:1669–75
Fowler CA, Rosenblum LD. 1990. Duplex perception: a comparison of monosyllables and slamming doors. *J. Exp. Psychol. Hum. Percept. Perform.* 16:742–54
French CA, Groszer M, Preece C, Coupe A-M, Rajewsky K, Fisher SE. 2007. Generation of mice with a conditional *Foxp2* null allele. *Genesis* 45:440–46
Frey R, Riede T. 2003. Sexual dimorphism of the larynx of the Mongolian gazelle (*Procapra gutturosa* Pallas, 1777) (Mammalia, Artiodactyla, Bovidae). *Zool. Anz.* 242:33–62
Friederici AD. 2017. Evolution of the neural language network. *Psychon. Bull. Rev.* 24:41–47
Fritzsch B, Pauley S, Feng F, Matei V, Nichols DH. 2006. The molecular and developmental basis of the evolution of the vertebrate auditory system. *Int. J. Comp. Psychol.* 19:1–25
George SL. 1978. A longitudinal and cross-sectional analysis of the growth of the post-natal cranial base angle. *Am. J. Phys. Anthropol.* 49:171–78
Geschwind N. 1970. The organization of language and the brain. *Science* 170:940–44
Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK. 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* 25:5004–12
Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. 2007. Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr. Biol.* 17:425–30
Gil-da-Costa R, Hauser MD. 2006. Vervet monkeys and humans show brain asymmetries for processing conspecific vocalizations, but with opposite patterns of laterality. *Proc. R. Soc. Lond. B* 273:2313–18
Goodale MA, Milner AD. 1992. Separate visual pathways for perception and action. *Trends Neurosci.* 15:20–25
Gopnik M. 1990. Feature-blind grammar and dysphasia. *Nature* 344:715
Groszer M, Keays DA, Deacon RMJ, de Bono JP, Prasad-Mulcare S, et al. 2008. Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. *Curr. Biol.* 18:354–62
Guilloud NB, McClure HM. 1969. Air sac infection in the orang-utan. In *Proceedings of the 2nd International Congress of Primatology*, ed. CR Carpenter, 3:143–47. Basel, Switz.: Karger

Haesler S, Rochefort C, Geogi 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, Gavrilov N, Nieder A. 2016. Developmental changes of cognitive vocal control in monkeys. *J. Exp. Biol.* 219:1744–49
Hammerschmidt K, Schreiweis C, Minge C, Paabo S, Fischer J, Enard W. 2015. A humanized version of *Foxp2* does not affect ultrasonic vocalization in adult mice. *Genes Brain Behav*. 14:583–90
Harley T. 2014. *The Psychology of Language: From Data to Theory*. Sussex, UK: Psychology
Harnad SR. 1987. *Categorical Perception: The Groundwork of Cognition*. Cambridge, UK: Cambridge Univ. Press
Harris TR, Fitch WT, Goldstein LM, Fashing PJ. 2006. Black and white colobus monkey (*Colobus guereza*) roars as a source of both honest and exaggerated information about body mass. *Ethology* 112:911–20
Hauser M, Chomsky N, Fitch WT. 2002. The language faculty: What is it, who has it, and how did it evolve? *Science* 298:1569–79
Hayes C. 1951. *The Ape in Our House*. New York: Harper
Heffner HE, Heffner RS. 1986. Effect of unilateral and bilateral auditory cortex lesions on the discrimination of vocalizations by Japanese macaques. *J. Neurophysiol.* 56:683–701
Heffner RS. 2004. Primate hearing from a mammalian perspective. *Anat. Rec.* 281:A1111–22
Heimbauer LA, Beran MJ, Owren MJ. 2011. A chimpanzee recognizes synthetic speech with significantly reduced acoustic cues to phonetic content. *Curr. Biol.* 21:1210–14
Hemilä S, Nummela S, Reuter T. 1995. What middle ear parameters tell about impedance matching and high frequency hearing. *Hear. Res.* 85:31–44
Herbst CT, Stoeger AS, Frey R, Lohscheller J, Titze IR, et al. 2012. How low can you go? Physical production mechanism of elephant infrasonic vocalizations. *Science* 337:595–99
Hewitt G, MacLarnon A, Jones KE. 2002. The functions of laryngeal air sacs in primates: a new hypothesis. *Folia Primatol*. 73:70–94
Hienz RD, Jones AM, Weerts EM. 2004. The discrimination of baboon grunt calls and human vowel sounds by baboons. *J. Acoust. Soc. Am.* 116:1692–97
Holloway RL, Broadfield DC, Yuan MS, Schwartz JH, Tattersall I. 2004. *The Human Fossil Record, Brain Endocasts—The Paleoneurological Evidence*. Hoboken, NJ: Wiley
Iwatsubo T, Kuzuhara S, Kanemitsu A, Shimada H, Toyokura Y. 1990. Corticofugal projections to the motor nuclei of the brainstem and spinal cord in humans. *Neurology* 40:309–12
Janik VM, Slater PJB. 2000. The different roles of social learning in vocal communication. *Anim. Behav.* 60:1–11
Jungers WJ, Pokempner AA, Kay RF, Cartmill M. 2003. Hypoglossal canal size in living hominoids and the evolution of human speech. *Hum. Biol.* 75:473–84
Jürgens U. 1994. The role of the periaqueductal grey in vocal behaviour. *Behav. Brain Res.* 62:107–17
Jürgens U. 2002. Neural pathways underlying vocal control. *Neurosci. Biobehav. Rev.* 26:235–58
Jürgens U, Alipour M. 2002. A comparative study on the cortico-hypoglossal connections in primates, using biotin dextranamine. *Neurosci. Lett.* 328:245–48
Kay RF, Cartmill M, Balow M. 1998. The hypoglossal canal and the origin of human vocal behavior. *PNAS* 95:5417–19
Kellogg WN. 1968. Communication and language in the home-raised chimpanzee. *Science* 162:423–27
Kluender KR, Lotto AJ, Holt LL, Bloedel SL. 1998. Role of experience for language-specific functional mappings of vowel sounds. *J. Acoust. Soc. Am.* 104:3568–82
Knörnschild M. 2014. Vocal production learning in bats. *Curr. Opin. Neurobiol.* 28:80–85
Kojima S. 1990. Comparison of auditory functions in the chimpanzee and human. *Folia Primatol.* 55:62–72
Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, et al. 2007. The derived *FOXP2* variant of modern humans was shared with Neandertals. *Curr. Biol.* 17:1908–12
Kuhl PK, Meltzoff AN, Williams KA, Lacerda F, Stevens KN, Lindblom B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255:606–8
Kuhl PK, Miller JD. 1975. Speech perception by the chinchilla: voiced–voiceless distinction in alveolar plosive consonants. *Science* 190:69–72
Kuypers HGJM. 1958a. Corticobulbar connections to the pons and lower brainstem in man: an anatomical study. *Brain* 81:364–88

Kuypers HGJM. 1958b. Some projections from the pericentral cortex to the pons and lower brain stem in monkey and chimpanzee. *J. Comp. Neurol.* 110:221–55
Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 413:519–23
Larson CR, Sutton D, Taylor EM, Lindeman R. 1973. Sound spectral properties of conditioned vocalizations in monkeys. *Phonetica* 27:100–12
Lemon RN, Griffiths J. 2005. Comparing the function of the corticospinal system in different species: organizational differences for motor specialization? *Muscle Nerve* 32:261–79
Lenneberg EH. 1967. *Biological Foundations of Language*. New York: Wiley
Li G, Wang J, Rossiter SJ, Jones G, Zhang S. 2007. Accelerated *FoxP2* evolution in echolocating bats. *PLOS ONE* 2:e900
Liberman AM. 1957. Some results of research on speech perception. *J. Acoust. Soc. Am.* 29:117–23
Liberman AM, Isenberg D, Rakerd B. 1981. Duplex perception of cues for stop consonants: evidence for a phonetic mode. *Percept. Psychophys.* 30:133–43
Liberman AM, Mattingly IG. 1985. The motor theory of speech perception revised. *Cognition* 21:1–36
Lieberman DE, McCarthy RC, Hiiemae K, Palmer JB. 2001. Ontogeny of postnatal hyoid and larynx descent in humans. *Arch. Oral Biol.* 46:117–28
Lieberman PH, Crelin ES, Klatt DH. 1972. Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *Am. Anthropol.* 74:287–307
Lieberman PH, Klatt DH, Wilson WH. 1969. Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science* 164:1185–87
Lombard RE, Bolt J. 1979. Evolution of the tetrapod ear: an analysis and reinterpretation. *Biol. J. Linn. Soc.* 11:19–76
MacLarnon AM, Hewitt GP. 2004. Increased breathing control: another factor in the evolution of human language. *Evol. Anthropol.* 13:181–97
Manley GA. 2000. Cochlear mechanisms from a phylogenetic viewpoint. *PNAS* 97:11736–43
Maricic T, Günther V, Georgiev O, Gehre S, Curlin M, et al. 2013. A recent evolutionary change affects a regulatory element in the human *FOXP2* gene. *Mol. Biol. Evol.* 30:844–52
Marshall AJ, Wrangham RW, Arcadi AC. 1999. Does learning affect the structure of vocalizations in chimpanzees? *Anim. Behav.* 58:825–30
Martínez I, Rosa M, Arsuaga J-L, Jarabo P, Quam R, et al. 2004. Auditory capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *PNAS* 101:9976–81
Martínez I, Rosa M, Quam R, Jarabo P, Lorenzo C, et al. 2013. Communicative capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Quat. Int.* 295:94–101
McGurk H, MacDonald J. 1976. Hearing lips and seeing voices. *Nature* 264:746–48
Nelson DA, Marler P. 1989. Categorical perception of a natural stimulus continuum: birdsong. *Science* 244:976–78
Nishimura T, Mikami A, Suzuki J, Matsuzawa T. 2006. Descent of the hyoid in chimpanzees: evolution of face flattening and speech. *J. Hum. Evol.* 51:244–54
Nottebohm F. 1971. Neural lateralization of vocal control in a passerine bird. I. Song. *J. Exp. Zool.* 177:229–62
Ohms VR, Gill A, van Heijningen C, Beckers GJL, ten Cate C. 2009. Zebra finches exhibit speaker-independent phonetic perception of human speech. *Proc. R. Soc. Lond. B.* **https:/doi.org/10.1098/rspb.2009.1788**
Pääbo S. 2014. The human condition—a molecular approach. *Cell* 157:216–26
Paracchini S, Scerri T, Monaco AP. 2007. The genetic lexicon of dyslexia. *Annu. Rev. Genom. Hum. Genet.* 8:57–79
Pastore RE, Schmuckler MA, Rosenblum L, Szczesiul R. 1983. Duplex perception with musical stimuli. *Percept. Psychophys.* 33:469–74
Perrodin C, Kayser C, Logothetis NK, Petkov CI. 2011. Voice cells in the primate temporal lobe. *Curr. Biol.* 21:1408–15
Petersen MR, Beecher MD, Zoloth SR, Green S, Marler P, et al. 1984. Neural lateralization of vocalizations by Japanese macaques: Communicative significance is more important than acoustic structure. *Behav. Neurosci.* 98:779–90

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. 2008. A voice region in the monkey brain. *Nat. Neurosci.* 11:367–74

Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* 346:1256846

Pisanski K, Mora EC, Pisanski A, Reby D, Sorokowski P, et al. 2016. Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Sci. Rep.* 6:34389

Ploog DW. 1988. Neurobiology and pathology of subhuman vocal communication and human speech. In *Primate Vocal Communication*, ed. D Todt, P Goedeking, D Symmes, pp. 195–212. Berlin: Springer

Poremba A, Malloy M, Saunders RC, Carson RE, Herscovitch P, Mishkin M. 2004. Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature* 427:448–51

Rao MRKM, Choudhary C, Ali S. 1984. Case of sudden death in a panther (*Panthera pardus*) with choke. *Indian Vet. J.* 61:618–19

Rauschecker JP, Scott SK. 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12:718–24

Rauschecker JP, Tian B. 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *PNAS* 97:11800–6

Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton-Brock T. 2005. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proc. R. Soc. Lond. B* 272:941–47

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

Remez RE, Rubin PE, Pisoni DB, Carrell TD. 1981. Speech perception without traditional speech cues. *Science* 212:947–50

Repp BH. 1982. Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychol. Bull.* 92:81–110

Reynolds Losin EA, Russell JL, Freeman H, Meguerditchian A, Hopkins WD. 2008. Left hemisphere specialization for oro-facial movements of learned vocal signals by captive chimpanzees. *PLOS ONE* 3:e2529

Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, et al. 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. *Nat. Neurosci.* 11:426–28

Rödel RMW, Laskawi R, Markus H. 2003. Tongue representation in the lateral cortical motor region of the human brain as assessed by transcranial magnetic stimulation. *Ann. Otol. Rhinol. Laryngol.* 112:71–76

Rödel RMW, Olthoff A, Tergau F, Simonyan K, Kraemer D, et al. 2004. Human cortical motor representation of the larynx as assessed by transcranial magnetic stimulation (TMS). *Laryngoscope* 114:918–22

Rogers LJ, Andrew JR. 2002. *Comparative Vertebrate Lateralization*. Cambridge, UK: Cambridge Univ. Press

Rosen SM, Howell P. 1981. Plucks and bows are not categorically perceived. *Percept. Psychophys.* 30:156–68

Sasaki CT, Levine PA, Laitman JT, Crelin ES. 1977. Postnatal descent of the epiglottis in man: a preliminary report. *Arch. Otolaryngol.* 103:169–71

Savage-Rumbaugh ES, Murphy J, Sevcik RA, Brakke KE, Williams SL, Rumbaugh DM. 1993. Language comprehension in ape and child. *Monogr. Soc. Res. Child Dev.* 58:1–221

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* 366:2124–40

Schenker NM, Hopkins WD, Spocter MA, Garrison AR, Stimpson CD, et al. 2010. Broca's area homologue in chimpanzees (*Pan troglodytes*): probabilistic mapping, asymmetry and comparison to humans. *Cereb. Cortex* 20:730–42

Schreiweis C, Bornschein U, Burguière E, Kerimoglu C, Schreiter S, et al. 2014. Humanized *Foxp2* accelerates learning by enhancing transitions from declarative to procedural performance. *PNAS* 111:14253–58

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. *Nature* 457:818–23

Simonyan K. 2014. The laryngeal motor cortex: its organization and connectivity. *Curr. Opin. Neurobiol.* 28:15–21

Simonyan K, Horwitz B. 2011. Laryngeal motor cortex and control of speech in humans. *Neuroscientist* 17:197–208

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

Sinnott JM, Adams FS. 1987. Differences in human and monkey sensitivity to acoustic cues underlying voicing contrasts. *J. Acoust. Soc. Am.* 82:1539–47
Sinnott JM, Williamson TL. 1999. Can macaques perceive place of articulation from formant transition information? *J. Acoust. Soc. Am.* 106:929–37
Skead DM. 1980. Whitebreasted cormorant *Phalacrocorax carbo* chokes on fish. *Cormorant* 8:27
Striedter GF. 1994. The vocal control pathways in budgerigars differ from those of songbirds. *J. Comp. Neurol.* 343:35–56
Suga N. 1990. Cortical computational maps for auditory imaging. *Neural Netw.* 3:3–21
Suthers RA, Fitch WT, Popper AN, Fay RR, ed. 2016. *Vertebrate Sound Production and Acoustic Communication.* New York: Springer
Taylor A, Reby D. 2010. The contribution of source-filter theory to mammal vocal communication research. *J. Zool.* 280:221–36
Thexton AJ, Crompton AW. 1998. The control of swallowing. In *The Scientific Basis of Eating: Taste and Smell, Salivation, Mastication and Swallowing, and Their Dysfunctions*, ed. RWA Linden, pp. 168–222. London: Karger
Tian B, Reser D, Durham A, Kustov A, Rauschecker JP. 2001. Functional specialization in rhesus monkey auditory cortex. *Science* 292:290–93
Titze IR. 2006. *The Myoelastic Aerodynamic Theory of Phonation*. Denver, CO: Natl. Cent. Voice Speech
Trout JD. 2003. Biological specializations for speech: What can the animals tell us? *Curr. Dir. Psychol. Sci.* 12:155–59
Tucker AS. 2017. Major evolutionary transitions and innovations: the tympanic middle ear. *Philos. Trans. R. Soc. Lond. B* 372:20150483
Vallortigara G, Rogers LJ. 2005. Survival with an asymmetrical brain: advantages and disadvantages of cerebral lateralization. *Behav. Brain Sci.* 28:575–633
van den Berg J. 1958. Myoelastic-aerodynamic theory of voice production. *J. Speech Hear. Res.* 1:227–44
Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. 2005. *FOXP2* and the neuroanatomy of speech and language. *Nat. Rev. Neurosci.* 6:131–38
Vernes SC. 2017. What bats have to say about speech and language. *Psychon. Bull. Rev.* 24:111–17
Wall CE, Smith KE. 2001. Ingestion in mammals. In *Encyclopedia of Life Sciences*, pp. 1–6. London: Macmillan
Wang R, Chen C-C, Hara E, Rivas MV, Roulhac PL, et al. 2015. Convergent differential regulation of *SLIT-ROBO* axon guidance genes in the brains of vocal learners. *J. Comp. Neurol.* 523:892–906
Wang VY, Hassan BA, Bellen HJ, Zoghbi HY. 2002. *Drosophila atonal* fully rescues the phenotype of *Math1* null mice: new functions evolve in new cellular contexts. *Curr. Biol.* 12:1611–16
Wang X, Kadia SC. 2001. Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J. Neurophysiol.* 86:2616–20
Watkins KE, Dronkers NF, Vargha-Khadem F. 2002a. Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain* 125:452–64
Watkins KE, Vargha-Khadem F, Ashburner J, Passingham RE, Connelly A, Friston KJ. 2002b. MRI analysis of an inherited speech and language disorder: structural brain abnormalities. *Brain* 125:465–78
Weissengruber GE, Forstenpointner G, Peters G, Kübber-Heiss A, Fitch WT. 2002. Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyx jubatus*), and domestic cat (*Felis silvestris* f. *catus*). *J. Anat.* 201:195–209
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
Wild JM. 1997. Neural pathways for the control of birdsong production. *J. Neurobiol.* 33:653–70
Wohlgemuth S, Adam I, Scharff C. 2014. *FoxP2* in songbirds. *Curr. Opin. Neurobiol.* 28:86–93
Yerkes RM, Yerkes AW. 1929. *The Great Apes*. New Haven, CT: Yale Univ. Press
Zatorre RJ, Evans AC, Meyer E, Gjedde A. 1992. Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256:846–49
Zoloth SR, Petersen MR, Beecher MD, Green S, Marler P, et al. 1979. Species-specific perceptual processing of vocal sounds by monkeys. *Science* 204:870–72