HUMAN EVOLUTION

Monkey vocal tracts are speech-ready

W. Tecumseh Fitch,[1,2]* Bart de Boer,[3] Neil Mathur,[4,5] Asif A. Ghazanfar[4,5,6]*

For four decades, the inability of nonhuman primates to produce human speech sounds has been claimed to stem from limitations in their vocal tract anatomy, a conclusion based on plaster casts made from the vocal tract of a monkey cadaver. We used x-ray videos to quantify vocal tract dynamics in living macaques during vocalization, facial displays, and feeding. We demonstrate that the macaque vocal tract could easily produce an adequate range of speech sounds to support spoken language, showing that previous techniques based on postmortem samples drastically underestimated primate vocal capabilities. Our findings imply that the evolution of human speech capabilities required neural changes rather than modifications of vocal anatomy. Macaques have a speech-ready vocal tract but lack a speech-ready brain to control it.

INTRODUCTION

Despite repeated attempts, no nonhuman primates have ever been trained to produce speech sounds, not even chimpanzees raised from birth in human homes (*1*). Humans appear to be the only primates with a capacity to flexibly control their vocalizations and to integrate respiration, phonation, and vocal tract movements in an intricate manner as required for speech (*2*–*4*). Since Darwin's time, two hypotheses have been considered to be the likely explanations for this fact. The first "neural" hypothesis is that other primates lack the brain mechanisms required to control and coordinate their otherwise adequate vocal production system; Darwin favored this hypothesis, and it was widely accepted until the 1960s (*5*). The second "peripheral" hypothesis, in contrast, identifies the basis of primate vocal limitations as the anatomy and configuration of the nonhuman primate vocal tract. This hypothesis is widely accepted today, largely due to a seminal 1969 *Science* paper by Lieberman *et al.* (*6*), which used a computer program to explore the phonetic capability of a rhesus macaque and, by extension, other nonhuman primates. They concluded that "the vocal apparatus of the rhesus monkey is inherently incapable of producing the range of human speech" [(*6*), p. 1187]. Later work used the same methods and reached the same conclusions for chimpanzees (*7*), and thus inaugurated the reign of the "peripheral" hypothesis, which today remains a widely accepted "textbook fact" concerning human speech (*8*–*13*). For example, "early experiments to teach chimpanzees to communicate with their voices failed because of the insufficiencies of the animals' vocal organs" (*9*). This now-traditional hypothesis has an important implication for the evolution of human language: that the broad phonetic range used in modern human speech required key changes in peripheral vocal anatomy during recent human evolution. Here, we present new data, based on x-ray images from living monkeys, that sharply challenge this hypothesis and thus its implication concerning language evolution.

Lieberman and colleagues (*6*) first made a plaster cast of the oral cavity of a rhesus macaque cadaver and sectioned it to derive an estimate of its resting shape. They then created a computer model of the monkey vocal tract, roughly estimating its boundary conditions by manipulating the tongue of an anesthetized animal, and finally explored the possible acoustic range of this computer model to outline the possible vowel range of a monkey. The use of a computer model is appropriate because animals do not necessarily exploit the full phonetic range of their vocal apparatus when vocalizing, and recordings of their vocalizations thus indicate only what the animal does, rather than what it could do. However, a computer model based on a plaster cast of the vocal tract of a monkey cadaver does not necessarily provide an adequate indication of the range of vocal tract shapes produced in living animals. More recent research suggests that investigations of postmortem anatomy drastically underestimate the flexibility of the mammalian vocal tract (*14*, *15*).

To remedy this inadequacy, we first examined the vocal anatomy of behaving macaques using x-ray videos. To gain a full estimate of the dynamic flexibility in the vocal tract, we examined the configuration of the upper respiratory tract not only during vocalization but also during facial displays (such as lip smacking and teeth chattering) and during feeding and swallowing. On the basis of the tracings of 99 observed vocal tract configurations, we next constructed a computer model of the macaque vocal tract. Crucially, we never extrapolated beyond the observed anatomical range: Every data point is based on an actual observed configuration. We then used a maximization approach to choose—from the observed vocal tract configurations—five maximally distinct "monkey vowels" that make the best use of the observed space and synthesized these vowels using a monkey grunt vocalization as a source signal. These monkey vowels were finally played in a discrimination test to human listeners to evaluate the listeners' ability to discriminate among the five monkey vowels. We also used a nearest-neighbor approach to find the closest approximation to various human vowels producible by the monkey phonetic model (scaled for differences in overall vocal tract length). This approach provides a highly conservative estimate of potential acoustic output: Only macaque vocal tract configurations we actually observed are used in our model.

Brief methods

Our study used standard methods in speech science, similar to those used in earlier studies, but replaced the original cadaver estimates with more realistic and accurate input x-ray data from living animals. The methodology used is illustrated in Fig. 1. We first made x-ray videos of monkeys in various vocal tract configurations (Fig. 1A; shown is a macaque producing a threat call), extracted digitized still images of the most extreme vocal tract configurations observed, and digitally traced the vocal tract outlines for 99 configurations (Fig. 1B shows one example). With custom Matlab (version 2011b) scripts and C++ code,

[1]Department of Cognitive Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria. [2]Haidlhof Research Station, University of Vienna/University of Veterinary Medicine Vienna, Bad Vöslau, Austria. [3]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [4]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA. [5]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. [6]Department of Psychology, Princeton University, Princeton, NJ 08544, USA. *Corresponding author. Email: tecumseh.fitch@univie.ac.at (W.T.F.); asifg@princeton.edu (A.A.G.)

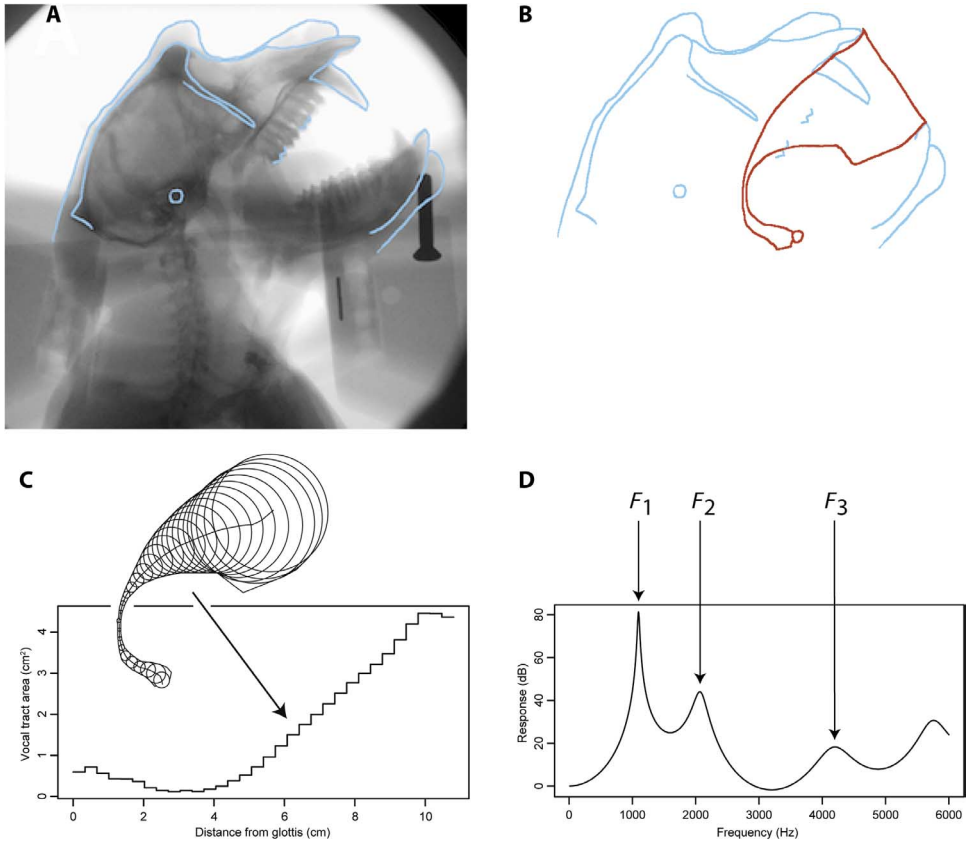

Fig. 1. Methodology for constructing a single vocal tract configuration. (**A**) We first made x-ray videos of monkeys and extracted still images of various vocal tract configurations (the example shown is a macaque producing a threat call). (**B**) We then traced the vocal tract outlines. (**C**) We used custom Matlab scripts to extract the diameter of the vocal tract along the glottis-to-lip midline (medial axis transform), straightened this diameter function, and converted it to a vocal tract area function. (**D**) Finally, the resultant area function was used to compute the vocal tract transfer functions for the observed vocal tract configuration [using Flanagan's lossy tube model (39)], and the first three formant frequencies (F1, F2, and F3) were extracted via peak picking. The set of all 99 vocal configurations, each computed in this fashion, was then used to estimate the monkey's phonetic space.

tracings were converted to vocal tract area functions (where x represents the distance from the glottis to the lips and y represents the vocal tract cross-sectional area at that point) by extracting the diameter of the vocal tract along the glottis-to-lip midline, straightening this diameter function, and converting it to a vocal tract area function based on magnetic resonance imaging (MRI) data for a male macaque (Fig. 1C). This anatomical description of the vocal tract geometry allows a direct calculation of the formant frequencies that are the main acoustic cues to phonetic identity in human speech. The area function was then used to compute the vocal tract transfer functions for the observed vocal tract configuration, and the first three formant frequencies (F1, F2, and F3) were extracted via peak picking (Fig. 1D). The set of formants for all 99 vocal configurations, each computed in this fashion, was then used to estimate the monkey's potential phonetic space.

RESULTS

Our core findings are shown in Fig. 2, which plots the large phonetic space generated by the observed macaque vocal tracts in the standard F1 versus F2 vowel space. Each point represents the formant data for an observed vocal tract shape. A convex hull enclosing all points is outlined in black, and the vocal tract configurations corresponding to the extreme points along the perimeter of this convex hull are illustrated in the margins.

The most extreme values that define this space might stem from the subset of our data where the monkey was eating: Food in the mouth might make tongue deformations possible that would not be available with the mouth empty. To evaluate this possibility, we reran the analysis omitting all feeding contours and obtained the same result; the chewing data yielded nonextreme vocal tract configurations (centered and close to a schwa articulation). Thus, all the extreme points in Fig. 2 stem from nonfeeding behaviors that were, or could have been, accompanied by vocalization.

Figure 3 provides a comparison with human vocal production, plotting the human female data for English vowels and the monkey convex hull enclosing all formant data (F1 versus F2 and F2 versus F3). Most of the vowels of American English have relatively close equivalents in the observed monkey transfer functions: For example, the point vowel /ae/ can be produced, along with a good approximation of the point vowel /a/. For comparison, the macaque formant values calculated by Lieberman *et al.* (*6*) are also shown in Fig. 3 as black dots. These traditional, cadaver-based estimates are clearly a considerable underestimate of the true phonetic potential of the living macaque monkey.

To illustrate the lack of restrictions imposed upon speech by the monkey vocal tract, we have produced the nearest monkey equivalents of human speech utterances; one of these, for the phrase "Will you

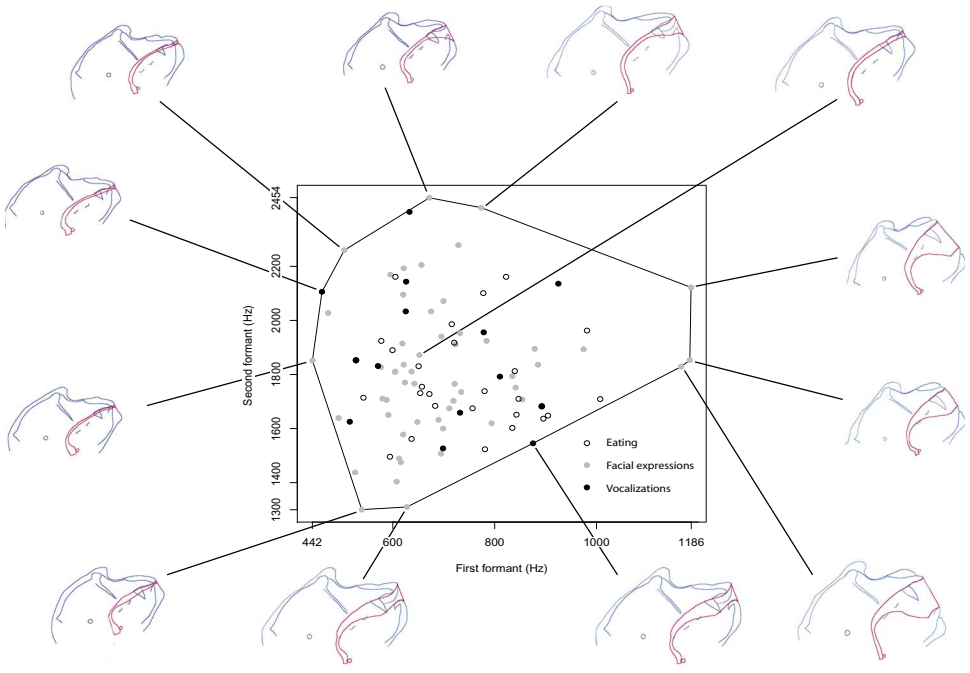

Fig. 2. Attested macaque monkey formant space. Formant plot (F1-F2) for all 99 observed monkey vocal tract configurations, enclosed in a convex hull to show total phonetic space, with corresponding tracings of extreme vocal tract configurations. Eating-related outlines (**open circles**) are all the outlines in which a food item (banana or orange slice, grape or raisin) was involved. Facial expressions (**gray dots**) are yawns and various lip smacks. Vocalizations (**black dots**) involved production of sound through coos and grunts.

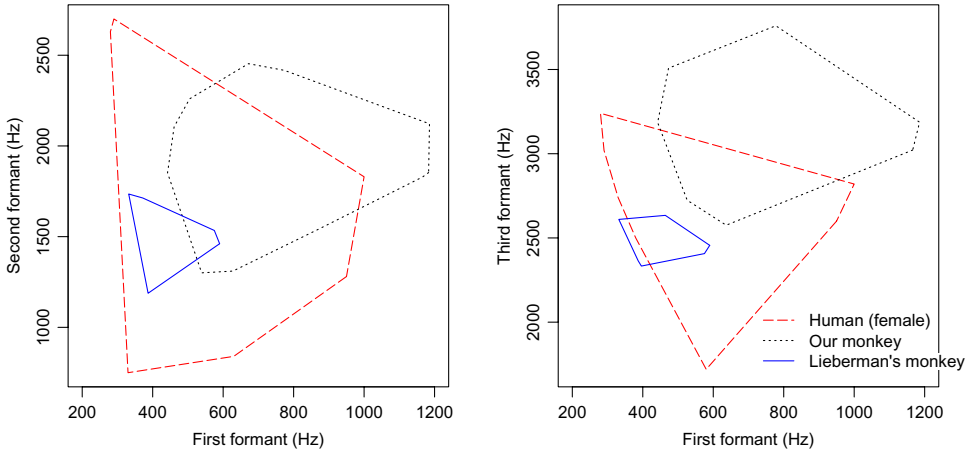

Fig. 3. Formant plot comparisons. Macaque monkey (**black dotted line**) versus human female vowel space [**red dashed line**; American English from Peterson and Barney (*18*)]. The left panel shows F1 against F2, and the right panel shows F1 against F3. For comparison, the blue outline shows the previous macaque monkey estimates from Lieberman *et al.* (6).

marry me?," is illustrated in Fig. 4 (Fig. 4A is female human speech, and Fig. 4B is the closest monkey equivalent). As can be seen, the sounds are similar in most ways, except that the monkey formants are somewhat more evenly spaced than the human formants. In particular, F2 does not cover as broad a frequency range and never approaches as closely to F3. Although this effect is audible and sounds like a slight indistinctness or foreign accent, the monkey version of such phrases is still clearly understandable (audio file S2).

Finally, to provide a more formal evaluation of the phonetic potential of the macaque's vocal tract, we used an iterative procedure to generate five "ideal" vowels, which used the macaque's observed formant space in a maximally differentiated fashion. Random pairs of these five vowels were then played to 10 adult human listeners in an ABX discrimination task: Listeners heard first vowel A and then vowel B and had to decide whether a third stimulus X was the same as A or B. This is a more demanding discrimination task than simple same/different judgments of two vowels and is often used in speech perception research (*16*). Each vowel was paired with the remaining four vowels, and each pair was presented twice (in both possible vowel orders), yielding 40 trials per participant. Human participants answered between 36 and 39 trials correctly (90 to 98% correct); the mean error rate was only 7% (differing significantly from 50% chance by a binomial

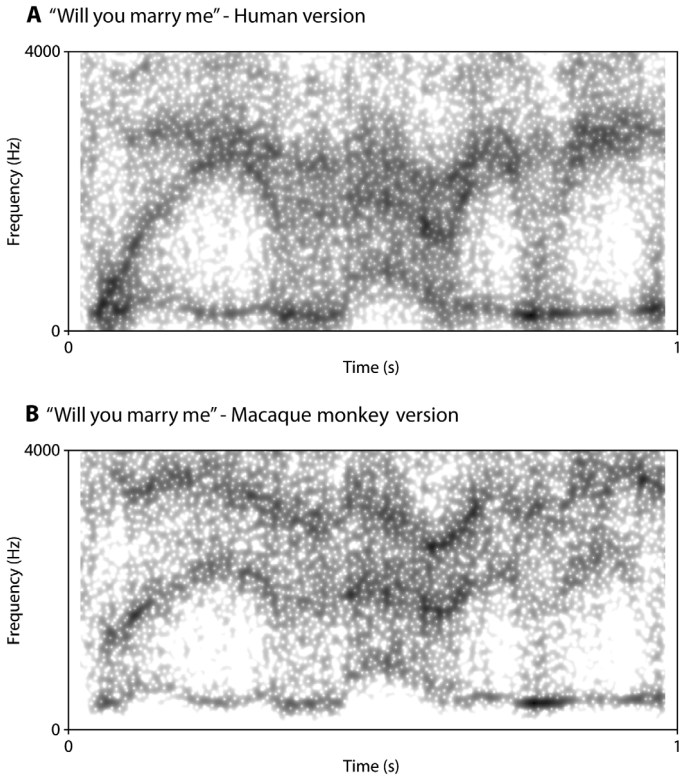

Fig. 4. Spectral comparison. Spectrograms of original speech (the English phrase "Will you marry me?," spoken by a human female) (**A**) and a synthesized version of the same phrase using formant values available, based on observed vocal tract configurations, to a macaque monkey (**B**).

test, $P < 0.000001$). These results are comparable to previous results examining perception of English vowels (*17*, *18*). Thus, a monkey that could use the full formant range inherent in its vocal tract would be able to produce a set of five easily discriminable vowels.

DISCUSSION

Our data indicate that the actual phonetic potential of the primate vocal tract is not highly restricted as previously thought (*6*, *19*). Because we used essentially the same modeling approach as in the classic studies by Lieberman and colleagues (*6*), but based our analysis on radiographs from actual living monkeys rather than reconstructions from cadavers, we believe that our approach provides a much more accurate estimate of the vocal sounds that a macaque monkey could potentially produce.

The key conclusion from our study is that the basic primate vocal production apparatus is easily capable of producing five clearly distinguishable vowels (for example, those in the English words "bit," "bet," "bat," "but," and "bought"). Five vowels are the worldwide norm for human languages (*20*), and many of the world's languages make do with only three vowels. Although it is more challenging to estimate the range of consonants that would be producible by a monkey (*21*), the common stop consonants (/p/, /b/, /k/, and /g/) along with a variety of other consonantal sounds (for example, /h/, /m/, and /w/) would be easily attainable by a macaque monkey. In any case, consonant production has never been proposed as a major restriction on the phonetic potential of nonhuman primates [see the reviews by Lameira *et al.* (*21*) and Fitch (*22*)].

We do not of course argue that a talking macaque would sound precisely the same as a human or that a macaque could create every possible vowel. For instance, our macaque monkey never produced a vocal tract shape corresponding to the /i/ vowel (that in "beet"), which has been suggested to play a special role in human speech (*19*, *23*). Of course, the restriction of our approach to vocal tract configurations that were actually observed remains very conservative, and it is possible that a monkey could produce more extreme vocal tract shapes, such as /i/, given suitable neural control and/or training. There is no reason to believe that linguistic communication requires /i/ or any other particular vowel, because the role of extreme vowels (for example, for vocal tract normalization) could be played by whatever vowel does represent the extreme "corner" of the phonetic space available to that species. Thus, we do not claim that macaque speech would sound precisely like human speech. Rather, our results definitively show that the phonetic range inherent in a macaque vocal tract, based on actual observed vocal tract configurations, would itself pose no impediment to linguistic communication if macaques had human-like neural control systems.

We conclude that if a macaque monkey had a brain capable of vocal learning and combinatoric operations over speech sounds, its vocal tract would be able to produce clearly intelligible speech. Our data join those based on computer models of the human vocal tract, showing that the importance of human vocal anatomy for speech has been overestimated (*24*–*26*). Our data are also consistent with anatomical data from apes and monkeys, indicating that human vocal anatomy and the descended larynx are not as exceptional as widely thought (*27*, *28*). These findings refute the widespread opinion that nonhuman primate vocal tracts are "unsuited to speaking" [(*29*), p. 59]. We conclude that the inability of macaques and other primates to speak is a reflection not of peripheral vocal tract limitations but of their lack of neural circuitry enabling sophisticated vocal control (*30*, *31*). In short, primates have a speech-ready vocal tract but lack a speech-ready brain to take advantage of its latent operating range.

MATERIALS AND METHODS

All experiments followed national and international regulations. Monkey experiments were performed in compliance with and were approved by the Princeton University Institutional Animal Care and Use Committee guidelines for the care and use of laboratory animals (www.princeton.edu/ria/animal-research-protectio/committee-information/). Human experiments were approved by the Princeton University Institutional Review Board for Human Research (www.princeton.edu/ria/human-research-protection/guidelines/).

Monkeys were provided with a specially designed plastic collar (after an ordinary metal collar used on the first day of acquisition was found to partially obscure laryngeal movements) and placed in a plastic primate chair (Crist Instrument), which allowed free movement of the head and complete 360° body rotation.

Behaviors

X-ray videos of a range of different orofacial movements were acquired, including vocalizations, lip smacking (a common macaque affiliative facial gesture), yawning, and feeding on preferred food items. Animals were periodically fed a 2% barium sulfate suspension (Readi-Cat 2, E-Z-EM Canada) to outline the vocal tract contour. The main vocalizations the monkeys produced were low-amplitude "grunt" vocalizations and, occasionally, coo and "threat" calls (*32*–*34*). Grunts are

short, broadband pulsatile vocalizations, produced during affiliative and food-related situations; this vocalization class provided excellent resolution of the vocal tract formant frequencies that were our primary interest. Coos are longer and more tonal than grunts but otherwise similar. Threats are noisy aperiodic sounds that also provided excellent formant resolution. In some cases, the animals were induced to vocalize in the fluoroscope, typically upon presentation of an edible treat. Lip smacks were elicited either by bringing another monkey into the room or by the human experimenter facing the monkey and simulating a lip smack.

Radiograph acquisition

Three adult male long-tailed macaque monkeys (*Macaca fascicularis*, ages 7 to 9 years; Franco, Patrice, and Emiliano) were examined by video fluoroscopy [Philips BV Pulsera system; 12-inch image intensifier, software R2.2.6 (22 July 2008)] in a 5.9 m × 2.9 m concrete room sound-treated with Acoustiblok acoustic panels (4 feet × 8 feet) and SONEX foam to reduce noise and reverberation. The Pulsera was set in cardiac mode, the Auto kV varied from 57 to 62 kV, and the frame rate was 30 frames per second (fps). Monkeys were seated in restraint chairs for placement within the x-ray field of view, but were not head-fixed and were allowed a complete range of head motion, essential to capturing natural orofacial gestures. In addition to the real-time audio/video acquisition described above, individual video clips and still images of higher resolution were also acquired directly into the Philips Pulsera system and were transferred via Ethernet to a DICOM server on a PC. Although this still format lacked synchronized sound, these images were used to provide higher-resolution still images for x-ray tracings, for quality comparison. Our digitized videos were of equivalent quality.

Videos were acquired in real time, direct to a PC hard disk (on a Dell OptiPlex 960 computer running Microsoft Vista Home Basic Service Pack 1) using Adobe Premiere software and Canopus ADVC110 (Thomson; www.thomsongrassvalley.com) audio/video digitizer connected via an IEEE 1394 FireWire interface as NTSC AVI files (US mode, 7.5 IRE; 720 × 480 pixels, NTSC format, 29.97 fps). For analysis, all AVI video files were segmented into their component still frames as PNG files to produce a large set of still images. A 5 × 5 wire calibration grid was periodically placed into the field of view beside or above the monkey's head to allow calibration of the recorded video images (~13 mm^2; 63.5 mm × 63.5 mm outside diameter for the entire grid with all five squares); calibration was performed for each x-ray session with each monkey.

Monkey phonetic vocal tract model

To derive the vocal tract potential for a representative macaque, we selected a subset of 99 still images from a single monkey (Emiliano) that exhibited a wide range of behaviors and provided excellent video images. All x-ray stills were midsagittal views, where the long axis of the monkey's skull was in the plane of the image; stills were selected for high image quality and to represent a broad range of observed vocal tract configurations. We selected a mixture of vocalizations (coos, grunts, and threats; $n = 34$), facial expressions (lip smacks and yawns; $n = 37$), and feeding on various food types (grapes, raisins, orange slices, and banana pieces; $n = 28$). These selected stills were hand-traced using a graphics tablet (Wacom Cintiq 12WX) and Adobe Photoshop CS4. For calibration and alignment purposes, we traced the outline of the skull and various landmarks in one image layer, including the canine teeth, orbital bars, and auditory meatus. In a separate layer, we traced the vocal tract outline (formed by the palate and posterior pharynx wall dorsally and tongue surface ventrally). Image size calibration (converting pixel values to millimeters) used the digitized x-ray images of the 63.5 mm × 63.5 mm wire grid, held adjacent to a metal screw within the primate chair. This reference was used to measure the collar fixation screw (35 mm) that was present in all images. This was then used to measure the skull length of the monkey in six different x-ray images, where the monkey's head and jaws were in diverse positions, using the Fiji image processing package (fiji.sc). These six values were then averaged (14.1 cm; SD, 0.15 cm). We then used a 14-cm skull length value to scale all other images. These tracings were then exported as bitmapped PNG images, which served as the input for further analysis.

Hand-traced vocal tract lines were imported as bitmap images into Matlab, and three hard bony landmarks (the skull reference line from the anterior tip of the maxilla to the opening of internal ear canal, and a third point at the posteriormost point of the sagittal crest) were selected using a custom Matlab interface. These landmarks were used to normalize the images to have identical skull dimensions and rotation, compensating for the minor changes due to head rotation and/or the distance from the x-ray system. Because the restraining collar and laryngeal movements often made the vocal folds difficult to visualize precisely, the location of the vocal folds was conservatively estimated at just below the laryngeal inlet (using the location of the bullate hyoid bone just above the larynx as a reference, indicated by the small circle in Fig. 1, A and B); this location and the lip exit were also selected (two points each, dorsal and ventral). A fifth point roughly indicating the transition from the pharyngeal to oral vocal tract was additionally selected; this was used only to aid the construction of an anatomical reference grid and played no role in the acoustic analysis. The x and y coordinates of the dorsal and ventral vocal tract outlines were then extracted. To produce a straightened vocal tract, we used a Cartesian reference grid, with one line for every 10 pixels, starting with the x axis parallel to the skull reference line. The grid was then rotated 90° at the pharyngeal-oral transition point, interpolating two intermediate grid lines at 30° intervals. The grid then proceeded in this orientation to the lip exit. Using this reference grid, the center point of each dorsal and ventral outline tracing was calculated. Points were chosen along each vocal tract boundary, with one point for every 10 pixels, and output to a normalized text file as x and y coordinates, with roughly 50 automatically extracted points per vocal tract edge.

These isometrically normalized vocal tract outlines were then used to calculate vocal tract area functions in three steps. In the first step, the vocal tract diameters were extracted along the length of the vocal tract (in two dimensions), using two different methods, which produced comparable results. One method was a variant of that used by Goldstein (*35*), but we did not calculate the side branches of the vocal tract because these were not present in our data (the nasal tract entrance was closed off). The program uses the medial axis transform (*36*) and was calculated using custom C++ programs, implemented in Apple Xcode. This method works intuitively by attempting to progressively fit a circle of maximal diameter along the length of the vocal tract. The vocal tract midline is given by the center of the fitted maximal circle, and its diameter gives the cross-sectional diameter. The alternative method for estimating vocal tract geometry followed Mermelstein (*37*) and was calculated using custom Matlab scripts. First, the midline contour, midway between the dorsal and ventral edges, was calculated as a simple average of these two contours. Then, starting at the glottis, and proceeding every 10 pixels along this midline, a line perpendicular

to the midline contour was drawn, and its intersection with the dorsal and ventral vocal tract contours was extracted. The vocal tract diameter at this point was the length of this line. In 9 of 99 cases, at the termination of the vocal tract (lip exit), the final midline perpendicular did not necessarily intersect both edges; in this case, the diameter line was clamped at the last viable point on the contour, and the final diameter measurement ran between the clamped point, the midline point, and the intersection point on the opposite side.

In the second step, the cross-sectional diameter functions were converted to the three-dimensional area functions necessary for acoustic calculations. The diameter functions, along with normalization information, were read into Matlab and converted into centimeters using the reference 14-cm skull length. These diameters were then converted to areas using the following formula: $1.75 \times \text{diameter}^{1.4}$. These values were empirically determined, based on the MRI images of a monkey vocal tract of the same species, sex, and approximate body size (Franco), as illustrated in fig. S1. Our 1.4 exponent was very close to the value used for humans by Mermelstein (*37*), but we used the same conversion function along the entire vocal tract, whereas Mermelstein altered his human area conversion value slightly depending on vocal tract location. To avoid unnaturally small constrictions that might generate turbulence in a real vocal tract, a fixed minimum offset of 0.1 cm^2 ("min diam") was added to the resulting vocal tract area function, resulting in a minimum area (0.138 cm^2) closely corresponding to the minima observed during human vowel production (*38*).

In the third and final step, these vocal tract area functions were used to calculate vocal tract resonances ("formants" hereafter), using two standard methods in speech science (lossless and lossy tube); formant center frequencies were nearly identical in the two methods. Because the lossy model, which includes losses and radiation, produces more realistic formant bandwidths, these data are presented here. This method (Toeter2formants3.m) implements Flanagan's lossy vocal tract model [(*39*), p. 24], which uses a lossy transmission line model: The vocal tract is modeled as a series of tubelets, where each tubelet is approximated by a simple RLC circuit with capacitance, resistance, and inductance, and these equivalent circuits are connected in series. The frequency response of this transmission line was then calculated, between 10 Hz and 6 kHz, at a resolution of 1 Hz. The three lowest-frequency peaks of this vocal tract transfer function were then determined by the points where the first derivative changed from positive to negative sign and were denoted formants one, two, and three. The other simpler method (implemented in code file Toeter2formants.m) calculates the transfer function directly by assuming uniform lossless tubes [based on the method described by Rabiner (*40*)], calculating the frequency of the poles of the transfer function by finding the zeroes of the polynomial that constitutes its denominator. These poles correspond directly to the formants of the lossless tube model.

Visualization and plotting

Finally, for comparison with a human female speaker, the average length of all observed monkey vocal tract shapes (11.4 cm) was multiplicatively scaled to an average value appropriate for an adult female human (14 cm). For visual comparison with human vocal tract potential, the monkey formant values are plotted against those of a single typical female (with near-average formant values) from Peterson and Barney's data for American vowel formants [speaker 36 in the works of Peterson and Barney (*18*) and Watrous (*41*)]. For comparison (see the Supplementary Materials), we also performed a similar monkey/human comparison with Dutch vowel data from Pols *et al.* (*42*) and van Nierop *et al.* (*43*), in which the identity of the back vowel is clear and the front-rounded vowels (which exist in Dutch but not in English) are present. Again, the results were equivalent (fig. S2). To evaluate the quantitative differences between the space, we calculated the "articulatory efficiencies" (a standardized measure of the area of the acoustic space a vocal tract can reach). The F1/F2 efficiency was increased nearly an order of magnitude, from 1406 to 10,986 Hz^2 m^2 (*6*).

Perceptual experiments

Using the monkey vowels derived from the vocal tract outlines, we calculated the convex hull. In this convex hull, we ran the Liljencrants and Lindblom (*44*) simulation 100 times with five vowels and obtained the same final vowel system 100 times (indicating that there is only one optimal vowel system in this case). These five vowels were used as ideal monkey vowels. We scaled the formant frequencies of these vowels from a vocal tract length of 11.4 cm to a human female's tract length of 14 cm, and then used Praat to synthesize them, using a recording of a macaque grunt vocalization as the source. These were then played in pairs to participants (Princeton University students) in an ABX discrimination task, with each of the $5 \times 4 = 20$ possible different pairs played twice in each order. The task was implemented using LiveCode (www.livecode.com); participants listened via headphones (Grado Labs SR-225) and made their choice using the computer track pad (MacBook Air). The perceptual experiment was approved by the Princeton University Institutional Review Board for Human Subjects (protocol #7783).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/12/e1600723/DC1

audio file S1. Audio file of an adult human female saying "Will you marry me?," resynthesized with a noisy source.

audio file S2. Audio file of our macaque vocal model uttering the same phrase "Will you marry me?," synthesized with the same noisy source.

fig. S1. Monkey MRIs used to estimate the conversion factor from linear midsagittal diameter measurements to two-dimensional areas.

fig. S2. Comparison of human female vowels in Dutch, which has vowels not present in English (red dashed line), with the macaque vocal tract model (gray dotted line).

REFERENCES AND NOTES

1. W. N. Kellogg, Chimpanzees in experimental homes. *Psychol. Rec.* **18**, 489–498 (1968).
2. W. T. Fitch, The evolution of speech: A comparative review. *Trends Cogn. Sci.* **4**, 258–267 (2000).
3. V. M. Janik, P. J. B. Slater, Vocal learning in mammals. *Adv. Study Behav.* **26**, 59–99 (1997).
4. S. Nowicki, W. A. Searcy, The evolution of vocal learning. *Curr. Opin. Neurobiol.* **28**, 48–53 (2014).
5. C. F. Hockett, The origin of speech. *Sci. Am.* **203**, 89–96 (1960).
6. 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).
7. P. Lieberman, E. S. Crelin, D. H. Klatt, Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *Am. Anthropol.* **74**, 287–307 (1972).
8. G. Yule, *The Study of Language* (Cambridge Univ. Press, ed. 3, 2006), xxx pp.
9. D. Crystal, *The Cambridge Encyclopedia of Language* (Cambridge Univ. Press, ed. 2, 2003), xxx pp.
10. P. Lieberman, S. E. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge Univ. Press, 1988), xxx pp.
11. L. J. Raphael, G. J. Borden, K. S. Harris, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech* (Lippincott Williams & Wilkins, 2007), xxx pp.
12. A. Barney, S. Martelli, A. Serrurier, J. Steele, Articulatory capacity of Neanderthals, a very recent and human-like fossil hominin. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 88–102 (2012).

13. A. MacLarnon, in *The Oxford Handbook of Language Evolution,* M. Tallerman, K. R. Gibson, Eds. (Oxford Univ. Press, 2011), pp. 224–235.
14. W. T. Fitch, The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals. *Phonetica* **57**, 205–218 (2000).
15. A. A. Ghazanfar, D. Y. Takahashi, N. Mathur, W. T. Fitch, Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. *Curr. Biol.* **22**, 1176–1182 (2012).
16. A. M. Liberman, K. S. Harris, H. S. Hoffman, B. C. Griffith, The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* **53**, 358–368 (1957).
17. A. Cutler, A. Weber, R. Smits, N. Cooper, Patterns of English phoneme confusions by native and non-native listeners. *J. Acoust. Soc. Am.* **116**, 3668–3678 (2004).
18. G. E. Peterson, H. L. Barney, Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **24**, 175–184 (1952).
19. P. Lieberman, *The Biology and Evolution of Language* (Harvard Univ. Press, 1984).
20. I. Maddieson, *Patterns of Sounds* (Cambridge Studies in Speech Science and Communication, Cambridge Univ. Press, 1984).
21. A. R. Lameira, I. Maddieson, K. Zuberbühler, Primate feedstock for the evolution of consonants. *Trends Cogn. Sci.* **18**, 60–62 (2014).
22. W. T. Fitch, in *The Cradle of Language*, R. P. Botha, C. Knight, Eds. (Oxford Univ. Press, 2009), pp. 112–134.
23. P. Lieberman, The evolution of human speech: Its anatomical and neural bases. *Curr. Anthropol.* **48**, 39–66 (2007).
24. L.-J. Boë, J.-L. Heim, K. Honda, S. Maeda, The potential Neandertal vowel space was as large as that of modern humans. *J. Phonetics* **30**, 465–484 (2002).
25. L.-J. Boë, S. Maeda, J.-L. Heim, Neanderthal man was not morphologically handicapped for speech. *Evol. Commun.* **3**, 49–77 (1999).
26. B. de Boer, W. T. Fitch, Computer models of vocal tract evolution: An overview and critique. *Adapt. Behav.* **18**, 36–47 (2010).
27. 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).
28. T. Nishimura, A. Mikami, J. Suzuki, T. Matsuzawa, Development of the laryngeal air sac in chimpanzees. *Int. J. Primatol.* **28**, 483–492 (2007).
29. T. Harley, *The Psychology of Language: From Data to Theory* (Psychology Press, ed. 4, 2014).
30. U. Jürgens, Neural pathways underlying vocal control. *Neurosci. Biobehav. Rev.* **26**, 235–258 (2002).
31. G. Holstege, H. H. Subramanian, Two different motor systems are needed to generate human speech. *J. Comp. Neurol.* **524**, 1558–1577 (2016).
32. T. E. Rowell, R. A. Hinde, Vocal communication by the rhesus monkey (*Macaca mulatta*). *Proc. Zool. Soc. Lond.* **138**, 279–294 (1962).
33. M. D. Hauser, Sources of acoustic variation in rhesus macaque (*Macaca mulatta*) vocalizations. *Ethology* **89**, 29–46 (1991).
34. D. Rendall, M. J. Owren, P. S. Rodman, The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. *J. Acoust. Soc. Am.* **103**, 602–614 (1998).
35. U. G. Goldstein, thesis, Massachusetts Institute of Technology (1980).
36. H. Blum, in *Models for the Perception of Speech and Visual Form*, W. Whaten-Dunn, Ed. (MIT Press, 1967), pp. 362–380.
37. P. Mermelstein, Articulatory model for the study of speech production. *J. Acoust. Soc. Am.* **53**, 1070–1082 (1973).
38. B. H. Story, I. R. Titze, E. A. Hoffman, Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.* **100**, 537–554 (1996).
39. J. L. Flanagan, *Speech analysis, synthesis and perception* (Springer, 1965), 317 pp.
40. L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, 1978).
41. R. L. Watrous, Current status of Peterson-Barney vowel formant data. *J. Acoust. Soc. Am.* **89**, 2459–2460 (1991).
42. L. C. W. Pols, H. R. C. Tromp, R. Plomp, Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **53**, 1093–1101 (1973).
43. D. J. P. J. van Nierop, L. C. W. Pols, R. Plomp, Frequency analysis of Dutch vowels from 25 female speakers. *Acustica* **29**, 110–118 (1973).
44. J. Liljencrants, B. Lindblom, Numerical simulations of vowel quality systems. *Language* **48**, 839–862 (1972).

Acknowledgments: We thank S. Steckenfinger for help with data collection and management and L. Kelly for animal care. **Funding:** This work was supported by European Research Council (ERC) Advanced Grant SOMACCA 230604 (to W.T.F.), NIH grant R01NS054898 (to A.A.G.), and ERC starting grant ABACUS 283435 (to B.d.B.). **Author contributions:** A.A.G. and W.T.F. designed the project and gathered the x-ray data; N.M. converted the raw movie files to computer-ready tracings; B.d.B. and W.T.F. performed the computer-based analysis and generated the figures; and B.d.B. wrote the software. W.T.F., A.A.G., and B.d.B. wrote the paper, which was approved by N.M. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 5 April 2016
Accepted 8 November 2016
Published 9 December 2016
10.1126/sciadv.1600723

Citation: W. T. Fitch, B. de Boer, N. Mathur, A. A. Ghazanfar, Monkey vocal tracts are speech-ready. *Sci. Adv.* **2**, e1600723 (2016).