TRACKING THE EVOLUTION OF LANGUAGE AND SPEECH

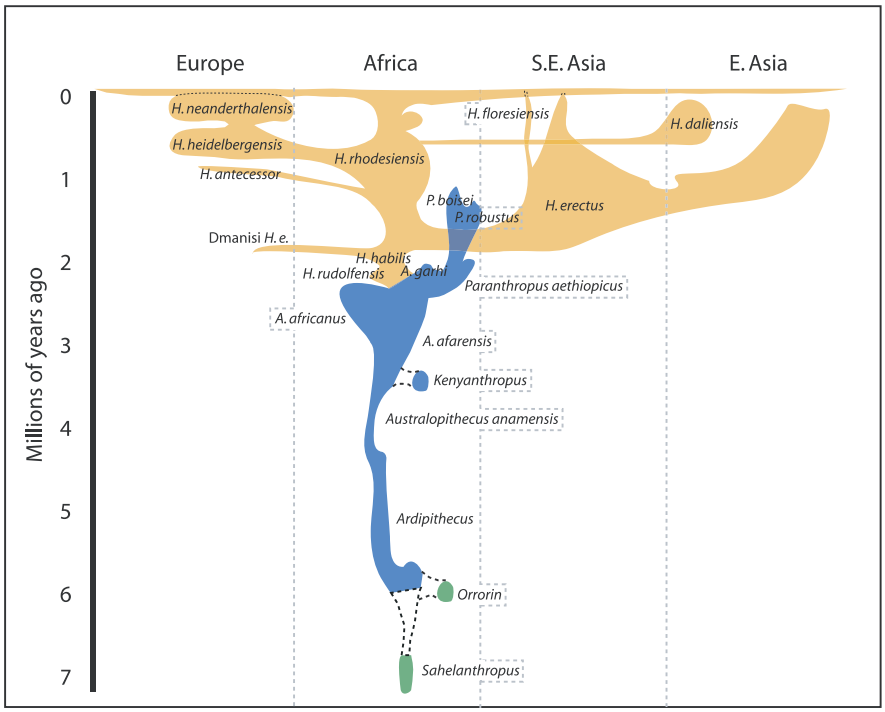
Comparing Vocal Tracts to Identify Speech Capabilities

BY PHILIP LIEBERMAN & ROBERT MCCARTHY

In 1973 THEODOSIUS DOBZHANSKY wrote that "nothing in biology makes sense except in the light of evolution." This dictum applies equally well to human language and speech, which have an evolutionary history that has yet to be fully discovered. Unfortunately, apart from their sometimes fossilized bones and archaeological traces of their behavior, nothing remains of our distant ancestors. Yet the mark of our evolution may be discerned in our modern bodies, brains, and even our vocal tracts.

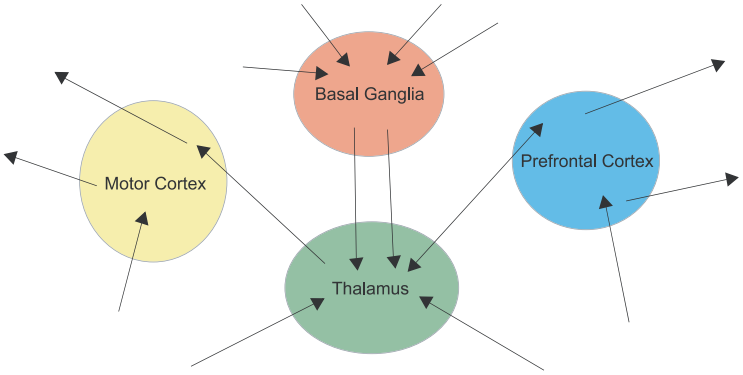
Evidence from seemingly unrelated disciplines suggests that the specialized anatomy and neural mechanisms that confer fully human speech, language, and cognitive ability reached their present state sometime between 100,000 and 50,000 years ago. The appearance of these attributes relatively late in our evolution—well after our species originated about 200,000 years ago—has important implications for how we think about ourselves, our ancestors, and our collateral relatives (including the Neanderthals who evolved separately from our common ancestor starting about 500,000 years ago). In fact, the appearance of modern human bodies well before the appearance of what we consider to be modern human behavior—our higher mental processes such as complex thought, language, and symbolic behavior—suggests that there was something about our early modern ancestors that allowed them to develop into our more recent, fully modern selves. That building block may have been something as simple as speech, the vocal transmission of information at a very fast rate.

When we look at our brains the neural mechanisms necessary to produce fully articulate speech are intricately connected to the regulation of complex syntax and cognition. Rather than being localized in one part of our brain—as was traditionally thought in the 19th century—we now know that the neural bases of speech and language are actually found in the "circuits" that connect different parts of the brain. In most animals such circuits regulate the motor control of the body,



Human language and speech seem to have evolved between 100,000 and 50,000 years ago, a relatively late date in hominid evolution.

Robert McCarthy



Within our brains, independent groups of neurons in one part of the brain connect to distinct groups of neurons in other parts, forming "circuits" that regulate different aspects of behavior. This schematic shows neuronal populations from different regions of the outer brain, or cortex, that project into the putamen—a subcortical inner part of the brain—and from there indirectly into other regions of the cortex. These circuits through the putamen regulate the motor control of our bodies, for example, changing the direction of a thought process, comprehending the meaning of a sentence, and regulating our mood and emotions.

but in modern humans they also affect our cognitive abilities. For example, these circuits allow us to change the direction of our thought processes based on new stimuli such as the understanding of meaning conveyed by the syntax of language.

This is an important clue to understanding the evolution of human language because it indicates that our modern brains may actually have been shaped by an enhanced capacity for speech motor control that evolved in our ancestors. In other words, as our ancestors grappled with improved modes of speaking to each other, their brains gradually developed more complex language skills, allowing us to form and comprehend complex syntax. Over time, these changes made us "human"—we may have actually talked ourselves into being smarter!

SPEECH AND SPEECH PHYSIOLOGY

Speech is a special mode of communication, providing a rapid rate of information transfer necessary for complex language. Although there have been many attempts to devise systems that allow humans to communicate using sound—such as Morse code, tones, and musical notes—such systems require listeners to pay undivided attention in order to interpret the sequence of sounds and their meanings. This results in an information transfer rate that is agonizingly slow.

Speech is different. When a person listens to speech, he or she decodes, or unscrambles, melded acoustic cues through a complex perceptual process that relies on the listener's unconscious "knowledge" of the physiology of speech production. In its most basic form, speech is a vocalization produced during the expulsion of air from the lungs while breathing. In reptiles and mammals, the larynx, or voicebox, converts the turbulent energy of air coming from the lungs into higher, audible frequencies through a process called phonation.

In humans and other mammals, the larynx is a complex structure made of cartilage, muscle, and other soft tissues. The vocal cords of the larynx—which are situated in the thyroid cartilage that forms the laryngeal prominence known as the "Adam's apple"—act as a valve that rapidly opens and closes during phonation, releasing puffs of air at a frequency determined by the rate of airflow from the lungs and the degree of tension in the laryngeal muscles. The rate at which these puffs of air are released is known as the fundamental frequency of phonation (F0), which is closely related to the perceived pitch of a person's voice.

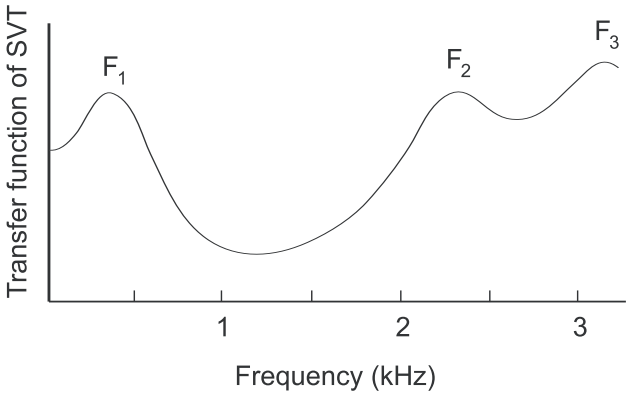
In humans, vocalizations are modified in the airway above the vocal cords—the supralaryngeal vocal tract, or SVT—by positioning the tongue, lips, and larynx. This alters the overall shape of the SVT, allowing it to filter local energy peaks as they

A REMARKABLE DISCOVERY

In the 1960s at the Haskins Laboratories at Yale University researchers attempted to build a machine that would read books aloud to the blind. At the outset, this seemed to be a trivial problem, akin to typing letters on a keyboard corresponding to each sound. However, the Haskins researchers soon discovered that phonemes—the functional units of language—could not be isolated and strung together like an alphabet to produce intelligible sentences. The results of such experiments were the verbal equivalent of a ransom note, with the different-sized letters barely intelligible together in the same sentence.

In order to be understood, the acoustic signals delineating phonemes must be carefully strung together, a process requiring coordinated muscle movements of the tongue, lips, soft palate (which opens the airway to the nose), and larynx. Such complex gestures cannot be isolated because they are not discrete entities, but are instead blended together during speech. For example, the words "tea" and "too" contain the same phoneme [t], but they are produced by positioning the lips differently at the beginning of the word. Therefore, the acoustic signal is different because the quality of [t] varies based on the position that the lips take at the start of the syllable to enunciate the vowel following [t]. In other words, the resulting sound pattern is *encoded* so that the acoustic cues that convey the initial consonant and vowel are transmitted in the same time frame.

Philip Lieberman



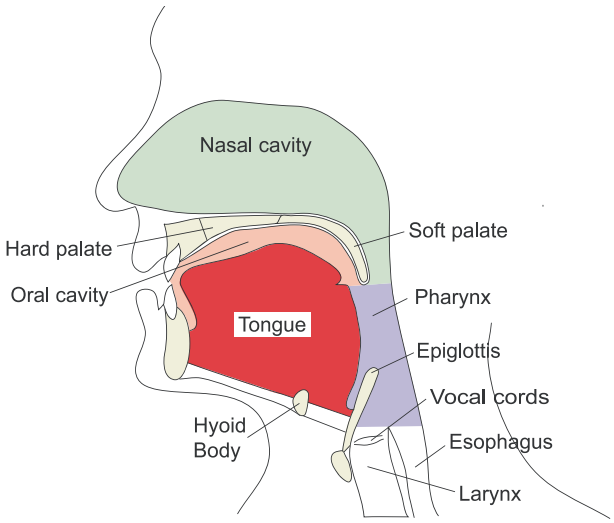
The supralaryngeal vocal tract, or SVT, determines the phonetic quality of speech sounds by serving as a filter for acoustic energy produced in the larynx. It does this by generating formant frequencies that differentiate vowels and consonants in much the same way as a pipe organ produces different notes. The lowest formant frequency is identified by the notation F1, the next highest as F2, the third as F3. Different vowels can have identical fundamental frequencies but very different formant frequencies (e.g. the vowels [i] and [u] of the words "see" and "sue"). Here the filter function of a 17 cm long SVT for the vowel [i] (as in "see") is produced by a larynx phonating at 100 Hz, in addition to the distribution of acoustic energy at the speaker's lips at different frequencies. The horizontal axis represents frequency in KHz, while the vertical axis represents the degree to which sound energy passes through the SVT at a given frequency.

pass through it. As a result, the SVT acts on the acoustic signal in much the same way that a pipe organ of a particular length and shape determines the frequency of acoustic energy in a musical note. However, whereas all notes produced by organ pipes occur at mathematical multiples of the lowest frequency, the human SVT is extremely malleable and constantly changing shape. Therefore, humans can produce a wide range of formant frequency patterns that form the basis for human speech.

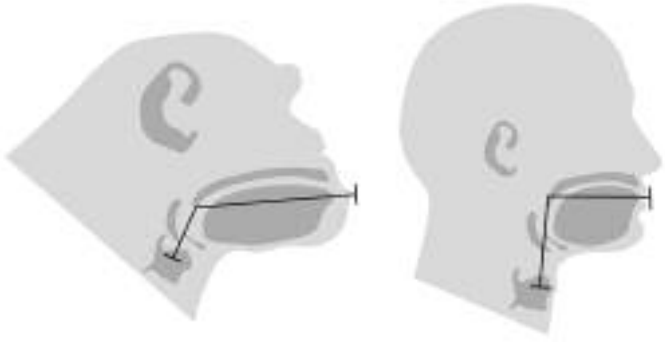
THE UNIQUE VOCAL TRACT OF MODERN HUMANS

In his *Origin of Species* Charles Darwin noted that the human vocal tract differs from that of other living primates in a way that increases the likelihood of choking. Our vocal tracts are divided into two sections—a "horizontal" portion (SVT_H) in the oral cavity, which includes the mouth and oropharynx, and a "vertical" portion in the throat called the pharynx, which is located behind the tongue and above the larynx. This vertical portion of the vocal tract (SVT_V) extends from the palate down to the vocal cords.

In normal adults these two portions of the SVT form a right angle to one another and are approximately equal in length—in a 1:1 proportion. Movements of the tongue within this space, at its midpoint, are capable of producing tenfold changes in the diameter of the SVT. These tongue maneuvers produce the abrupt diameter changes needed to produce the formant frequencies of the vowels found most frequently



The adult human supralaryngeal vocal tract (SVT) has a horizontal portion (SVT_H) associated with the oral cavity and a vertical portion (SVT_V) associated with the pharynx, of almost equal lengths. The natural discontinuity formed by the intersection of SVT_H and SVT_V in modern humans enables speakers to form abrupt changes in the cross-sectional area of the SVT at its midpoint, allowing the production of a wide range of sounds.

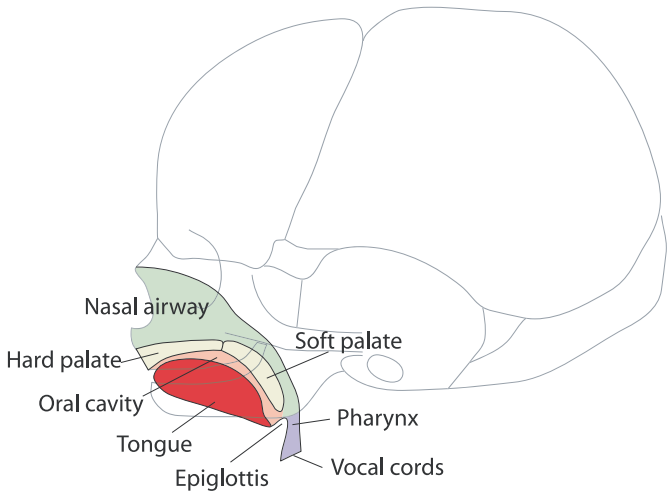


In chimpanzees the hyoid bone and larynx are positioned high in the throat, at or near the base of the mandible, and the tongue is long and largely restricted to the oral cavity, resulting in a disproportionately shaped SVT. In modern humans, the hyoid bone and larynx are low in the throat, well below the lower border of the mandible, and the tongue is large and only partially located in the oral cavity, resulting in an equally proportioned SVT.

among the world's languages—the "quantal" vowels [i], [u], and [a] of the words "see," "do," and "ma." In contrast, the vocal tracts of other living primates are physiologically incapable of producing such vowels. Besides having relatively small tongues, these primates have a SVT that is disproportionately long in the horizontal dimension compared to its vertical portion.

The peculiar vocal tract configuration of modern humans develops slowly during our lifetime. As newborns we actually start life with a vocal tract similar to those of most non-human primates and other mammals. As infants, our tongues are positioned almost entirely in the oral cavity, allowing our larynx to lock into the nose and form a sealed airway so we can

Robert McCarthy (top left and right), Elizabeth Murdoch (bottom right)



The tongues of human infants are positioned almost entirely in the oral cavity, allowing the larynx to lock into the nose and form a sealed airway so they can simultaneously suckle and breathe.

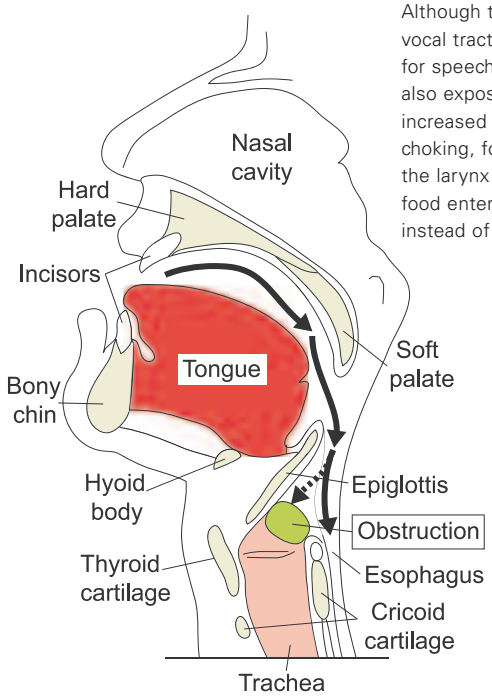
simultaneously suckle and breathe. As a result, human infants, like most other mammals, are able to ingest both air and liquid at the same time.

However, unlike most other mammals, during the first two years of our lives, the roof of our mouth flexes relative to the base of our cranium, limiting the room available in the pharynx for the SVT and esophagus. Besides constraining the length of our mouths (and therefore the horizontal portion of our vocal tracts), the most obvious consequence of this aspect of our anatomy is that our faces appear "flat" and "tucked in" compared to apes (as well as our early ancestors, the australopithecines).

Throughout childhood our tongues gradually descend into the pharynx, below the level of the lower jaw. As the tongue descends it carries the larynx down with it, a process that is not complete until we reach 6–8 years of age and we achieve the fully human 1:1 vocal tract. Only at this point are we able to produce the quantal vowels [i], [u], and [a], whose formant frequency patterns make them resistant to auditory confusion and, paradoxically, have stable formant frequency patterns that resist slight errors in articulation. In particular the vowel [i] is an ideal acoustic index of the length of a speaker's vocal tract—a factor necessary for deriving the phonemes encoded in the flow of speech. Without these quantal vowels, speech would still be possible, but less effective.

The advantages derived from having a vocal tract with 1:1 proportions are balanced, however, by a serious biological cost—the threat of death resulting from a blocked larynx. There is very little doubt that many thousands of incidents of fatal choking occurred in the human past. Even today about 500,000 Americans suffer from swallowing disorders (dysphagia), and, despite the invention of the Heimlich maneuver, death due to choking is the fourth-largest cause of accidental deaths in the U.S. (http://www.nsc.org/library/report_injury_usa.htm).

Given this high biological cost of possessing a 1:1 SVT it is likely that some form of speech already existed before our distinctly modern human vocal tracts evolved. In order for evolution to favor the development of a vocal tract configuration that also increases the likelihood of accidental choking, the evolving vocal tract was probably functioning in a way that bestowed its own advantages—presumably increased speech capabilities.



Although the adult human vocal tract is advantageous for speech production, it also exposes us to an increased risk of death by choking, for example, when the larynx gets blocked by food entering the trachea instead of the esophagus.

RECONSTRUCTING VOCAL TRACTS FROM FOSSILS

So when did fully human vocal tracts appear, and what can this tell us about our ancestors? To answer these questions we must first understand how the vocal tract is positioned in the head and neck, then see if we can identify when in the fossil record it "arrived" there.

As noted above, to produce the full range of human speech sounds the vocal tract must have horizontal and vertical portions of approximately equal length. But that is not the complete story. We also have to remember that those parts of our anatomy involved in speech—our tongue, the larynx, and the hyoid bone to which they attach—fulfill a more basic function in that they allow us to eat.

During swallowing, our hyoid bone moves upward and forward about 13 mm to open the esophagus, the pathway to the

Robert McCarthy

THE *FOXP2* GENE

Genetic research also provides insight into the evolution of human language capabilities. By studying members of an extended family who share a number of speech and language disorders (as well as cognitive and linguistic disabilities) researchers in England have identified the *FOXP2* gene.

This regulatory gene—sometimes labeled the "language gene" in the press—has been found to govern the embryonic development of neural structures that regulate motor control, aspects of cognition, emotions, and even the development of lung tissue. Individuals lacking the normal human variant of this gene are unable to position their tongues in a manner that allows for clear speech.

This evidence might suggest that modern human speech only appeared after this gene evolved into its modern normal variant. Interestingly, comparisons with the version of the gene found in mice and chimpanzees indicate a high degree of similarity to the human version. Only three mutations separate mice from humans, while only two separate us from chimpanzees.

Based on molecular genetic techniques, the human form of the gene seems to have appeared sometime in the last 200,000 years. This time frame corresponds with the emergence of anatomically modern humans, suggesting that this genetic variant may have conferred the increased motor control over speech that led to the later evolution of the specialized anatomy that makes modern human speech possible.

stomach. By doing so, it moves the larynx into a position that keeps food from falling into it. A larynx located in the neck can execute these maneuvers. In primates, including humans, the hyoid and larynx must reside in the neck, below the level of the jaw but above the sternum and clavicle bones of the chest. All else being equal, a hyoid positioned too high would rearrange the strap muscles that connect the hyoid to the underside of the cranium and interfere with their ability to elevate the hyoid. This configuration would seriously affect swallowing. Likewise, a hyoid and larynx positioned too low in the throat would cause the sternum and clavicle bones to impede the upward and forward motion of the tongue and hyoid, while the strap muscles connecting the hyoid and larynx to the sternum would cease to depress and stabilize the hyoid. This configuration would also seriously affect swallowing.

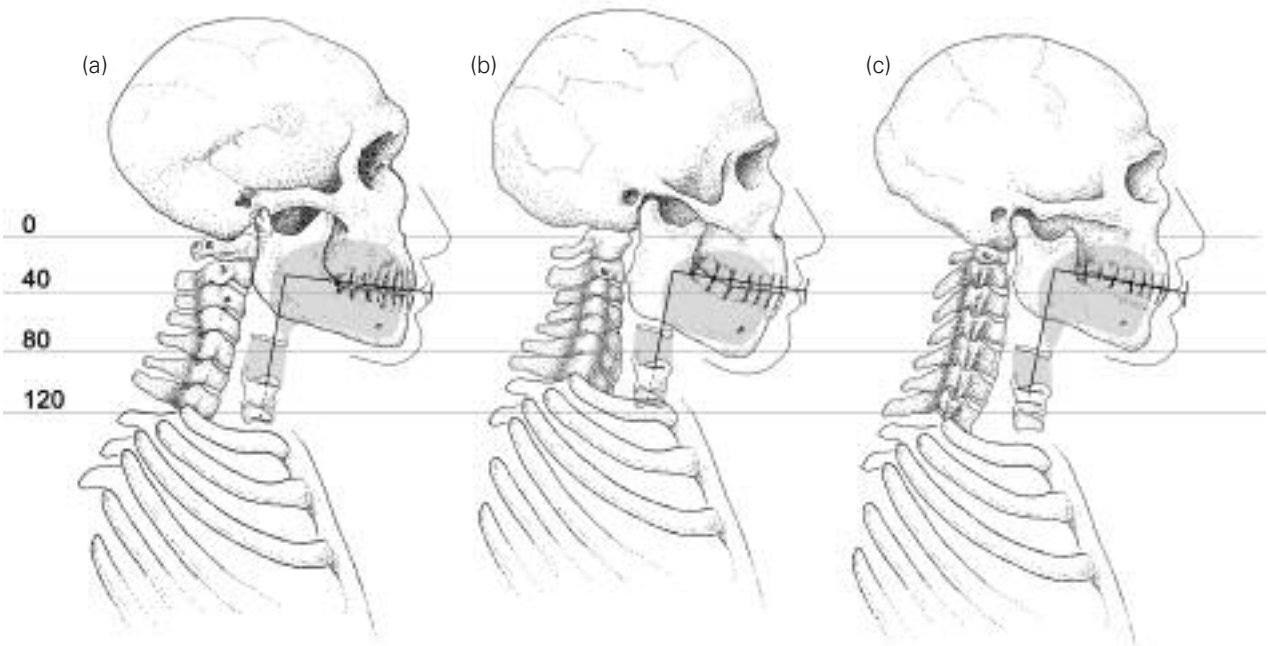
Using this information it should be possible to reconstruct the probable location of the vocal tracts of our hominid ancestors and such collateral relatives as the Neanderthals. To do this we examined the undersides of fossil skulls in order to determine the length of the horizontal portion of their SVT. Similarly, we used fossilized cervical vertebrae to reconstruct the length of their necks, which provides an important clue about the length of the vertical portion of their SVT.

The specific fossils measured included (1) a 1.6 million-year-old *Homo erectus* specimen believed to be ancestral to both Neanderthals and modern humans; (2) three Neanderthal specimens ranging in date from 70,000 to 40,000 years ago that are associated with Middle Paleolithic stone tools; (3) a 100,000-year-old early modern human specimen from Israel associated with Middle Paleolithic stone tools; and (4) eight more recent modern human specimens dating between 40,000 and 10,000 years ago that are found in association with more complex Upper Paleolithic tools. As a control, we also determined the relevant measurements for a large sample of modern chimpanzees and contemporary humans from seven different populations.

What we found was that Neanderthal necks were too short and their faces too long to have accommodated equally proportioned SVTs. Although we could not reconstruct the shape of the SVT in the *Homo erectus* fossil because it does not preserve any cervical vertebrae, it is clear that its face (and underlying horizontal SVT) would have been too long for a 1:1 SVT to fit into its head and neck. Likewise, in order to fit a 1:1 SVT into the reconstructed Neanderthal anatomy, the larynx would have had to be positioned in the Neanderthal's thorax, behind the sternum and clavicles, much too low for effective swallowing. Instead, these hominids likely possessed SVTs that had a horizontal dimension longer than its vertical one, suggesting that they would have been incapable of producing the full range of sounds made by humans today. Early hominids like *Homo erectus* and Neanderthals, therefore, would most likely have had SVTs intermediate in shape between those of chimpanzees and humans.

Surprisingly, our reconstruction of the 100,000-year-old specimen from Israel, which is anatomically modern in most respects, also would not have been able to accommodate a SVT with a 1:1 ratio, albeit for a different reason. Although it had only a moderately long face, its extremely short neck would have also placed its larynx too low in the chest if its SVT were equally proportioned. Again, like its Neanderthal relatives, this early modern human probably had an SVT with a horizontal dimension longer than its vertical one, translating into an inability to reproduce the full range of today's human speech.

It was only in our reconstruction of the most recent fossil specimens—the modern humans postdating 50,000 years—that we identified an anatomy that could have accommodated



Given the cranial reconstructions of (a) a Neanderthal dating to about 70,000 years ago, (b) a 100,000 year-old early modern human, and (c) a 26,000-year-old modern human from the Upper Paleolithic, we can reconstruct where the larynx would have to fit in order to achieve a 1:1 proportioned SVT. For the Neanderthal and the early modern human, their short necks and long faces would position the larynx in their thorax, or chest cavity, rather than their neck. This is clearly an untenable position (known as a "ghosted" larynx) and suggests that these two crania instead most likely possessed SVTs in which the horizontal portion was 30–60% longer than the vertical portion. As a result, neither would have been capable of producing the full range of speech sounds available to humans today. In contrast, the head and neck of the Upper Paleolithic human can accommodate a 1:1 SVT, indicating that by this date our ancestors were probably capable of fully modern speech.

a fully modern, equally proportioned vocal tract. Interestingly, the date of these specimens coincides with the appearance of the Upper Paleolithic tool kit, which is often associated with a florescence in modern human cognitive capacities.

If we assume that such fossil evidence indicates a 1:1 SVT that was capable of producing a full range of modern speech sounds, it seems logical to suggest that these Upper Paleolithic humans also had brains that were capable of sequencing the complex gestures necessary to produce speech. Taking this one step further (beyond what little hard evidence exists) it is likely that a brain so similar to ours would have possessed not only the capability to produce languages with complex syntax, but also cognitive flexibility. Therefore, we think that the presence of modern human vocal tracts sometime between 100,000 and 50,000 years ago marks the appearance of people with whom we might have had something to talk about.

PHILIP LIEBERMAN *is the Fred M. Seed Professor of Cognitive and Linguistic Sciences at Brown University, where he also is Professor of Anthropology.*

ROBERT MCCARTHY *is Assistant Professor in the Department of Anthropology at Florida Atlantic University.*

For Further Reading

Gardner, R. Allen, Beatrix T. Gardner, and Thomas E. Van Cantfort. *Teaching Sign Language to Chimpanzees*. Albany, NY: State University of New York Press, 1989.

Klein, Richard G. *The Human Career: Human Biological and Cultural Origins*. 2nd ed. Chicago, IL: University of Chicago Press, 1999.

Lai, C. S., D. Gerrelli, A. P. Monaco, S. E. Fisher, and A. J. Copp. "FOXP2 Expression during Brain Development Coincides with Adult Sites of Pathology in a Severe Speech and Language Disorder." *Brain* 126 (2003):2455-62.

Lieberman, Philip. *Toward an Evolutionary Biology of Language*. Cambridge, MA: Harvard University Press, 2006.

Negus, Victor. *The Comparative Anatomy and Physiology of the Larynx*. New York: Haffner, 1949.

Vargha-Khadem, Faraneh, K. E. Watkins, C. J. Price, J. Ashbruner, K. J. Aalcock, A. Connely, R. S. Frackowiak, K. J. Friston, M. E. Pembry, M. Mishkin, D. G. Gadian, and R. E. Passingham. "Neural Basis of an Inherited Speech and Language Disorder." *Proceedings of the National Academy of Sciences* 95 (1998):2695-2700.

Acknowledgments

Much of Philip Lieberman's research reported here was sponsored by NASA under grant NCC9-58 with the National Space Biomedical Research Institute. We are grateful to Elizabeth Murdoch for providing two of our figures (http://www.emurdoch.com/illustration.html).

Elizabeth Murdoch (top), Marcia Lieberman (middle), Cynthia Wilson (bottom)