Human speech
Thinking about the how and the why

Björn Lindblom
Department of Linguistics, Stockholm University, Sweden
lindblom@ling.su.se

Abstract

This text gives a brief account of why speech and hi-speed phonology are uniquely human. It provides a new view of sound structure – new in that it derives speech sounds from behavioural antecedents. In real-time phonology phonetic segments arise as *emergents* of open-ended vocabulary growth. The classical 'inescapable' form-substance distinction plays no role in this account.

An early start

The last common ancestors with the chimpanzees (LCA's) lost their rain forest habitat (> 4-6 Ma) and came under tremendous pressure to adapt to the new conditions. An entry point to speech seems to have been the tacit demand for improved cooperative communication, a cognitive skill that took on a life-saving role.

Mechanisms of change

For biologists the principal mechanism of evolutionary change is *natural selection* which builds cumulatively on existing capacities. It favors individuals with traits useful for survival. It resembles an amplifier operating also on traits that are only weakly present in the population.

When our ancestors acquired the ability to imitate and learn from each other they added a cultural dimension to selections. Culture is part of biology since cultural traits can determine the success and survival of individuals and groups, e.g., having language or not (Richerson & Boyd).

Habitat loss: A homeless ape

Long ago a period of major geological unrest occurred that drastically transformed East Africa. It went from a flat and jungle-like landscape to a drier and more open region with deserts, deep lakes, and high plateaus, volcanos and savannas. The disappearance of the rain forest made many species homeless - among them our ancestors, a population of ancient chimp-like apes, known as the first *hominins*. This transition forced our ancestors to come up with new routines for managing basic things such as finding food and water, coping with the hot sun, and protecting themselves and their young.

Since the savanna was sparse in familiar foods (fruits and plants) they significantly expanded the range of foraging. They made treks that took them far afield as indicated by the spread of tools left behind at feeding sites. Tools were carried along, handy should an abandoned carcass come in their way. Analyses of teeth microwear suggest a diet including more meat.

Natural selection appears to have favored those who were better at taking the heat and walking long distances. In response to the challenges their anatomy changed – (and their behavior, more anon). Over generations they adopted bipedal gait, got longer legs and taller bodies, lost body hair and added more sweat glands – all adaptations helping them expand their foraging range and manage the heat better.

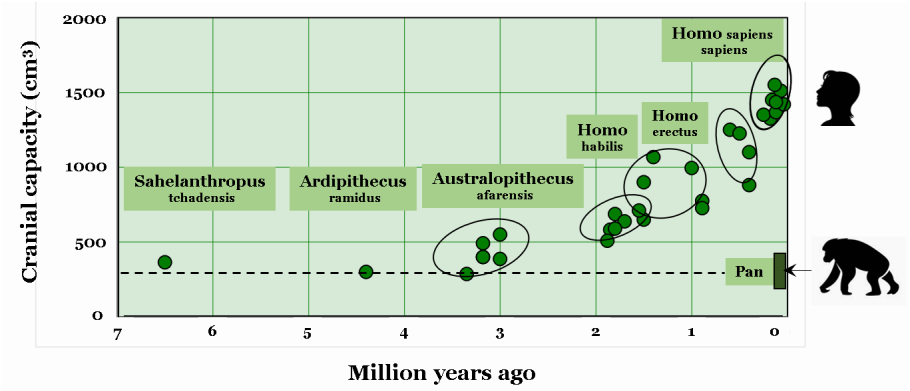

Fig 1. Fossil data on brain volume for five bipedal species. *H sapiens* data points are about three times as large as the average value for modern chimpanzees (Lieberman).

Where to hide and hunt

The oldest fossils (> 1,8 Ma) have been found along the East African Rift (EAR). The EAR is a gigantic gash in the Earth's shell with both deep lakes and higher grounds. It runs south from present Ethiopia down into Mozambique. The region shows an unusual number of volcanos and its soil is rich in eroded sediments and volcanic ashes, a favourable environment for the preservation of ancient fossil remains. One hypothesis suggests that our ancestors actively sought out places where the ground had been radically broken up by tectonic shifts and had been replaced by rough, hilly and mountainous terrain. That rough topography gave them places where they could hide and use for hunting. Animals could be tricked into entering paths in the rocky terrain where they would be easier to catch and kill. For instance, there would be plenty of opportunities for *ambush hunting.*

A disproportionately large brain

Brain tissue is metabolically expensive. Human brains use about 20-25% of the energy supplied by what we eat. For other animals' brain metabolism is considerably lower, less than 10 % for other primates and 5 % for non-primate mammals. These costs arise because potentials across axonal membranes must be maintained and the continual synthesis of neurotransmitters requires energy. It follows that the larger the number of neurons in a brain, the greater the need for metabolic fuel. The human brain does not get its energy from an increase in the basal metabolic rate. Neither does it "steal" energy from other expensive organs such as heart, kidneys, liver and the gastro-intestinal tract.

Rather, for evolution to produce a disproportionately large brain - a brain that is larger than what would be expected from the animal's body weight - a drastic increase in the nutritional value of the diet would have been needed. In short, that is exactly what evolution delivered. The hunting scenario just mentioned indicates how hominins may have come across the necessarily large amounts of animal foods needed for this disproportionate brain growth.

From innate to learned signals

How did the hominins communicate? What were they like? Fossil and molecular evidence suggests that they were similar to today's chimpanzees in anatomy and lifestyle (Pilbeam & Lieberman). Psychologists have shown that young children use pointing communicatively before they can express themselves in speech. This behavior reflects a

cognitive maturity that chimps lack. Chimpanzees are not motivated to cooperate and share; they do not understand communicative intentions; they follow gaze, but they do not establish the joint attention and common communicative ground needed to understand the meaning of gestures. Hence chimp-like gestures would not have been a useful platform for improving communication.

The vocal system of the common chimpanzee consists of a fixed, small number of innate signals that are reflexively triggered by internal or external stimuli. Vocal learning and vocal imitation do not seem to occur. Remarkably, all of these characteristics are also found in the vocal behaviour of other apes. The similarity includes signal categories labeled: *hoots*, *whimpers*, *screams* and *squeaks*, *grunts* and *pants* (Slocombe & Zuberbühler). This cross-species match is surprising in view of the millions of years that separate their respective speciations. It implies considerable evolutionary stability. That is an important observation because it strongly supports assuming that *the vocal skills of our earliest forebears did not include vocal learning or an ability to imitate vocal signals.*

Volitional VT motor control

We cannot speak to the details of how our ancestors' cognitive skills and social behavior evolved. But, in broad strokes, we can suggest a few stages of a speculative, but plausible scenario.

Under strong threat hominins probably behaved like modern apes: They put their competitive, individualistic personalities on hold and responded by increasing troop cohesion. Their talent for 'togetherness' was reinforced by the enlargement of their brains which were fuelled by the new diet. New neurons and synapses were pruned by genes and experience and maintained according to the 'use-it-or-lose-it' rule. Evolution seems to have met the tacit demand for improved communication by building an extension to the existing sound production mechanism. The result was two coordinated modules both connected to the motoneurons of vocalization: one emotionally triggered (old) and one providing full volitional control via direct cortical connections to a broad array of motoneurons including those for articulation, phonation and respiration.

In retrospect, this development was a milestone along the path to speech. It was a *game changer* in that it paved the way for new skills such as vocal imitation and vocal learning. Once in place it reshaped human cognition promoting a collective behavior and mentality with intersubjectivity, mutuality and shared intentionality. It also gave culture a significant role to play in the evolutionary process.

Still far from speech but ….

At this point in our narrative, let us assume that our ancestral ape is a bipedal with a smaller, retracted jaw, less posterior foramen magnum, lowered larynx and rounded (non-flat) tongue (Lieberman). He has a bigger brain and a more advanced cognition. He has just begun to experiment with his VT, making noises different from his own reflexive calls. He has also started to imitate others. He is more social and multi-modally communicative than his predecessors, but he and his conspecifics are still far from modern speech.

Symbolic reference

Animals communicate vocally. Vervet monkeys make alarm calls. Male humpback whales sing, but they do not invent symbolic signals (Deacon). Human children do. Our infants come highly motivated to interact socially. They seek people's attention. Not yet able to speak, they often use pointing in combination with grunts and other vocalizations. Over time the gestural component is omitted and the voice alone does the "pointing". This vocal activity is first triggered by context. The object referred

to has to be present. Then the behavior becomes more abstract. When the child uses her words without a contextual prompt, she has had the 'nominal insight'. She says 'ball or 'doll' also in the absence of these objects. She understands that 'things have names' and has thus grasped the notion of *symbolic reference* (Vihman).

It does not seem unreasonable to picture our ancestors' first use of signals going through a similar process. As their cognitive capacities grew, they could represent their ecological and social environment in greater detail. Expressive needs became stronger. A shared, open-ended space of semantic information emerged. How could a matching open-ended production mechanism evolve – one that linked a distinct sound pattern to every possible meaning? Problem: The current view is that primates show open-ended comprehension, but have highly inflexible production skills (Seyfarth & Cheney).

Exploration

Cortical control of the VT offered new ways of making acoustic signals. Initially all modalities may have been used but, as hand and body messages grew more elaborate, they eventually reached a complexity that favored faster and more precise ways of communicating. The vocal/auditory modality offered an independent, omnidirectional channel useful at a distance and in the dark. It did not impede locomotion, gestures, or manual work (Donald).

New meaning-carrying signals were derived and tested. All troop members would participate contributing their own attempts and imitating each other. From this collective activity there emerged a distribution of shared signals, *selected*, I assume*, on the basis of their articulatory, perceptual and cognitive merits.* This explorative activity revealed that the VT is not an indivisible whole, but is biomechanically made up a number of semi-independent subsystems (larynx, lips, soft palate, tongue blade and body). They came across different sound sources (voicing, fricative noise, transients) and found that their acoustic output could be varied by manipulating articulators and modifying the shape of the VT.

Segmentation

Out of the large number of signals that our ancestors are certain to have produced, evolution seems to have singled out the *jaw-based babble* as a template for expanding vocabularies.

Fig 2. Stylized 'chunky' waveform envelope produced by opening and closing the VT. Claim: This quantal singularity is the source of modern vowels and consonants (MacNeilage).

So where do discrete phonetic segments come from? Because (i) up-and-down jaw movement had a uniquely large quantal and salient effect on the acoustic signal; (ii) because co-opting the jaw CPG network (Grillner) minimized motor programming; (iii) because using jaw oscillation as a carrier solved the serial organization problem (Lashley); (iv) because the segments of the babble were ideally suited for making combinatorially organized signals.

Boot-strapping imitative skills

Exploration also served the purpose of boot-strapping vocal imitation. It has been proposed that the child's babbling gives the brain an opportunity to map the possibilities of the vocal system (Vihman). Articulation is combined with phonation in a variety of ways. The motor learning that takes place then is absolutely basic to the acquisition of speech. An auditory-motor mapping takes place. As sound-producing movements are repeated again and again, a strong link is

forged between tactual and kinesthetic impressions and the auditory sensations that the child receives from his own utterances.

Deriving motor commands

When the child then wishes to replicate sounds spoken by others, she has at her disposal a store of orosensory and auditory associations to consult in deciding how to program her own vocal system. For example, when imitating a certain vowel, the motor commands are shaped by the afferent pattern stored for that particular vowel - its AFF-ID. In words, the brain's message is: Move articulators so as to produce a match between the incoming afferent pattern and the vowel's stored AFF-ID.

Invariants

Invariant attributes for a prototypical speech sound are specified at the orosensory (tactual and kinesthetic) level. This choice solves the classical invariance problem because, in the speech chain, the orosensory stage comes before the brain imposes coarticulation on the motor commands.

Differentiation

If mandibular oscillation served as the template for vocabulary expansion, how were template slots phonetically differentiated?. A primary constraint on these selections was perceptual: *Different meanings must sound different*. Phoneticians have concluded that the distribution of the world's vowel qualities is governed by a *principle of dispersion* which tends to drive vowels apart in the acoustic vowel space. Vowels tend to be single-constriction articulations. Articulatory modeling and physiological analyses have shown that the space for vowels is largely determined by the tongue's three extrinsic muscles plus the jaw, the larynx height and the lips. Numerical models of tongue shapes accurately generate arbitrary vowels as a linear combination of the relative activities in m. genioglossus, m. styloglossus and m. hyoglossus. Informally, these three could be called the [i]-muscle, the [u]-muscle and [a]-muscle.

The UPSID data base lists a total of about 650 consonants observed in 451 languages. Individual languages converge on inventory sizes between 20 and 37 consonants. Remarkably they make similar selections. The bilabial, dental/alveolar and velar places of articulation is a preferred choice (Maddieson). As systems get bigger more places get added, but the stronger trend is to reuse those three places in combination with different sources and manners. Differentiation of the dental/alveolar place has a particularly long lists of possibilities. The basic mechanism in building a consonant inventory is reuse which modifies existing patterns by adding new dimensions (e g, phonation types, as in Fig 3, bottom), or by making new small incremental changes (enhancing stops by making releases aspirated / ejective, or making their voicing implosive, Fig 3, top).

NAMBIQUARA

	bil	alv	velar	bil velar
voiceless	p	t	k	kW
vl aspirated	ph	th	kh	kWh
vl ejective	p'	t'	k'	kW'
vd implosive	b<	d<		

Hindi-Urdu

	bil	dent
plain	m	n̪
breath	m^h	$n̪^h$

Fig 3. Reuse in stop system (top), nasal system (bottom).

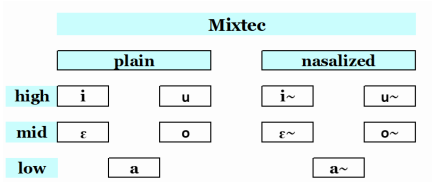

Fig 4. Dispersion and reuse. 10-vowel system of Mixtec.

The above remarks suggest that vowel and consonant inventories may be constructed differently. That is in not

necessarily the case. When a large vowel system is needed differentiation uses both dispersion and reuse (Fig 4).

Neural reuse

Neuroscience shows that it is quite common for neural circuits established for one purpose to be exapted during evolution, or recycled during normal development. Adding a new function need not require drastic re-wirings, only new connections and/or sharing of existing circuitry (Anderson). A question for future research: Is reuse during phonetic learning facilitated by neural reuse?

Easy way sounds OK.

All languages have syllables. All languages have speech sounds built at specific places of articulation and differentiated by their manners of articulation; Manner is here construed as a specific combination of articulatory subsystem activities. Possible interpretation: Syllables and the patterning of speech sounds offer phenomena that children are likely to come across by chance during their VT exploration. Hence, they are universally present in languages to make phonetic learning easier. *Easy way sounds OK*!

Devil in the msec: Coarticulation

In adult speech, articulatory movements have time constants that make the duration of a vowel gesture (activation + deactivation) longer than its corresponding acoustic segment duration. In real-time phonology that difference shows up as soon as articulatory position control of segment content is applied to the template. Hence coarticulation is present early in life and is a very ancient property of speech.

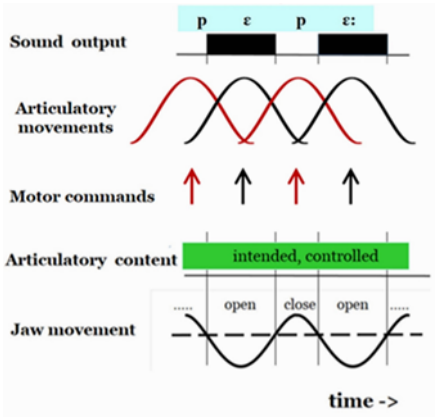

Fig 5 A schematic of a 9-month-old child's attempt at saying '*baby'*. The result was [pɛpɛː]. Note the timing of motor commands (arrows). Also note that the movements (bell-shaped curves) are longer than the acoustic segments.

References

Anderson M L (2010). Neural reuse: A fundamental organizational principle of the brain, *Behavioral and Brain Sciences* 33, 245–313.

Deacon T W (1997). *The symbolic species and the co-evolution of language and the brain*, Norton:New York

Donald M (2001). *A mind so rare*, Norton, New York

Grillner S (2023). *The brain in motion*, MIT Press, Cambridge MA

Lieberman D E (2011). *The evolution of the human head*, Harvard University Press, Cambridge MA

MacNeilage P F (2008). *Origins of speech*, Oxford University Press, New York.

Maddieson I (1984). *Patterns of sound*, Cambridge University Press, Cambridge.

Pilbeam D R & Lieberman D E (2017). Reconstructing the Last Common Ancestor of chimpanzees and humans, 22-141 in Muller M N et al (eds), *Chimpanzees and human evolution,* Harvard University Press, Cambridge MA

Richerson P J & Boyd R (2005). *Not by genes alone,* Chicago University Press: Chicago.

Seyfarth R M & Cheney D L (2010). "Production, usage, and comprehension in animal vocalizations", Brain & Language 115, 92–100

Slocombe K & Zuberbühler K (2010). “Vocal communication in chimpanzees”, pp 197-207 in Lonsdorf E V et al (eds): *The mind of the chimpanzee*, Chicago University Press:Chicago.

Stevens K N (1998). *Acoustic Phonetics*, MIT Press, Cambridge MA

Tinbergen N (1963). "On aims and methods of ethology," *Zeitschrift für Tierpsychologie*, 20: 410–433.

Vihman M M (2014). *Phonological Development*, WILEY & Blackwell