CellPress

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 Reby[1,*]

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.

Trends

Source–filter theory is the unifying methodological framework in the study of nonverbal vocal communication in humans and other vertebrates.

Numerous studies have implicated fundamental frequency (*F*0) and vocal tract resonances (formants) in mammalian social communication. In humans, these sexually dimorphic source and filter voice features reliably indicate sex, age, body size, and dominance.

By focusing on static rather than dynamic vocal processes, this literature has largely overlooked the human capacity to volitionally modulate *F*0 and formants to express or exaggerate ecologically relevant traits, often in response to specific social contexts.

Emerging research suggests that *F*0 and formant scaling is widespread in humans and present in some other mammals, potentially representing an evolutionary path from simple voice-frequency scaling 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 (*F*0)** (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 *F*0 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

[1]Mammal Vocal Communication and Cognition Research Group, School of Psychology, University of Sussex, Brighton, UK
[2]Institute of Psychology, University of Wrocław, Wrocław, Poland
[3]Royal Holloway Vocal Communication Laboratory, Department of Psychology, Royal Holloway, University of London, Egham, UK

*Correspondence: reby@sussex.ac.uk (D. Reby).

 http://dx.doi.org/10.1016/j.tics.2016.01.002

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 *F*0, 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 *F*0 and perceived pitch; however, regardless of mass, *F*0 increases when the vocal folds are stretched and become tenser [16,17]. Thus, the effective mass, length, and tension of the vocal folds affect *F*0. 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 IA,B). Both *F*0 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 *F*0 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 *F*0 [92,93].

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 *F*1 and *F*2 [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 *F*0. 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.

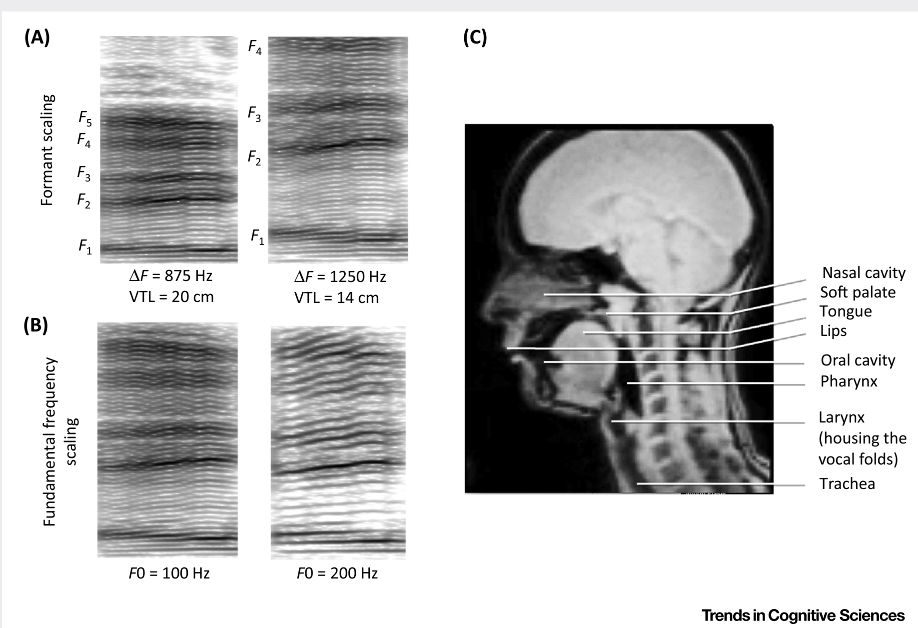

Figure I. Voice Frequency Scaling. Each spectrogram illustrates the vowel /el/ spoken by the same man, whose voice frequencies were manipulated using Praat acoustic software [95]. (A) Formant scaling, wherein lower, more closely spaced formants (left) correspond to a relatively longer vocal tract and a perceptually more resonant voice. Apparent vocal tract length (VTL) was estimated from ΔF; for algorithms see [91,92]. (B) Fundamental frequency (*F*0) scaling, wherein a lower *F*0 and more closely spaced harmonics (multiple integers of *F*0) (left) correspond to a perceptually lower-pitched voice. Note when comparing panels (A) and (B) that formants and *F*0 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.

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 (*F*0): 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 *F*0. Voice pitch is the perceptual correlate of *F*0 (along with its harmonics, integer multiples of *F*0).

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 *F*0) or the supralaryngeal vocal tract (affecting the production of formants) in a flexible or voluntary manner, for