Review

Voice Modulation: A Window into the Origins of Human Vocal Control?

Katarzyna Pisanski,1,2 Valentina Cartei,1 Carolyn McGettigan,3 Jordan Raine,1 and David Reby1,*

An unresolved issue in comparative approaches to speech evolution is the apparent absence of an intermediate vocal communication system between human speech and the less flexible vocal repertoires of other primates. We argue that humans’ ability to modulate nonverbal vocal features evolutionarily linked to expression of body size and sex (fundamental and formant frequencies) provides a largely overlooked window into the nature of this intermediate system. Recent behavioral and neural evidence indicates that humans’ vocal control abilities, commonly assumed to subserve speech, extend to these nonverbal dimensions. This capacity appears in continuity with context-dependent frequency modulations recently identified in other mammals, including primates, and may represent a living relic of early vocal control abilities that led to articulated human speech.

The Dynamic Human Voice

Recent research examining the communicative function of the human voice has effectively established the roles of two key acoustic components – the fundamental frequency (F0) (see Glossary) and formants (Box 1) – in the expression of many biological and psychological dimensions, including sex and age, body size and shape, hormonal condition, dominance, masculinity or femininity, and attractiveness (for reviews see [1–5]). For example, men and women whose voices are characterized by low frequencies are typically judged by listeners as more dominant, physically larger and stronger, and more masculine than are speakers with higher-frequency voices [5], and these stereotypes appear partly driven by the physical, anatomical, and physiological mechanisms that influence and constrain F0 and formant production. In addition to typically having larger bodies than women, men also tend to have relatively larger larynges, longer vocal tracts, and lower-frequency voices (Box 1). As a consequence, listeners often associate lower frequencies with stereotypically male traits [6,7].

By revealing the many ways in which the human voice is linked – physically or perceptually – to ecologically relevant traits, this body of research has greatly advanced our understanding of the evolutionary functions of nonverbal vocal communication in humans. In particular it provides compelling evidence that the human voice, and the strong sexual dimorphism that characterizes its main frequency components, has been largely shaped by sexual selection [8] and continues to play a substantial role in human mate choice and mate competition [1,3]. At the same time, by focusing on vocal communication of mate quality, recent studies have overwhelmingly described human voice production and perception in terms of static cues that are almost exclusively studied in the absence of social context. This is clearly an oversimplification, as sexual selection may also favor the ability to flexibly manipulate the voice to exploit listeners’ tendency to
Box 1. Source–Filter Theory of Speech Production

In humans and most other mammals, vocal signals are produced by the larynx (the source) and subsequently filtered by the supralaryngeal vocal tract (the filter) [88,89]. When humans phonate, air that is expelled from the lungs and forced through the closed glottis causes oscillation of the vocal folds within the larynx, which determines F0, also perceived as voice pitch. The sound waves produced by this oscillation travel up from the larynx and through the pharynx and oral and nasal cavities, which comprise the supralaryngeal vocal tract, to the mouth, through which vocal sounds are radiated (Figure IC). In the process, the vocal tract filters the sound, attenuating certain frequencies and not others, thereby producing resonant frequencies or formants that affect our perception of voice timbre [90].

Following the principles of biomechanics and sound physics, larger vocal folds vibrate at a slower rate than do smaller vocal folds, resulting in a relatively lower F0 and perceived pitch; however, regardless of mass, F0 increases when the vocal folds are stretched and become tenser [16,17]. Thus, the effective mass, length, and tension of the vocal folds affect F0. Independently of this, longer vocal tracts produce relatively lower, more closely spaced formants [formant spacing (ΔF)] than do shorter vocal tracts [91,92] and hence a more perceptually resonant voice (Figure 1A,B). Both F0 and formants are inversely related to body size in most mammals [5,10,11,92]. Among humans, these voice-size relationships are particularly salient between sexes, where F0 and formants are substantially lower among men than women [29]. However, controlling for sex and age, formants explain several times more variation in body size and shape than does F0 [90,93].

Manipulations of the tongue, lips, jaw, and soft palate can also alter the shape (rather than length) of the vocal tract, affecting the relative distribution (rather than absolute scaling) of formant frequencies and thereby giving rise to different speech sounds. The relative positions of formants (especially F1 and F2) [90]; see, for example, Figure 1D, Key Figure) play a critical role in speech production and perception and in non-tonal languages play a much greater role than F0. The lowered position of the human larynx also allows humans to produce a wider range of sounds than any other primate [27,28]. However, whether the human larynx descended due to selection for a more complex vocal repertoire [94] or selection for apparent larger body size [27] remains unclear.

Glossary

Dual-pathway model: current neural models implicate two pathways in human vocal control: sensorimotor cortical systems that support the production of learned vocalizations such as speech and song; and the limbic system that supports innate vocalizations such as laughter. Although these two pathways have traditionally been thought of as separate, recent research on vocal control suggests otherwise.

Enculturation: the process by which an individual is exposed to and learns the traditional content of a culture or assimilates its practices and values; for instance, nonhuman primates that have been raised by humans in a human-like cultural and social environment.

Flexible: characterizes a behavior that can change as a function of various (e.g., social, temporal, environmental) factors.

Formant: resonant frequencies of the supralaryngeal vocal tract that affect our perception of voice timbre. The shape of the vocal tract influences the relative positions of formants and formant spacing, whereas its length (VTL) affects formant scaling. Longer vocal tracts produce lower, more closely spaced formants.

Fundamental frequency (F0): the rate of vocal fold vibration. All else being equal, longer, more massive, and less tense vocal folds tend to vibrate more slowly and produce a relatively lower F0. Voice pitch is the perceptual correlate of F0 (along with its harmonics, integer multiples of F0).

Language: a symbolic, rule-governed system of communication shared within a group that need not involve verbal communication or speech (e.g., written or sign language).

Speech: a vocal communication system involving precise coordination of vocal anatomical structures (larynx, supralaryngeal vocal tract, and articulators such as tongue and lips) required to articulate specific sounds.

Vocal control: the capacity to control the larynx (affecting the production of F0) or the supralaryngeal vocal tract (affecting the production of formants) in a flexible or voluntary manner, for

Figure 1. Voice Frequency Scaling. Each spectrogram illustrates the vowel /æ/ spoken by the same man, whose voice frequencies were manipulated using Praat acoustic software [95]. (A) Formant scaling, wherein lower, more closely spaced formants (left) correspond to a relatively longer vocal tract and a perceptually more resonant voice. Apparent vocal tract length (VTL) was estimated from ΔF; for algorithms see [91,92]. (B) Fundamental frequency (F0) scaling, wherein a lower F0 and more closely spaced harmonics (multiple integers of F0) (left) correspond to a perceptually lower-pitched voice. Note when comparing panels (A) and (B) that formants and F0 can vary independently. (C) Labeled sagittal MRI of the human vocal apparatus during the production of the sustained vowel /u:/, Note the back position of the tongue and protrusion of the lips. Spectrogram parameters: y-axis, 0-5 kHz; x-axis, 0.25 s; window length, 0.04.
associate sex-typical voice patterns with sex-typical traits (i.e., the frequency or gender code [6,7,9]). Although several animal species are known to scale their vocal frequencies to exaggerate their body size or to express threat [10–12], the unusually strong sexual dimorphism in formants and F0 that differentiates humans from other animals, including nonhuman primates [8], may offer a particularly effective platform by which humans can exploit sex-typical vocal patterns above and beyond size exaggeration. Hence, we suggest that men and women routinely take advantage of the salient sex differences in patterns above and beyond size exaggeration. Hence, we suggest that men and women routinely take advantage of the salient sex differences in F0 and formants, using the broad acoustic space between (and within) these male and female categories to express various sex-related attributes such as dominance, masculinity or femininity, and threat [9,13].

It is well known that, with vocal training, actors and voice impersonators can significantly raise or lower the key frequency components of their voices (F0 and formants), often adopting vocal ranges characteristic of the opposite sex [2]. In singing, sopranos can raise their F0 and match their first formant to frequencies above 1200 Hz [14], six times the average pitch of a woman’s voice. Some politicians (such as Margaret Thatcher) have had extensive vocal coaching to lower the frequency of their voice to exude a more authoritative, powerful persona [15]. Although these examples may be extreme or anecdotal, a growing body of empirical evidence suggests that people routinely and volitionally modulate their voice in everyday social contexts such as first dates and job interviews, when making a compelling argument, or when attempting to deceive someone: the human voice is a dynamic and flexible channel for self-expression and, beyond its role in encoding linguistic information, functions also as a social tool.

In this review we suggest that the ability of humans to flexibly control the size-related source–filter dimensions of our vocal signals (i.e., vocal control) is likely to predate our ability to articulate the verbal dimensions of speech and therefore may provide an evolutionary pathway from nonhuman primate vocal communication to human speech. We show that this largely overlooked modulation of nonverbal components is not only highly prevalent and functional in human vocal communication but also present in the vocal systems of other, nonhuman mammal species. We propose that voice modulation can provide key insights regarding the evolution of articulated human speech, focusing on how comparative explorations of the behavioral and neural bases of vocal control have the potential to lead to important advances in our understanding of the evolution of human vocal communication and the origins of speech.

**Behavioral Evidence of Human Vocal Modulation and its Social Functions**

Vocal control involves manipulation of the vocal folds and larynx or the supralaryngeal vocal tract, including the articulators such as tongue and lips. Mechanistically, F0 is controlled by manipulating the tension and effective length or surface area of the vocal folds; for example, by contracting or relaxing the thyroarytenoid and cricothyroid muscles or increasing subglottal pressure. By contrast, formants are altered by manipulating the length of the vocal folds; for example, by lowering the larynx or protruding the lips – within limits imposed by the skull and body size [2,10,16,17] (Box 1).

Recent studies investigating voice modulation in humans have focused on its functional role, typically in a mating context, and provide evidence that people dynamically alter sexually dimorphic and size-related vocal features (F0 and formants) to advertise or exaggerate traits that are relevant to mate selection and competition. For example, in studies involving simulated dating scenarios, men who perceived themselves as physically dominant relative to their male competitor lowered their F0 in conversations with him (whereas men who felt subordinate raised their F0 [13]) and produced more monotone speech with lower F0 variability when describing themselves to a potential female partner [18]. Men exhibiting masculine traits such as relatively large body size, high levels of androgens, or low formant frequencies also appear to articulate vowels less clearly than do more feminine men by differentially modulating the lower formants example as a function of social context.

**Vocal learning:** the capacity to produce novel vocalizations through imitation or experience (which does not necessarily involve vocal control).

**Voice modulation:** manipulation of any nonverbal property of the voice including but not limited to F0 and formant frequencies.

**Votional:** characterizes a voluntary and intentional behavior. Votional vocal control involves the capacity to control the larynx and supralaryngeal vocal tract independent of immediate context and physiological state.
The authors of this latter study suggest that some men may opt for communicatively less intelligible phonetic variants, typically associated with lower social economic status but also with toughness, as a means of communicating their masculinity [19]. Other studies show that men and women volitionally and spontaneously decrease F0 and formants when asked to sound masculine and increase these frequencies to sound feminine [9]. A similar ability has been observed in children aged 6–9 years [20], suggesting that prepubescent boys and girls learn to express their gender through the voice by imitating adult models. Our own ongoing research further indicates that men and women from distinct cultures spontaneously lower their F0 and formants to sound physically large and raise these vocal frequencies to sound small. These modulations are in some cases extreme, as men can increase their apparent vocal tract length (VTL) by 25% to sound larger or raise their voice pitch to frequencies characteristic of a small child when attempting to imitate a small body size.

Sex-typical voice patterns (i.e., relatively lower frequencies in men’s voices and higher frequencies in women’s voices) are generally considered attractive [3,5]. Recent research efforts to determine whether men and women can effectively manipulate vocal attractiveness have produced mixed results [21–25], but unequivocally indicate that both sexes modulate their F0 when speaking to attractive members of the opposite sex. Men and women also appear to volitionally raise their F0 to express confidence and intelligence [21]. At a perceptual level, listeners should be selected and/or learn to detect vocal modulation, given that misattribution of speaker traits such as dominance and attractiveness can be costly. Evidence for this hypothesis is also mixed: although listeners correctly discriminate deliberately modulated from habitual speech [21], modulated voices can affect listeners’ assessments of target traits including intelligence, dominance, and attractiveness [21,24,25], suggesting that even when detectable, voice modulation can remain effectual (cf. [24]).

The behavioral studies reviewed above show that humans can spontaneously manipulate vocal frequencies to deemphasize or accentuate various biosocially relevant traits with meaningful variation across social contexts, within the limits imposed by various anatomical or mechanistic constraints on vocal production (see, for example, Box 1). Volitional modulation of F0 and formants is often achieved by exploiting robust perceptual correspondences (e.g., gender [9] or frequency code [6,7,12]). Such vocal modulation has obvious potential benefits. For example, between- and within-individual variation in F0 and formants can influence others’ social attributions and mate-choice decisions [21,25] and predicts socioeconomic outcomes including success in a job interview [22] and political votes [26]. We suggest that similar advantages associated with vocal modulation during social interactions are likely to have led to the selection of this capacity in our ancestors [7,12,27].

**Is Voice Modulation the Origin of Vocal Control?**

Vocal modulation of source–filter components is important for conveying (or exaggerating) various indexical characteristics of the speaker, as outlined above, but is also necessary for speech production (Box 1). Thus, although behavioral studies of humans offer compelling evidence for vocal modulation and its influence on others, they do not alone indicate whether F0 and formant modulation emerged to subserve speech [27,28], to exploit apparent indexical features of the speaker [7,12,27], or potentially, for both purposes.

To answer this question, we must consider at least two additional, comparative lines of research. First, we need to examine vocal modulation in other mammals that lack articulated speech, including our closest primate relatives as well as other mammals known to scale their vocal frequencies, and to compare this with human nonverbal vocal communication as well as speech. Here it is important to distinguish between involuntary vocal variation (i.e., automatically elicited by environmental stimuli or different levels of arousal) and more controlled vocal modulation.
(more goal-directed and less directly dependent on arousal or external stimuli, but not necessarily implying intentionality). Second, we must understand the underlying neural mechanisms of nonverbal vocal control and, crucially, determine whether the neural correlates involved in nonverbal vocal modulation in humans (including emotional vocalizations such as laughter) and in other primates, differ from those involved in volitional human speech production. Here we review initial work in these two areas that, together with the behavioral studies outlined above, suggests that control over $F_0$ and formants in our ancestors may have emerged to exaggerate socially relevant traits such as body size and threat potential even before the advent of articulated speech.

Vocal Flexibility in Other Mammals

Between- and within-individual differences in $F_0$ and formants play an important role in the social communication of many animals [10,11,28,29]. Several mammalian species scale or modulate their vocal frequencies in social interactions (Figure 1, Key Figure) (e.g., [30–32]). Although, in the examples given in Figure 1, formant modulation in nonhuman mammals appears largely linked to an external stimulus or arousal state, we suggest that such formant scaling nevertheless represents a potential path for the evolution of vowel articulation via an increasing complexity in formant modulation. In particular we argue that vocal exaggeration of apparent body size, which has been documented in animals ranging from red deer (Cervus elaphus) to humans (Figure 1), may have paved the way for more volitional forms of source and filter modulation and could have been the main driver in the evolution of vocal control [7,27].

Humans are highly unique in the advanced nature of our voice modulation abilities for two key reasons. First, we can voluntarily and independently control the source and filter properties of our vocalizations, and second we can perform these modulations in the complete absence of an associated inducing experience or state [33–35]. It is important to note that vocal control as defined above is a phenomenon distinct from vocal learning, which does not necessarily involve flexible manipulation of the vocal anatomy but rather entails the capacity to acquire or converge call types or entire vocal repertoires through imitation or learning [36,37]. The use of specific and pre-existing vocalizations in different or novel contexts, and the ability to respond differentially to the vocalizations of others through experience, is considered distinct from vocal production learning, namely because these behaviors are present in a broader range of animals and are likely to require comparatively less complex neural control [38]. These latter forms of learning are typically referred to as usage and contextual learning, respectively [39]. Like humans, many marine mammals and songbirds are capable of complex vocal learning. Indeed, vocal learning evolved independently in multiple lineages [34,38]. Until recently it was considered to be largely absent in nonhuman primates [40,41]; however, new evidence suggests that chimpanzees (Pan troglodytes) and orangutans (Pongo) may be capable of some forms of vocal learning [42–44] (cf. [41]).

Although human vocal control capabilities are unparalleled, below we present recent evidence that nonhuman primates may share our capacity to modulate $F_0$ and formants to perhaps a greater extent than previously thought. Other mammals, such as red deer, are capable of simple, uniform scaling of formants (Figure 1), but the comparative literature examining flexible vocal control has focused on nonhuman primates due to their relative intelligence and phylogenetic proximity to humans, and also the long-held notion that vocal control subserves spoken language acquisition [27,28].

Vocal Control in Nonhuman Primates?

Much of the evidence for control of $F_0$ and formants and imitation in nonhuman primate vocalizations has surfaced in the past 5 years, most of it from captive or enculturated great apes. For example, captive orangutans spontaneously imitate voiceless whistles produced by
Spectrograms Illustrating Progressive Functional Complexity in Formant Modulation Across Several Mammalian Species

Figure 1. The figure illustrates how formant scaling (formants F1–F4 labeled) may have first emerged in mammals for size exaggeration and threat displays, becoming increasingly complex over evolutionary time and ultimately allowing for the sophisticated formant modulation that characterizes articulated human speech. (A) Calls lacking formant modulation produced by a male bison (*Bison bison*), a baboon (*Papio hamadryas ursinus*), and a female rhesus macaque (*Macaca mulatta*). (B) Male red deer (*Cervus elaphus*, left), whose highly mobile larynges are positioned unusually low in the vocal tract [96], scale the formant spacing of their roars downward to project a large body size when encountering a threatening male competitor [30]; humans (*Homo sapiens*, right) scale formants similarly when instructed to sound ‘large’. (C) Diana monkeys (*Cercopithecus diana*) produce alarm calls in which the lowest two formants vary differentially in response to aerial versus terrestrial predators. Reproduced, with permission, from [31]; see also [32] in meerkats (*Suricata suricatta*). (D) Humans modulate relative formant positions, particularly F1 and F2, to produce different speech sounds; here we illustrate the differences in formant spacing between the vowels /i/ and /u/.
humans, indicating voluntary control over the lips and respiratory musculature [45], whereas chimpanzees produce novel and apparently flexible attention-seeking grunts toward humans [46]. Although this latter call type, with extended duration, is not produced in the wild, it demonstrates a latent capacity to control vocal fold adduction and airflow that is required to produce sustained laryngeal vibration, vocal behavior rarely observed in nonhuman primates. Many other primates also possess a rich repertoire of voiceless calls used in the wild, usually involving articulator manipulation without vocal-fold vibration homologous to human voiceless consonants [40]. Case studies of enculturated great apes even document some latent abilities to voluntarily control the vocal musculature, thus manipulating $F_0$ and individual formants [47–50], and in some cases producing novel rudimentary speech sounds beyond species-typical repertoires, with phonological features paralleling those observed in human speech [47,49,50].

There is mounting evidence for behavioral and contextual flexibility in great ape vocalizations. For instance, wild chimpanzees are capable of voluntary inhibition of affectively triggered food grunts [51], alarm calls [52], greetings [53], and responses to unfamiliar conspecifics [54]. The prefrontal cortex is thought to play a key role in such behaviors [55]. Chimpanzees also preferentially produce snake alarm calls in the presence of group members unaware of the threat [56]. These alarm calls appear flexible and potentially volitional [57], as do their food grunts [58] and joint-travel calls [59]. Despite occupying similar habitats, different populations of wild orangutans are now known to use different calls in the same context [44], as well as functionally arbitrary calls [60], suggesting a detachment between the production of the acoustic signal and physiological influence. Finally, among wild bonobos (Pan paniscus), ‘peeps’ and ‘contest hoots’ with the same acoustic structure are used in various social contexts [61,62], providing the first evidence of vocal signals used by a nonhuman primate to express a range of emotional states independent of context and biological function (i.e., functional flexibility).

These studies offer preliminary evidence that vocal adaptations important in human speech production are also present or latent in other mammals, and that among nonhuman primates manipulation of the larynx and vocal tract may be more flexible than once thought. In some cases, vocal control among nonhuman primates appears independent of affective state, an important prerequisite for articulated speech [55]. Studies of enculturated apes provide evidence that the potential for frequency modulation beyond simple, uniform scaling of formants exists in nonhuman primates, wherein great apes are able in some cases to manipulate individual formants to produce articulatory contrasts with speech-like rhythm and some apparent degree of voluntary, self-initiated control of the glottis and supralaryngeal vocal tract. However, critical work is now needed to determine whether nonhuman primates modulate $F_0$ and formants within call types to express different levels of aggression/subordination, as posited by several authors [7,12,27]. In this regard, mammals whose $F_0$ and formant scaling is clearly tied to communicating internal states or exaggerating physical traits, such as deer (Figure 1), may provide better convergent models for the study of early vocal modulation in humans.

A comprehensive understanding of vocal control also demands greater insight into the neural correlates of this behavior. Although neocortical control of the vocal apparatus during volitional speech production has traditionally been seen as setting humans apart from other primates [63,64], the behavioral studies reviewed above suggest that nonhuman primates may possess some cognitive control over vocal behavior. Unfortunately, the neural networks responsible for modulation of vocal motor systems in animals are poorly understood, and those involved in source and filter modulation are particularly understudied, even in humans. However, studies comparing volitional and spontaneous vocal production in humans can offer some insight.
Box 2. The Neuroanatomy of Human Vocal Control

The prevailing understanding of the neural systems controlling human vocal behavior owes much to extensive investigation of nonhuman primates such as the squirrel monkey (Saimiri), as well as to brain lesion and stimulation studies in human patients. A dual-pathway model is posited where midline and lateral motor systems are associated with the control of (innate) affective and learned vocalizations, respectively [55,63,97–99].

The production of vocal calls of various types is associated with a complex set of midbrain and brainstem structures including the basal ganglia, the periaqueductal grey (PAG), and brainstem nuclei such as the nucleus ambiguus (which houses motoneurons innervating the muscles of the larynx). The PAG has been implicated in the gating of vocalization patterns, receiving inputs on the motivational and affective states of the organism and activating a motor response. The neural initiation of innate calls and vocalizations (including, in humans, emotional sounds such as laughter and crying) implicates the frontal lobe in human and nonhuman primates and other mammals (specifically, the ACC).

Vocal flexibility sufficient for human speech requires the control of several source- and filter-related elements: (i) fine control of the extrinsic and intrinsic laryngeal muscles; (ii) regulation of airflow from the lungs through the oral and nasal cavities; and (iii) the capacity for precise configuration of the supralaryngeal articulators (lips, tongue, jaw, soft palate) [97]. Numerous studies, ranging from cortical stimulation to functional imaging, have demonstrated evidence for somatotopic control of the articulators on the human precentral gyrus (e.g., [100–102]). The direct corticobulbar hypothesis argues that the flexibility of human vocal behavior, such as the rapid switching on and off of laryngeal signals, can be attributed to the presence of direct connections from M1 to brainstem motoneurons, which are indirect (via the reticular formation) in nonhuman primates. Interestingly, the first functional imaging study of cortical control of the human larynx implicated a more dorsal portion of the precentral gyrus than the ventral sites described for monkeys [103]. This alternative location, adjacent to both the articulatory and respiratory control centers (see also [104]), might facilitate coordinated processes subserving connected and coarticulated speech (Figure 2).

Controlled speech production implicates higher-order structures such as Broca’s area in the inferior frontal gyrus; this is argued to host a mental ‘syllabary’ important in the selection of learned motor programs [105]. Auditory and somatosensory representations have also been implicated in self-monitoring and error correction during voluntary vocal behavior (e.g., [106,107]).

Neural Correlates of Volitional and Spontaneous Vocal Control in Humans

The advent of functional neuroimaging (PET and fMRI) has brought with it the possibility to investigate the systems regulating speech and voice control in the intact brain and thus has yielded substantial advances in our understanding of human vocal production. This has led to the development of neurobiological models implicating a cortical network of inferior frontal, auditory, and sensorimotor regions in the selection and articulation of spoken syllables and the use of sensory feedback to guide output processes (e.g., [65]). Although the literature on mammalian vocalizations suggests two pathways for vocal control (i.e., the dual-pathway model; Box 2 and Figure 2), difficulties in imaging small and deep midbrain and brainstem centers with fMRI means that many existing accounts of vocal behavior in humans focus on cortical, and predominately perisylvian, sensorimotor systems. Recently, some authors have argued for greater interaction between the dual paths in human speech, positing meaningful involvement of subcortical regions (the basal ganglia) in the affective and motivational modulation of laryngeal speech signals [63,66,67]. There are numerous sources of evidence that challenge the independence of these two routes, or at least suggest a greater complexity to the system (see, for example, [68,69]), including evidence that patients with lesions in motor cortical regions retain the ability to swear [70].

Yet in the same way that the behavioral literature has largely neglected the para- or ‘supra’-linguistic aspects of flexible and voluntary human vocal behavior, so has the field of cognitive neuroscience. As a result, most studies have focused on what is being said (i.e., choice of phonemes, syllables, and words) and a very small number of studies have specifically focused on how speech is produced, or flexibly controlled (i.e., pronunciation, indexical cues, and affective content). Below we review this modest but growing area of research.

The investigation of novel speech sound imitation and pronunciation has revealed variation in both the activation and the structure of typically left-hemisphere frontal sites (e.g., inferior frontal
cortex, insula) as well as regions of the inferior parietal cortex (e.g., Rolandic operculum, supramarginal gyrus) and the superior temporal cortex, all associated with individual differences in task performance [71–74]. Studies have also highlighted the role of subcortical structures in voluntary phonetic learning contexts, where the dorsal striatum (bilateral caudate and putamen) is involved in initial attempts to imitate a novel sequence of sounds [74]. One experiment showed that the left anterior insula and inferior frontal gyrus were engaged during voluntary impersonations of speaking styles, accents, and familiar individuals [75]. A similar study showed increased responses in the left anterior insula (as well as the supplementary motor area, the anterior cingulate cortex (ACC), and regions of the superior temporal cortex) when participants voluntarily adopted a question intonation for a train of nonsense syllables [76].

Similar to research on speech production, work on the neural correlates of song – wherein singing often involves significant voice frequency modulation (see, for example, [14]) – has identified four key neural substrates of auditory–motor control, for example in the production and correction of melodic pitch: the auditory cortex, the intraparietal sulcus, the anterior insula, and the ACC [77]. Researchers have reported experience-dependent differences in the engagement of brain regions such as the anterior insula and sensorimotor cortices associated with the control of breathing (see [78]) and increased use of sensory feedback during vocal performance [79,80]. A recent study of vocal pitch processing showed greater engagement of the basal ganglia (bilateral putamen) in the imitation of melodies compared with non-imitative productions and melody perception [67]. The authors of that study argue that establishing how these subcortical

Figure 2. Main Brain Regions Implicated in Human Vocal Control. According to the dual-pathway model, sensorimotor cortical systems (and the cerebellum, shaded orange) support the production of learned vocalizations such as speech and song and involve direct innervation, whereas the limbic system in the anterior cingulate cortex (ACC, blue) is responsible for the initiation of innate vocalizations such as laughter. Several studies have identified an important role for the dorsal striatum (green) in vocal behavior, including foreign language learning (e.g., [74]). A first direct investigation of the control of the laryngeal muscles in humans identified a more dorsal location than had been observed in monkeys [103]. We argue that this represents an important evolutionary change in the control of vocalizations in humans, potentially supporting the emergence of articulated speech.
Box 3. The Special Case of Laughter: Novel Insight from an Animal-like Human Vocalization

Human laughter is ubiquitous in interpersonal interactions and serves various emotional and social signaling functions [108]. Importantly, spontaneous (i.e., authentic) laughter can be elicited in the laboratory with relative ease and can be compared with the volitional production of laughter (on demand). These two forms of laughter are developmentally, acoustically, and perceptually distinct [109–111] (Figure I), suggesting different production mechanisms that may align with the midline and lateral pathways outlined in Box 2.

Among humans, the hypothalamus appears to be more active during tickling laughter than during volitional, prompted laughter [68]. This suggests a potential role for the hypothalamus in conveying motivational and affective information to the PAG in the implementation of laughter motor patterns. The same study found ACC activation during volitional laughter and during the suppression of laughter during tickling, but not during spontaneous laughter, implicating the ACC in volitional vocal control [68]. Nonetheless, spontaneous tickling still engages a range of lateral sensorimotor regions equivalent to activation during volitional laughter; indeed, these appear even more strongly activated when tickling laughter is actively suppressed [88]. This finding suggests fairly continuous involvement of newer cortical control systems in voice production, even for affective and innate vocalizations, and argues against the rather clear demarcation of neural systems suggested in the literature (see also Box 2).

Spontaneous human laughter is argued to be similar in form and function to tickling-induced play vocalizations in other primates [112]. Both are often produced without voicing [113] and when spontaneous (but not volitional) human laughs are slowed down and pitch proportionally adjusted, they are largely indistinguishable from nonhuman primate vocalizations [109]. Spontaneous human laughter and homologous macaque and chimpanzee vocalizations are characterized by similar interval duration, serial organization, and high intra-bout variability in acoustic parameters [114]. The presence of regular vocal fold vibration and consistently egressive airflow in some ape laughter vocalizations – call characteristics previously described as markers of human laughter and speech – thus indicates homology in primate vocal control mechanisms [115].

In addition to spontaneous laughter, adult (but not infant) chimpanzees produce acoustically distinct (e.g., shorter) laugh replications in response to the laughter of conspecifics, with a comparatively delayed response latency [115], implicating at least some degree of non-automatic vocal control [108,110,115]. Hence, chimpanzees’ capacity to imitate or modulate laughter appears highly similar to volitional human laughter in both developmental trajectory and its function in social cohesion.

Figure I. Human Spontaneous Amusement Laughter (Left) and Volitional Laughter (Right) Produced by the Same Man. Volitional control of laughter as a normative social signal is refined throughout human development whereas spontaneous laughter emerges early in infancy and appears innate, and relatively more animal-like, in its acoustic structure. As illustrated by these waveforms and spectrograms, bouts of spontaneous laughter tend to be longer in overall duration with shorter individual calls, higher fundamental frequency (F0) (plotted in blue), and differing spectral profiles compared with volitional laughter [109,110]. Nonlinearities in spontaneous laughter vocalizations (e.g., wheezing, glottal whistles) reflect reduced control of vocal behavior during authentic emotional experience [110,116].
sites functionally connect with cortical structures is crucial to understanding how they assist in the auditory-to-motor transforms necessary for imitative behavior.

Other studies have examined the voluntary production of emotional inflections in speech to investigate the influence of limbic regions on vocal processes. A comparison of happy and neutral production of syllables revealed increased activation in many cortical and subcortical regions, crucially including the anterior cingulate and medial temporal cortex as well as the thalamus and midbrain [76]. Recent studies have linked involvement of the ACC and the amygdala to autonomic arousal during the production of angry speech vocalizations [81]. The induction of sad (but not happy) affect also predicts activation of the ACC as well as changes in F0 range [82]. A recent study argued for a more general role for the anterior cingulate in the control of pitch modulations by showing equivalent activation of this region (and of the laryngeal motor cortex) in the production of both volitional affective and non-affective syllables. These functional imaging studies have thus revealed clear roles for neocortical sensorimotor systems in the flexible control of human speech and song and our underlying relative expertise in these behaviors. However, reports of additional involvement of the limbic and striatal regions – part of an evolutionarily older control system (Box 2) – in comparatively demanding contexts requiring voluntary modulation of emotional prosody or in phonetic learning are consistent with the idea that both vocal pathways play a role in voice modulation.

Crucially, then, current evidence from neuroimaging research in humans seems to suggest that the limbic and cortical vocal pathways (Figure 2) are highly interactive and that both cortical and subcortical routes are involved in volitional and spontaneous vocal production [66]. Moreover, vocal flexibility in nonhuman primates suggests that other species have greater neuroanatomical elaboration of the direct lateral motor cortical route than previously thought or, alternatively, may be achieving flexibility with older neural structures (e.g., ACC, limbic system). Although there are issues associated with establishing genuine spontaneity in laboratory settings, laughter may offer one of the most promising signals with which to interrogate the influence of the different neural systems controlling vocal output: it allows us to directly compare what happens in the brain during volitional and spontaneous vocal production in humans and also to compare spontaneous vocalizations produced by humans and nonhuman primates (Box 3).

Concluding Remarks and Future Directions
Volitional and flexible control of the anatomical structures involved in speech production (the larynx producing F0 and the supralaryngeal vocal tract producing formants) has long been thought to set humans apart from other vocalizing animals and is generally considered a precursor to human speech acquisition and production. We have reviewed preliminary behavioral and neural evidence to suggest that the capacity for vocal control is not uniquely human but present to some degree in other mammals, including nonhuman primates (Box 4).

We suggest that flexible scaling of source–filter components of vocalizations (F0 and formants) may provide a path for the evolution of vocal control, wherein control of these acoustic components may have emerged in our ancestors to express, exaggerate, or even fake physical traits or internal motivational states. The fact that many animal species and humans from distinct cultures are capable of scaling their formants to exaggerate their apparent size imparts the possibility that formant scaling related to size/threat exaggeration (and perception) may constitute a precursor of more complex formant modulation involved in speech articulation [27]. As illustrated in Figure 1, linear formant scaling as observed in red deer roars, or slightly more complex formant modulations as observed in the predator calls of diana monkeys, effectively communicating size-related information, could have opened the way for increasingly greater vocal control, ultimately leading to full-blown articulated human speech. The sound-symbolic properties of high front-vowel sounds associated with smallness (such as /i/, in which F2 is
The current neuroimaging literature on vocal control is representative of a subfield influenced by the historical motivations of the linguistics and neuropsychology (i.e., aphasia) literatures, focusing largely on speech and language production. Thus, we are still lacking studies that primarily and systematically investigate the neural control of core source-filter modulations such as those that may have facilitated exaggeration of body size and cues to sex and identity in our preverbal evolutionary past [4,28] and those used ubiquitously in the voluntary vocal communica-
tion of attitude, emotional state, and even aspects of personality [84,85]. Hence, future work should make greater attempts to link neural activation data to the underlying articulatory dynamics and acoustic content of vocal behaviors, including imitation. Real-time anatomical imaging of the vocal tract using MRI (up to 80 frames/s [86]) offers the opportunity to image naturalistic vocal behavior while also measuring neural activation with echoplanar imaging of the brain. Using amodal analysis approaches such as representational similarity analysis (RSA) [87], it is possible to probe the neural representation of articulatory dynamics.

There exists an opportunity to strengthen the links between the behavioral and neural imaging literatures and form a more coherent cognitive neuroscience of the voice. We therefore encourage colleagues across disciplines to engage in research that directly examines voice modulation as a signal with tremendous social, as well as linguistic, significance. In our quest to uncover the origins of speech, understanding the behavioral functions and neural mechanisms of both involuntary and voluntary nonverbal modulation in humans, and looking for emerging equivalents in other mammals, is likely to be our best way forward (see Outstanding Questions).

Acknowledgments

The authors thank Tecumseh Fitch, Drew Rendall, and Megan Wyman for sharing vocal recordings used in the creation of Figure 1.


124. Sapir, E. (1929) A study in phonetic symbolism. J. Exp. Psychol. 12, 225