EVOLUTIONARY BIOLOGY

Comment on "Monkey vocal tracts are speech-ready"

Philip Lieberman

Monkey vocal tracts are capable of producing monkey speech, not the full range of articulate human speech. The evolution of human speech entailed both anatomy and brains. Fitch, de Boer, Mathur, and Ghazanfar in *Science Advances* claim that "monkey vocal tracts are speech-ready," and conclude that "…the evolution of human speech capabilities required neural change rather than modifications of vocal anatomy." Neither premise is consistent either with the data presented and the conclusions reached by de Boer and Fitch themselves in their own published papers on the role of anatomy in the evolution of human speech or with the body of independent studies published since the 1950s.

INTRODUCTION

The vowel inventory reported by Fitch and his colleagues (*1*) has a wider range than the 1969 Lieberman, Klatt, and Wilson study (*2*) owing to their modeling vocal tract shapes computed from cineradiographs. However, their studies nevertheless replicate the 1969 paper's principal finding: Monkey vocal tracts are incapable of producing the full range of human speech, including the "quantal" vowels [i], [u], and [a] (the vowels of the words "see," "do," and "ma"). Independent studies over the course of more than 70 years show that these vowels, which are present in virtually all human languages (*3*, *4*), contribute to the robustness of human speech as a medium of vocal communication (*5*–*9*). Fitch *et al.* (*1*) confirm that the monkey vocal tract is limited to the nonquantal vowels of the English words "bit," "bet," 'bat," and "bought." If monkeys had brains capable of learning and executing the motor commands involved in human speech, their "monkey speech" would not be as robust a means of vocal communication as that of fully modern human beings. Moreover, the limits imposed by neural control on speech production were noted in the 1969 paper (*2*), which was the second of a series that explored the limits placed by anatomy and brains in the evolution of human speech. We noted that the 1968 acoustic analysis of chimpanzee and monkey vocalizations (*10*) showed that they did not make use of the full possibilities of their vocal tracts.

The Lieberman and Crelin 1971 paper "On the speech of Neanderthal man" (*11*) used similar computer modeling techniques to estimate the acoustic consequence of vocal tract shapes based on a reconstruction of a Neandertal vocal tract. The reconstructed adult Neandertal vocal was similar to that of a large human newborn infant, and thus, the range of possible vocal tract shapes was guided by cineradiographs of newborn infant cry (*12*). The computed vowel range was virtually identical to that reported by Fitch *et al.* (*1*) and again did not include the quantal vowels [i], [u], and [a]. However, in light of the archeological record, we concluded that Neandertals had brains that could learn and execute the complex motor acts involved in talking and hence had language and could talk, albeit with less clarity than modern humans.

FORMANT FREQUENCIES

Because it is apparent that the relevant references may not be familiar to many readers, some background material may be helpful. The phonetic qualities of vowels are, in part, determined by their "formant frequencies." The supralaryngeal vocal tract (SVT), the airway above the larynx, acts as a malleable acoustic filter on the acoustic energy produced at the larynx, producing potential local energy maxima in the acoustic frequency spectrum at these formant frequencies. The musical notes produced by a pipe organ provide a rough analogy to this aspect of speech production. The quality of each musical note is determined by the length and shape of an organ pipe. Hence, if the range of vocal tract shapes that an animal or hominin can produce can be determined, the range of possible vowels can be modeled by constructing appropriately shaped pipes, or through computations that calculate the formant frequencies that a particular vocal tract shape will produce (*13*), or by means of appropriate computer-implemented algorithms (*14*). Consonant formant frequencies can also be determined, but because vowel vocal tract shapes can be held steady, the computational problems are simpler.

QUANTAL VOWELS AND VOCAL TRACT ANATOMY

The anatomy necessary for producing the quantal vowels and consonants that confer articulate human speech was determined by Stevens in his classic 1972 study (*5*). Stevens, using both computer modeling techniques and tubes that replicated the vocal tract shapes of quantal vowels, showed that the species-specific human tongue played a key role in the evolution of human speech. In human adults, half of the tongue, its "horizontal" segment, SVTh, rests in the mouth, whereas its vertical segment, SVTv, rests in the pharynx within the neck. The horizontal and vertical segments meet at an approximate right angle that permits the production of discontinuities in the cross-sectional area of the vocal tract that is necessary to produce the unique properties of quantal vowels. Two formant frequencies converge when producing a quantal vowel, yielding a spectral peak, analogous to a saturated color—other vowels being pastel-like owing to the absence of spectral peaks. At the same time, quantal vowels are relatively insensitive to small errors in tongue placement.

Human infants begin life with vocal tracts similar to those of apes and monkeys. Their tongues are "flat," positioned almost entirely in their mouths. When breathing, human infants, like most mammals, have a "patent" airway isolated from the pathway used to ingest solids and liquids. Through a gradual remodeling process over the course of the first 6 to 8 years of life, probably regulated by species-specific epigenetic factors, the human skull is remodeled and the human tongue changes its shape, descending into the pharynx as the neck lengthens, pulling the larynx down with it. At the end of this process, approximately one-half of the tongue—its "horizontal" portion, SVTh—rests in the mouth, whereas the "vertical" half, SVTv, rests in the pharynx.

Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912, USA.
Corresponding author. Email: philip.lieberman@gmail.com

At puberty in human males, the length of SVTv often increases somewhat (*15*).

PREVIOUS STUDIES OF DE BOER AND FITCH

The 2016 computer modeling study by Fitch and his colleagues (*1*), which was guided by cineradiographic data, shows a greater range of formant frequencies than the 1969 study (*2*) but replicates its central finding that the monkey vocal tract cannot produce the quantal vowels [i], [u], and [a]. That is not surprising because de Boer and Fitch themselves concluded in their 2010 study, "Computer models of vocal tract evolution: An overview and critique" (*14*), that Stevens's 1972 study was correct. Their discussion of Stevens's theory paraphrased that in my 1984 book, *The Biology and Evolution of Language* (*6*): The shape and proportions of the adult human tongues allow humans to produce articulate speech. As de Boer and Fitch also noted, independent studies came to the same conclusion.

Moreover, the 2010 paper by de Boer and Fitch (*14*) addressed the issue discussed here—the phonetic qualities of the vowels that a specific vocal tract could produce. Their paper discussed the ongoing debate concerning Neandertal capabilities. Many people who apparently never read the 1971 paper by Lieberman and Crelin (*11*) believe that it showed that Neandertals could not talk. However, we instead concluded that Neandertals had language and could talk, albeit with less clarity than modern humans (*12*). de Boer and Fitch correctly pointed out that "at issue, then, is not the question of whether a Neanderthal could speak. … but instead about the specific phonetic characteristics of the speech they could produce." The situation concerning monkey speech is similar; their phonetic repertoire is limited by their anatomy. Monkeys cannot produce the quantal vowels that contribute to the robustness of human speech. In the authors' 2016 paper, the formant frequency tracks in Fig. 4B for the macaque monkey version of the human speaking "Will you marry me" do not show the formant frequency sweeps and convergences apparent in the quantal human speech formant frequency plots in Fig. 4A (*1*). If monkeys had human brains, they would instead produce inarticulate monkey speech.

Paradoxically, de Boer's own published papers specifically dispute the claim made by Fitch *et al.* (*1*) that "the evolution of human speech capabilities required neural change rather than modifications of vocal anatomy." In de Boer's study comparing the speech capabilities of women with men (*16*), he states that his "results agreed with the hypothesis that modern human vocal anatomy has evolved because of speech." Using modeling techniques similar those in the study by Fitch *et al.* (*1*), de Boer noted that "…women's vowels were generally more 'quantal' because their vocal tracts better conform to the 1:1 oral to pharyngeal proportions noted by Stevens… and hence were more 'quantal.'" His paper concluded that "vocal tracts with approximately human anatomy and control over the tongue and jaw can produce the largest range of speech sounds if the vertical part is approximately as long as the horizontal part. This confirms the conjecture by Lieberman et al. (1972)" (*17*). In a second paper, de Boer's 2012 computer modeling study of the effects of air sacs present in chimpanzee vocal tracts (*18*), de Boer again concluded that the evolution of species-specific human vocal anatomy played a role in the evolution of human speech. de Boer showed that the absence of air sacs in human vocal tracts results in speech that is less easily confusing than the vocalizations that chimpanzee vocal tracts could potentially produce.

However, the most peculiar aspect of the 2016 *Science Advances* paper by Fitch *et al.* (*1*) is that they claim that their 2010 paper (*14*) that critiqued the computer modeling studies of Boë *et al.* instead "show[s] that the importance of human vocal tact anatomy for speech has been overestimated (24–26)." Their references 24 to 26 refer to their 2010 paper (*14*) and the computer modeling studies of Boë *et al.* (*19*, *20*). The abstract of the 2010 paper by de Boer and Fitch (*14*) instead states that the computer model used by the Boë group "…contains subtle but fatal flaws which invalidate the conclusions drawn from the model." Bart de Boer and Fitch in their 2010 paper concluded that "despite their strongly worded claims, the results of Boë and colleagues' simulations (*19*, *20*) do not, and in principle, cannot, demonstrate that vocal tract anatomy is 'irrelevant' to human speech production." The 2016 paper by Fitch *et al.* (*1*) has falsified the findings and conclusions of the 2010 paper by de Boer and Fitch that showed that the Boë modeling technique was meaningless and that, moreover, it endorsed the view that the species-specific human tongue and vocal tract played a major role in the evolution human speech.

THE HUMAN TONGUE AND VOCAL TRACT NORMALIZATION

Other errors and mischaracterizations mark the 2016 paper by Fitch *et al.* (*1*). The reference to the descent of the tongue in nonhumans is, as de Boer and Fitch (*14*) themselves pointed out, irrelevant—tongue shape and position instead characterize the species-specific human vocal tract. The human tongue descends because it is anchored to the root of the tongue, which moves down into the pharynx over the course of 6 to 8 years after birth. In the animals studied by Fitch and his colleagues, the tongue remains anchored in their mouths, whereas the larynx can either temporarily or permanently descend. The descent of the tongue does not increase the animals' vowel repertoire (*21*). The studies by Nishimura *et al.* (*22*, *23*) show that monkey fetal vocal tract morphology changes by birth to that noted in the 1969 study and has nothing to do with the later development of the vocal tract in human infants and young children. Contrary to the 2016 study by Fitch *et al.* (*1*), the vowel [i] (the vowel of the word "see") has a special status, providing an optimal cue to determine the length of a speaker's vocal tract. The formant frequencies that play a major role in specifying vowels and consonants are the frequencies at which local energy maxima can pass through a speaker's vocal tract, in much the same way as the frequencies at which maximum energy is generated by a pipe organ to determine musical notes. A shorter pipe produces a higher note. However, in contrast to listening to musical notes, when we hear a child speaking, we unconsciously take into account the child's shorter vocal tract length, "normalizing" it to recover the vowel or consonant that the child intended to produce—though its absolute frequencies are higher than any normal adult's. The vocal tracts of different adults also differ in length, and both human listeners and computer-implemented speech recognition systems must take normalization into account to correctly interpret the acoustic signal.

The 1952 Peterson and Barney study (*24*) directed at exploring the parameters for machine-implemented speech recognition first noted the special status of the vowel [i]. Panels of human listeners were placed in the position of a computer system that had to identify syllables having the form [hVd], presented in pseudorandom order as spoken by 76 different speakers who spoke different dialects of American English (*24*). In other words, without knowing who was talking or what was being said, they had to identify the vowels that differentiated these syllables. Out of more than 10,000 trials, two errors occurred for [i], slight confusion occurred for [u], and hundreds

of errors occurred for the vowels that a monkey vocal tract is capable of producing. Nearey in 1978 explained why the quantal vowel [i] was correctly identified in this situation, one that approximates the normal uses of speech (*25*). Nearey's analysis of cineradiographs of adult speakers speaking showed that speakers who had different SVT lengths can produce the same formant frequencies for nonquantal vowels such as [ae] by alternate SVT shapes. A speaker having a shorter SVT could, by protruding and/or constricting his lips, produce an [ae] that a speaker having a longer SVT normally would produce. In contrast, an [i] can only be produced by a particular SVT shape, thereby providing an optimal cue to determine the length of a speaker's SVT. In a speech perception task in which listeners were presented with a range of formant frequencies that would specify English nonquantal vowels produced by speakers who had different vocal tract lengths, Nearey showed that a listener hearing a single example of an [i] "recalibrated" his or her acoustic to phonetically "map" to the estimated SVT length provided by the [i]. The vowel [ae] cannot serve as an acoustic vocal tract length calibrating signal as Fitch *et al.* (*1*) claim.

CONCLUSION

In short, the evolution of human speech entailed both brains that could learn and execute voluntary complex acts and anatomy that enabled the production of the full range of human speech. Negus in 1949 speculated on the role and evolution of the human vocal tract, which he thought was not present in Neandertals and earlier extinct hominins (*26*). However, Negus was not the first person who commented on the species-specific morphology of the human supralaryngeal airway. Charles Darwin in *On the Origin of Species* repeatedly stated that small selective advantages drove the course of evolution. In 1859 in the first edition (*27*), he pointed out,

"The strange fact that every particle of food and drink which we swallow has to pass over the orifice of the trachea, with some risk of falling into the lungs…."

The species-specific anatomy that enables humans to produce the full range of quantal vowels, enhancing the robustness of speech, accounts for choking on food, which remains the fourth leading cause of accidental death in the United States (*15*).

REFERENCES AND NOTES

1. W. T. Fitch, B. Boer, N. Mathur, A. A. Ghaazanfar, Monkey vocal tracts are speech-ready. *Sci. Adv.* **2**, e1600723 (2016).
2. P. H. Lieberman, D. H. Klatt, W. H. Wilson, Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science* **164**, 1185–1187 (1969).
3. J. Greenberg, *Universals of Human Language* (MIT Press, 1963).
4. P. Maddiesson, *Patterns of Sounds* (Cambridge Univ. Press, 1984).
5. K. N. Stevens, Quantal nature of speech, in *Human Communication: A Unified View*, E. E. David Jr., P. B. Denes, Eds. (McGraw Hill, 1972), pp. 51–66.
6. P. Lieberman, *The Biology and Evolution of Language* (Harvard Univ. Press, 1984).
7. P. Lieberman, *Human Language and Our Reptilian Brain: The Subcortical Bases of Speech, Syntax, and Thought* (Harvard Univ. Press, 2000).
8. P. Lieberman *Toward an Evolutionary Biology of Language* (Harvard Univ. Press, 2006).
9. P. Lieberman, The evolution of language and thought. *J. Anthropol. Sci.* **94**, 127–146 (2016).
10. P. Lieberman, Primate vocalizations and human linguistic ability. *J. Acoust. Soc. Am.* **44**, 1574–1584 (1968).
11. P. Lieberman, E. S. Crelin, On the speech of Neanderthal man. *Linguist. Inq.* **2**, 203–222 (1971).
12. H. L. Truby, J. F. Bosma, J. Lind, *Newborn Infant Cry* (Almquist and Wiksell, 1965).
13. T. Chiba, J. Kajiyama, *The Vowel: Its Nature and Structure* (Tokyo-Kaisekan Publishing, 1941).
14. B. de Boer, W. T. Fitch, Computer models of vocal tract evolution: An overview and critique. *Adapt. Behav.* **18**, 36–47 (2010).
15. D. E. Lieberman, *The Evolution of the Human Head* (Harvard Univ. Press, 2011), pp. 281–339.
16. B. de Boer, Modelling vocal anatomy's significant effect on speech. *J. Evol. Psychol.* **8**, 351–366 (2010).
17. P. Lieberman, E. S. Crelin, D. H. Klatt, Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *Am. Anthropol.* **74**, 287–307 (1972).
18. B. de Boer, Loss of air sacs improved hominin speech abilities. *J. Hum. Evol.* **62**, 1–6 (2012).
19. L.-J. Boë, J.-L. Heim, K. Honda, S. Maeda, The potential Neandertal vowel space was as large as that of modern humans. *J. Phon.* **30**, 465–484 (2002).
20. L.-J. Boë, S. Maeda, J.-L. Heim, Neanderthal man was not morphologically handicapped for speech. *Evol. Commun.* **3**, 49–77 (1999).
21. P. Lieberman, Vocal tract anatomy and the neural bases of talking. *J. Phon.* **40**, 608–622 (2012).
22. T. Nishimura, T. Oishi, J. Suzuki, K. Matsuda, T. Takahashi, Development of the supralaryngeal vocal tract in Japanese macaques: Implications for the evolution of the descent of the larynx. *Am. J. Phys. Anthropol.* **135**, 182–194 (2008).
23. T. Nishimura, A. Mikami, J. Suzuki, T. Matsuzawa, Development of the laryngeal air sac in chimpanzees. *Int. J. Primatol.* **28**, 483–492 (2007).
24. G. E. Peterson, H. L. Barney, Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **24**, 175–184 (1952).
25. T. Nearey, *Phonetic Features for Vowels* (Bloomington Indiana University, 1978).
26. V. Negus, *The Comparative Physiology and Anatomy of the Larynx* (Haffner, 1949).
27. C. Darwin, *On the Origin of Species* (Harvard Univ. Press, facsimile edition 1859, 1964), p. 191.

Acknowledgments

Funding: No funding was involved. **Author contributions:** P.L. conceived the study and wrote the manuscript. **Competing interests:** The author declares that he has no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and in the references cited. Additional data related to this paper may be requested from the author.

Submitted 9 February 2017
Accepted 31 May 2017
Published 7 July 2017
10.1126/sciadv.1700442

Citation: P. Lieberman, Comment on "Monkey vocal tracts are speech-ready." *Sci. Adv.* **3**, e1700442 (2017).