Embedded Structure and the Evolution of Phonology*

J.C. Brown & Chris Golston

University of British Columbia & California State University, Fresno

J.C. Brown (corresponding author)
Department of Linguistics
Buchanan Building Block C
1866 Main Mall
Vancouver, BC V6T 1Z1
Canada

(jcb@interchange.ubc.ca)

Chris Golston
Department of Linguistics
California State University, Fresno
Peters Business Building, Room 383
5245 North Backer Avenue
Fresno, CA 93740-8001
USA

(chrisg@csufresno.edu)

Abstract

This paper explores a structure ubiquitous in grammar, the embedded tree, and develops a proposal for how such embedded structures played a fundamental role in the evolution of consonants and vowels. Assuming that linguistic capabilities emerged as a cognitive system from a simply reactive system and that such a transition required the construction of an internal mapping of the system body (cf. Cruse 2003), we propose that this mapping was determined through articulation and acoustics. By creating distinctions between articulators in the vocal tract or by acoustic features of sounds, and then embedding these distinctions, the various possible properties of consonants and vowels emerged. These embedded distinctions represent paradigmatic options for the production of sounds, which provide the basic building blocks for prosodic structure. By anchoring these embedded structures in the anatomy and physiology of the vocal tract, the evolution of phonology itself can be explained by extra-linguistic factors.

Keywords: language evolution, consonants, vowels, embedding, phonology

Running Title: Evolution of Phonology

Biographical Note: J.C. Brown is a graduate student in the department of linguistics at the University of British Columbia. Chris Golston is an Associate Professor in the department of linguistics at California State University, Fresno.

1. Introduction

Recent proposals in the literature on language evolution have suggested that much of syntax evolved from phonology (cf. Carstairs-McCarthy 1999). Such models have benefited the way in which the modular interactions in language evolution are approached; however, they still leave the ultimate origins of language to be explained. The spirit of this paper adopts the 'exapted structure' approach to the evolution of language, the fundamental difference being that the structure we propose is grounded in non-linguistic domains such as physiology. We attribute much of the evolution of speech production to the embedded tree, or what we term the 'treelet'. Here we try to show that embedded structures arise naturally from internal maps of the vocal tract and what one can profitably do with it. Not all parts of the vocal tract are well modeled with a treelet, but enough of them are to make treelets a good way of representing much of the speech apparatus and its output.

Embedded trees are ubiquitous in grammar and give it its hierarchical structure. We suggest that such treelets were exapted from articulation and acoustics into other grammatical spheres to lend coherence to the messages that the sound system was being used to communicate. The way this took place is probably through what Cruse (2003) terms a shift from a reactive system to a cognitive system. Reactive systems are simply that: their behavior is guided by reaction to external stimuli. Reactive systems have no planning capabilities, and cannot take measure of events over the course of time into the future. Cognitive systems, on the other hand, have the ability to plan their behavior ahead of time. This is what human speakers are: cognitive systems endowed with the capacity to plan speech events in real time, not merely systems that react to stimuli (Chomsky 1959). The big question is how speech could have evolved from a reactive to a cognitive system. The evolution from a reactive system to a cognitive system requires the construction of an internal map of the body (Cruse 2003), and we take this internal mapping to be the launch of the evolution of the treelet. New developments in the study of human and non-human primate neurology also indicate that such a system mapping is a plausible source for the origins of language. The discovery of 'mirror neurons' in primates (Gallese et al. 1996, Rizzolati et al. 1996), neurons which fire not only when an action is executed, but also when it is *observed*, provides a biological foundation for the mapping processes we are proposing. Furthermore, the areas in the brain where motor neurons have been discovered (ventral premotor cortex, area F5 in macaques[i]) control not only grasping actions of the hand, but also 'grasping' actions of the mouth

(Fogassi & Gallese 2002, Rizzolatti & Arbib 1998). This has deep implications for the model we are proposing, as the physiology of the vocal tract will become central to the function of mapping during the evolutionary process. The evolution and exaptation of the treelet is responsible for our ability to articulate consonants and vowels. Taking this step-by-step process, once the main building blocks of speech are in place, the rest follows by exapting the treelet into syllables, feet, and syntax; this is why it is crucial to determine how consonants and vowels arose.

Human speech packs an incredible amount of information into hierarchical structures. At the base we find the individual distinctive features (gestures) used to distinguish meanings: nasal, labial, low, round, etc.; nearer the top we find syllables. The preceding paragraph, for instance, has at least 4587 individual gestures packaged into 1529 speech sounds. If you can read the passage in 2 minutes you've produced at a minimum some 38 gestures per second, each of which is informative in deciding what the message means.

We're interested here in showing just how similar many of these trees are, specifically with how distinctions tend to embed in a similar way, with two binary branchings defining a three-way split. We begin here with phonetic and phonological distinctions used in speech and note that a large number of them involve a basic distinction between two categories (eg. [lingual = [coronal dorsal]]). Such dichotomies in phonetic and phonological distinctions are much more common than ternary distinctions with no sub-grouping, or quaternary distinctions with elaborations on both sides of the initial split.

Most of the distinctions we'll encounter here are paradigmatic, different optionals (like labial—coronal—dorsal) that one can take for a given parameter (like place of articulation). The little trees we'll now look at do not generally define syntagmatic, linear relations in language. Thus we will propose that both the paradigmatic and the syntagmatic aspects of language (Saussure 1916) have phonetic and phonological precursors, specifically consonants and vowels (paradigms) and syllables and feet (syntagms) (cf. Brown & Golston 2002); however, the focus of the present paper is on the former, and not the latter. For now, let us see how more basic phonetic and phonological distinctions break down.

2. Articulation of Consonants

Every sound must have a source and for most of us that source is pulmonic (from the lungs) and egressive (pushed out of the body). This is not limited to humans, for 'all vertebrates use their respiratory system to set the sound

production machine in motion' (Hauser 1997, 133). In anurans (frogs and toads) the source is the larynx (Rand 1988), as it is in mammals; birds have a different structure (the syrinx) but also use it as a sound source, the communicative purposes of which are in song. The sound waves produced at the source are of course filtered by the resonating cavities above the larynx, but this filtering need not have any communicative function. Thus in anuran advertisement calls there is effective use of pitch modulation with little or no filtering in the supralaryngeal part of the vocal tract:

> When one thinks of a calling frog, one typically imagines an individual with an inflated air sac. Although the air sac (which receives its air from expiratory forces generated by the lungs) plays a role in sound amplification…, the laryngeal cavity (muscles, vocal cords, cartilages), or sound source is sufficient for call production; experimentally puncturing the sac does not perturb the spectral properties of the call, but does reduce the sound-pressure level. (Hauser 1997, 115)

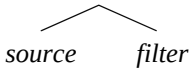
We might then graph the basic vertebrate sound production system like this, where 'source' stands for the larynx or syrinx:

(1) *sound*
|
source

The source (larynx or syrinx) makes different types of sound by being blown apart slowly or quickly by egressive lung air, giving a different *pitch* (measurable in terms of cycles per second, or Hz). As the size of the larynx correlates roughly with the size of the body in which it is housed, pitch height can be used to show aggression (low pitch) and submission (high pitch), and is so used in many species including humans.

Some kinds of birds can do more than just emit sounds. They can also *filter* them by manipulating the shape of the vocal tract above the syrinx (Nowicki et al. 1992; Westneat et al. 1993, Suthers & Hector 1988). Many mammals do this too and the gold medalists are humans. For us, sound is made like this:

(2)

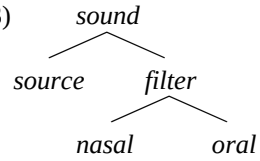


In most mammals there's not a lot of filtering you can do because the larynx empties directly into the nasal cavity; so the filter is essentially the nose.

Distinctions among nasal sounds are fairly limited, because there are no structures within the nose that you can move around. You can seal the mouth off at different locations, say the lips or palate, shunting some of the air into what Laver (1994, 211) calls a 'cul de sac behind the oral closure'. This will create a labial [m] or a lingual [n], but it's difficult to perceive distinct places of articulation among nasals, because the resonating chambers in the mouth get filtered again through the nose. So place of articulation, though articulatorily possible, probably wasn't very useful when humans only made nasal sounds.

But at some point in hominid evolution, the larynx descended so far into the throat that lung air could be forced through the mouth instead of the nose. Our species essentially developed a detachable nose. Once the larynx and nasal cavity were parted, sound-waves could be routed out of the nose *or* out of the mouth (Liberman 1984), and the filter split into a nasal (old) and an oral (new) part:

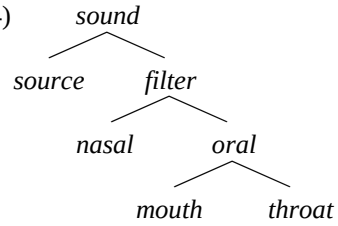
(3)



We now have a nicely embedded way of producing speech, modeled as the source/filter theory of speech production (Fant 1960). The main split is between (laryngeal) source and (supralaryngeal) filter, but the latter is itself split into a nasal and an oral part. The nasal/oral distinction is embedded under the filter part of the source/filter split.

As the larynx continued to drop, there opened up in its wake a new resonating cavity, the pharynx (throat). This changed the basically flat oral cavity into an L-shaped tract with two resonating cavities (Lieberman 1984), the old one above the tongue and a new one behind it (the pharyngeal cavity):

(4)



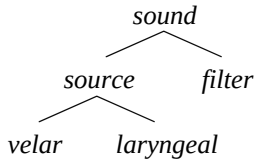
This changed everything. The vocal tract now had a source for continuous sound (the larynx) and *three* resonating chambers above the larynx to filter it (nose, mouth, throat). This had major consequences both for vowels and for consonants. And for choking to death: with a throat, there was now a shared space above the entrances to the stomach (esophagus) and lungs (larynx). Darwin noted

> the strange fact that every particle of food and drink we swallow has to pass over the orifice of the trachea, with some risk of falling into the lungs, notwithstanding the beautiful contrivance by which the glottis is closed (Darwin 189, 191).

But it also gave humans the ability to communicate clearly a huge number of phonetic distinctions that could be used to digitally code any simple meaning into an utterance. So let's go down the branches in the tree above and see what humans could do at this point. We start, appropriately, at the source.

Most speech is made by pushing air out of your lungs (pulmonic egressive), but you can also make noises by sucking air into your mouth (velaric ingressive). So the physical source of speech divides neatly in two[ii]:

(5)



Velaric ingressives are rare as speech sounds, but do occur in a number of south African languages as 'clicks' (Ladefoged & Trail 1984). Clicks are made by closing off the mouth at two points, the velum and one other place. While both articulations are held, the tongue body moves down, which lowers the air pressure in the oral cavity. When the secondary closure is released air gets rapidly sucked into the mouth creating a special sound where the

second closure occurred. Clicks can be made at a number of different places of articulation and come with their own phonetic symbols as follows:

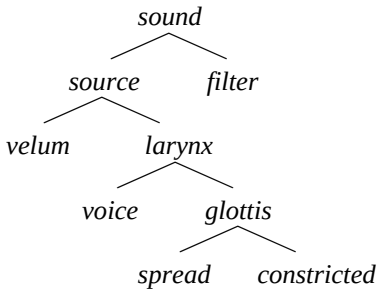
(6) Clicks

labial	⊙
dental	|
alveolar	!
post-alveolar	!
alveolar lateral	‖
palatal	ǂ

These are probably not the first six consonants of human language, of course (see for instance Herbert 1990), but it's important to see that even with little else than two moving articulators (lips and tongue) a number of distinctive sounds can be produced.

The primary source for speech is of course the larynx, and the contrasts one can make with the larynx have been extensively studied (Lombardi 1995; Iverson & Salmons 1995; Ladefoged & Maddieson 1996; Kehrein 2002). The laryngeal specifications used in contrasts are generally taken to be three: vibration of the vocal cords (voice), aspiration (spread glottis) and creakiness or ejection (constricted glottis). Spread and constricted cannot co-occur[iii], but each can co-occur with voice giving breathy voice (spread, voice) and implosion (constricted, voice). Thus it makes phonological sense to pair spread and constricted together and to make each the niece of voice, as is in fact commonly done in some models of laryngeal features:

(7)

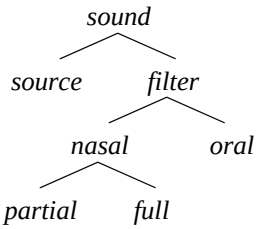


This treelet does not serve as a map of the larynx and thus is purely paradigmatic; indeed, it shows types of laryngeal feature that cannot all be distinctively ordered within the same speech sound. The lowered larynx probably played an important role in the evolution of language anyway (Lieberman 1984), but the present context

suggests an additional role it may have played: a precursor, along with other phonetic parameters, for the embedded treelets found elsewhere in grammar.

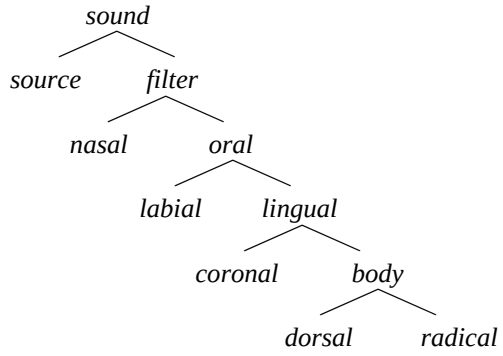
So let's turn now to the filter, with its three cavities: nasal, oral, and pharyngeal. The nasal cavity doesn't allow for a lot of distinctions, and most languages make no more than a two-way contrast between nasal (m) and oral (b) sounds. That said, there are languages that do more. Some languages contrast oral (b), prenasalized (^{m}b), and fully nasal (m). So it looks like the maximum use that can be made out of the nasal cavity is as follows:

(8)



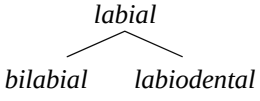
Returning now to the oral cavity, it is the mouth proper that has the most possibilities for clear distinctions because it has two major articulators, the lips and the tongue. Once the larynx has descended far enough to give us a throat, we can distinguish two parts of the tongue, crown and body, the latter divided into dorsum and root:

(9)



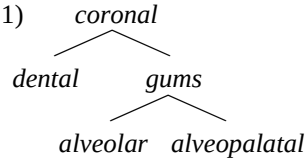
Labial consonants can be bilabial (p) or labiodental (f). In each case the moving articulator is the bottom lip: it touches either the top lip (p) or the top teeth (f):

(10)



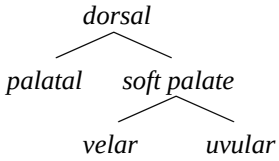
Moving back inside the mouth we find the other major articulator, the tongue. The crown of the tongue is the most versatile and it can touch a number of other places in the mouth including the teeth, and the (top) gums, either on the alveolar ridge (alveolar) or where the alveolar ridge contacts the palate (alveopalatal)[iv]:

(11)



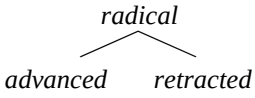
The dorsum is also fairly mobile and is used to make palatal sounds (Spanish *ca<u>ñ</u>on*, German *mi<u>ch</u>*), velars (English k, g) and uvulars (Hopi q, French ʁ):

(12)



The tongue root is much less flexible because it's connected to the mandible, but it can be advanced (ATR) and retracted (RTR) to make distinctions like those found in Arabic pharyngealized consonants.

(13)

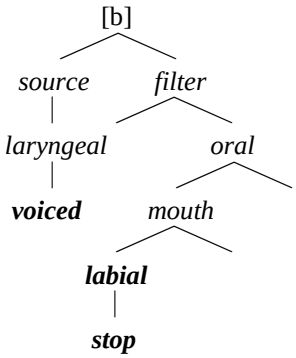


So far we have seen how the various components of the vocal tract have developed over the course of evolution: source (like anurans); filter (like birds); nasal, oral and pharyngeal cavities (unique to humans); lips and the various parts of the tongue and so on. We have seen that every sound requires a source and a filter; that the source and filter can both be modeled using treelets; and that various places in the mouth are used to filter sound ('places of articulation'). To round out the picture of how speech is produced we need to consider aperture as well, the way that constrictions are formed in the vocal tract. This issue is somewhat orthogonal to place of articulation, since it deals with the *how* of articulation rather than the *where*; but the two are of course intimately related since

you cannot make a constriction without making it somewhere. To show the connection we'll use little lines to connect a given place of articulation (lips, crown, etc.) with a given aperture.

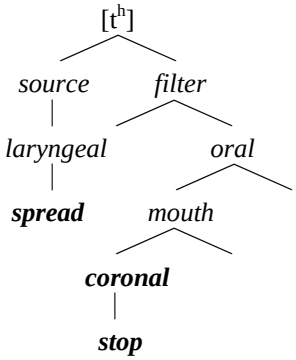
There are three degrees of openness or aperture: zero, fricated and maximal (Steriade 1994, etc.). Zero aperture involves a complete obstruction in the vocal tract that causes silence. Zero aperture at the lips gives you a labial stop (*apa*); at the teeth a dental stop (*ata*); at the soft palate a velar stop (*aka*). These plain stops can also be distinctly voiced (*aba*, *ada*, *aga*), aspirated (*ap^{h}a*, *at^{h}a*, *ak^{h}a*), or glottalized (*ap'a*, *at'a*, *ak'a*). To see how this works in the vocal tract, consider the medial sounds of *aba*, *at^{h}a*, *ak'a* below, where only the active articulations are shown. The medial consonant in *aba* has a voiced laryngeal source with zero aperture ('stop') at the lips:

(14)



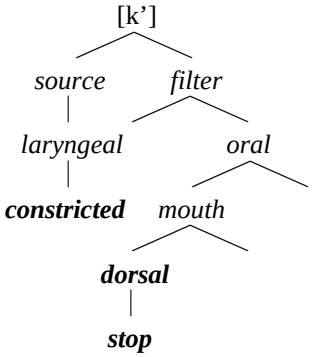
The medial consonant in *at^{h}a* has the glottis spread wide open for an aspirated laryngeal source of sound and has complete closure made by the crown of the tongue:

(15)



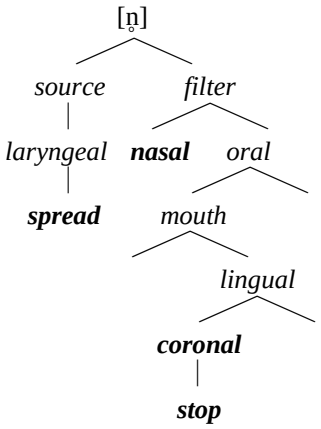
Finally, the ejective sound in *ak'a* has a constricted laryngeal source and complete closure made by the tongue dorsum:

(16)



Stops can also be made with nasal airflow by using zero aperture in the mouth to force air through the naso-pharyngeal port, giving us fully nasal sounds like the consonants in *moaning*. Nasal stops like this are usually simply voiced in language, but aspirated stops can be found in languages like White Hmong and Burmese (Ladefoged 1971) and glottalized nasals can be found in languages like Jalapa Mazatec (Kirk, Ladefoged & Ladefoged 1993) and Danish (Fischer-Jørgensen 1985). An aspirated coronal nasal stop (n̥) looks like this:

(17)



Children easily make stops before they make other consonants because they involve an extremely simple production strategy: throw an articulator against part of the upper vocal tract. What's true for children was most likely true for our ancestors in this regard; they must have started off producing stops, both oral and nasal. All modern languages have stops at multiple places of articulation. White Hmong makes use of some thirty-two distinct stops of this kind, including plain (t), aspirated (t^h), voiced (d), voiced aspirated (d^h), prenasalized (^{n}d), prenasalized aspirated ($^nd^h$), nasal (n) and nasal aspirated (n̥) consonants at six places of articulation:

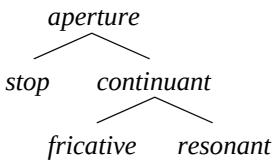
(18) White Hmong stops

Labial	Dental	Retroflex	Palatal	Velar	Uvular
p p^h	t t^h d d^h	ʈ $ʈ^h$	c c^h	k k^h	q q^h
^{m}b $_m$$b^h$	^{n}d ^{n}d^h	^{n}ɖ ^{n}$ɖ^h$	^{n}ɟ ^{n}$ɟ^h$	^{ŋ}g ^{ŋ}g^h	^{N}G ^{N}G^h
m m̥	n n̥		ɲ ɲ̥		

(White Hmong has more stops than this but they involve additional parameters we haven't discussed yet). When it was that humans became able to make many types of stops like those of Hmong is another question, but the individual gestures (spreading the glottis, letting air into the nose, touching the tongue to the teeth, etc.) are extremely old and controlled by separate groups of muscles. So there's no doubt that early hominids *could* have pronounced sounds of this complexity. The planning involved in making the sounds directly above is no more than the planning involved in swallowing (human swallowing being fairly complex as swallowing goes). So we can safely postulate a stage in the evolution of language where sequences like *bama* and *mebi* were all the rage. The up and down oscillation of the mandible in producing stops and then vowels is surely the precursor to the syllables that package sounds into groups. This is true for modern babies (MacNeilage & Davis 1990, 1993, 1999) and was surely true for early hominids for the same reasons (MacNeilage 1998; MacNeilage & Davis 2000). Thus, the first sounds children produce are zero-aperture stops (Locke 1983; Gildersleeve-Neumann, Davis & MacNeilage 2000) and maximal aperture vowels (including semi-vowels). We'll deal with maximal aperture sounds below, after we treat the non-stop consonants that children (and the earliest hominids) find so difficult to make.

Non-stop sounds are called continuants (because they can be produced continuously, with no cessation of sound), and there are two types of them, fricatives and resonants:

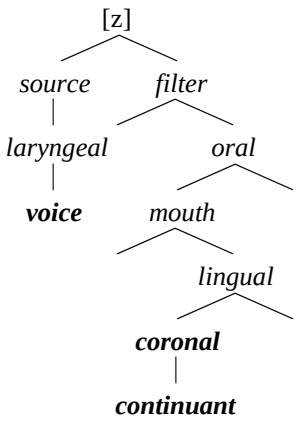
(19)



Fricatives are made with a fairly narrow aperture (AF) and include all of the consonants in *fifths* [fɪfθs]. They can be pronounced at many different places of articulation and can be produced with voicing, creaky voice and even aspiration (though aspirated fricatives are rare because fricatives require a lot of aspiration to begin with). English

has more fricatives than it does stops with two labio-dentals (f, v), two dentals (θ, ð), two alveolars (s, z), and two alveo-palatals (ʃ, ʒ). The first sound in *zen* looks like this:

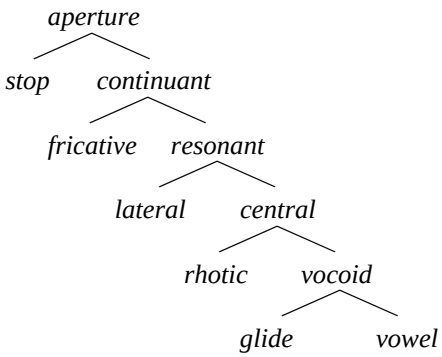
(20)



Little children have an aversion to fricatives because it's hard to control the fine-tuning that creates a jet of air with the moving articulator. Many children pronounce fricatives as stops saying *tee* for *see* and *tad* for *sad*, for instance. For this reason (articulator difficulty) we can safely assume that early hominids produced stops before they produced fricatives.

The other continuants are resonants, made with little or no occlusion in the vocal tract. The resonants branch off into lateral (l-sounds) and central resonants, the latter of which consist of r-sounds (rhotics), semi-vowels (glides) and vowels. The entire picture looks like this, with aperture increasing from zero (stop) to maximal (vowel) as you go down the tree:

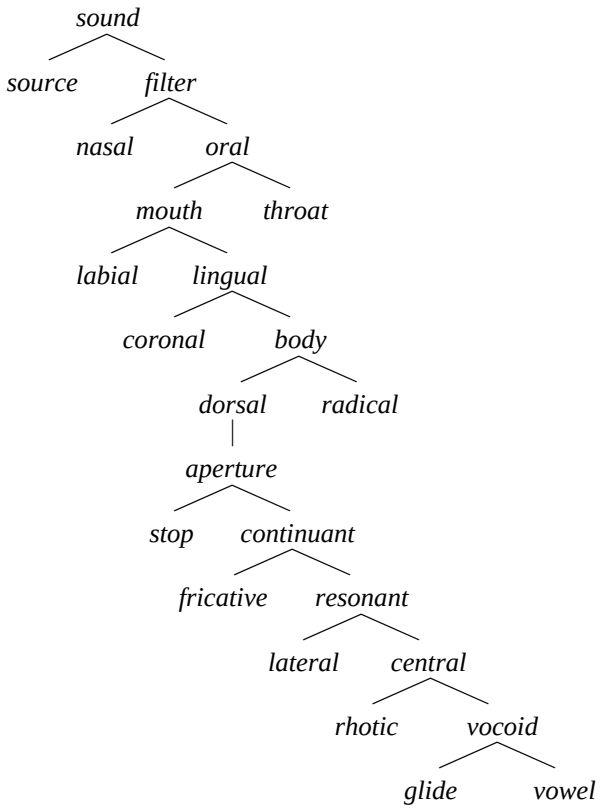
(21)



Laterals are made with a partial occlusion on one side of the mouth, while central resonants are made with a partial occlusion along the midline. Lateral and rhotic sounds aren't easy to make either: children often have great difficulty mastering them and languages usually have very few of them among their consonants. But as the vocal tract opens up more fully sounds get easier to say: glides and vowels are produced very early on in babbling, and we can safely assume that they were produced early on in the evolution of language.

Thus at one end of the spectrum we have silence (vocal tract completely closed), at the other end very loud sound (vocal tract wide open). We can assume that early hominids made only a two way split (stops vs. vocoids) and that the sounds made with partial constriction were tucked in between the two extremes as our vocal abilities evolved. To wrap up this section on consonant articulation, we have to see that place of articulation is tied to aperture, each dependent on the other at a point we might call the great connect:

(22)



the great connect

We can also consider the great connect a division point: the top half of the great divide is essentially a physiological map of the vocal tract in terms of a source and filter; the bottom half a map of the different degrees of aperture at

which the filter can be set. Linking the two halves of the great divide together is required for making any single speech sound.

We have seen so far how the ability to articulate consonants arose through the patterned distinctions that are made based on the treelet. In terms of our most basic communicative abilities, we share much in common with other animals like frogs and toads. Perhaps the quality that makes humans unique in this sense is the embedding that characterizes our categorical distinctions. The initial phonetic distinctions that were discussed above can probably be best thought of as the internal map that was constructed of the vocal tract. The development of this internal map allowed the human communicative system to evolve from a reactive to cognitive system, thus allowing more extensive planning for speech and at the same time giving us the necessary articulatory components to do it. To reflect back on the proposed role of the mirror system in the evolution of language, consider the importance of the neurological connection between the hand and the mouth. Regarding motor neurons which control actions in both the hand and the mouth, Fogassi & Gallese (2002, 26) note,

> Once neurons of this latter type acquired mirror properties, some of them possibly lost the anatomical connections with the circuit controlling the hand (indeed a feature of motor development is the progressive disappearance of hand and mouth synergism). In the adult these neurons would appear as "mouth" grasping neurons endowed also with the property to respond during the observation of hand actions.

Thus, although the basis of mirror systems may have originated with hand grasping actions, they are transferable to the mouth, and we presume, the vocal tract generally.

Of course, at some point the treelet was exapted from the physical domain into a paradigmatic function, so it is important to recognize that the evolution of the more complex trees involved in speech production aren't directly tied to the cartography of the vocal tract. We will now turn to the same evolutionary process to get us the other half of the articulatory components we need: vowels.

3. Articulation of Vowels

Vowels are of course made with the same articulators as consonants and it's possible to treat vowel articulation using the same primitives as consonants (Jakobson 1962; Catford 1977). We'll try and do this here, to maintain continuity with the rest of the proposal, though much of the literature on phonetics and phonology uses different vocabulary for vowels and consonants.

Starting at the source again, vowels are usually voiced, but they can also be breathy (spread) or creaky (constricted). A more common use of the larynx in vowels is tone and many languages make a three-way contrast among H, M and L tones. Contour ones (rising and falling) exist as well. A particularly rich seven-way contrast is found in various dialects of Hmong, which make meaningful distinctions using three simple tones (high, mid, low), two complex tones (falling and rising), and both breathy and creaky voice (Andruski & Ratliff 2000; Yang 2000):

(23) White Hmong laryngeal contrasts

pɔ́	'lump'	tɔ́	'deep'	(high)
pɔ	'pancreas'	tɔ	'bleed'	(mid)
pɔ̀	'thorn'	tɔ̀	'wait'	(low)
pɔ̂	'female'	tɔ̂	'hill'	(falling)
pɔ̌	'throw'	tɔ̌	'mix'	(rising)
pɔ̰	'see'	tɔ̰	'bite'	(creaky)
pɔ̤	'grandmother	tɔ̤	'sink'	(breathy)

Breathy voice is the vocalic equivalent of aspiration in consonants and creaky voice is the equivalent of glottalization in consonants; tones are not usually phonetically realized during consonants, though some languages allow this with nasals and resonants, for example, Kammu (Svantesson 1983) and Yoruba (Bambgbose 1967).

Moving down the tree, we find that the nasal cavity can be used to make vowel contrasts as well. Thus French contrasts oral and nasal vowels and Palantla Chinantec makes a surprising three-way contrast between oral, slightly nasal and fully nasal nuclei (Merrifield 1963, Ladefoged 1971):

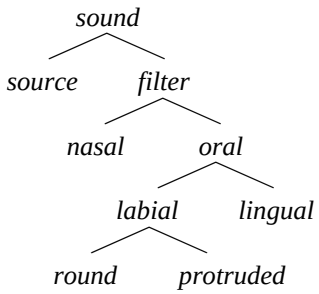
(24) Palantla Chinantec

háa	'so, such'	(oral)
háã	'(he) spreads open'	(slightly nasal)
hg̃ã	'foam, froth'	(fully nasal)

The contrast is reminiscent of the three-way nasal contrasts found in Hmong stops (oral, prenasalized, nasal); this three-way seems to be as far as people can go with nasal contrasts among vowels or consonants.

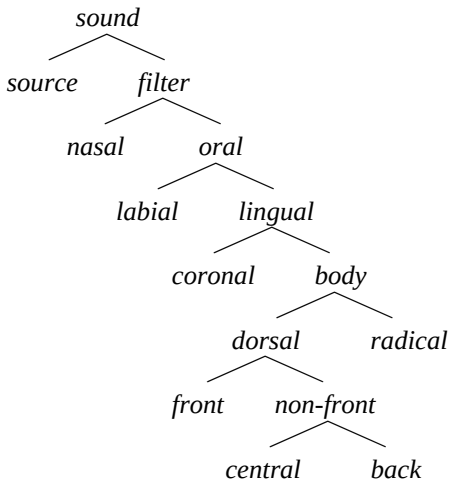
Moving down to the lips, we find that languages contrast unrounded vowels (i, e) and rounded vowels (y, ø). Rounded vowels are usually produced in the back of the mouth (o, u) for acoustic reasons and unrounded vowels are usually produced in the front (i, e), but in principle the rounding of the lips is completely independent of the backness of the tongue during a vowel. In most languages, the rounding of a vowel is a function of how far back the tongue is (back vowels being rounded and front vowels being unrounded), but many languages contrast the two. French and German, for instance, contrast front unrounded i and e with front rounded y and ø. And in Swedish there is a difference between round and protruded, as follows:

(25)



Despite the importance of the lips for producing vowels, the most important vocalic articulation is the tongue. And the most important part of the tongue is the dorsum, whose movement tends to carry the crown and root along with it. One of the fundamental distinctions among vowels has to do with the backness of the tongue dorsum. Many languages distinguish only front from non-front vowels (eg., Turkish) but some distinguish three degrees of backness. White Hmong, for instance, contrasts hi front [i], hi central [ɨ] and hi back [u]. Such languages motivate the following tree:

(26)



The height part of all of this is the vocalic analog of aperture: how far open the mouth is. Minimal aperture is a high vowel, partial aperture a mid vowel, and maximal aperture a low vowel. We have followed common usage and put these aperture distinctions in the tree alongside the front/back dimension, but we could also have pasted them in below the front/central/back branches of the dorsal part of the tree.

We find ourselves now at the tongue root, which can be advanced and retracted but not much else. When it is advanced (ATR) the throat cavity expands; when it is retracted (RTR) the throat cavity contracts. Tongue root distinctions[v] are found in many African languages (Archangeli & Pulleyblank 1994) and seem to be behind the distinction behind long and short vowels in English.

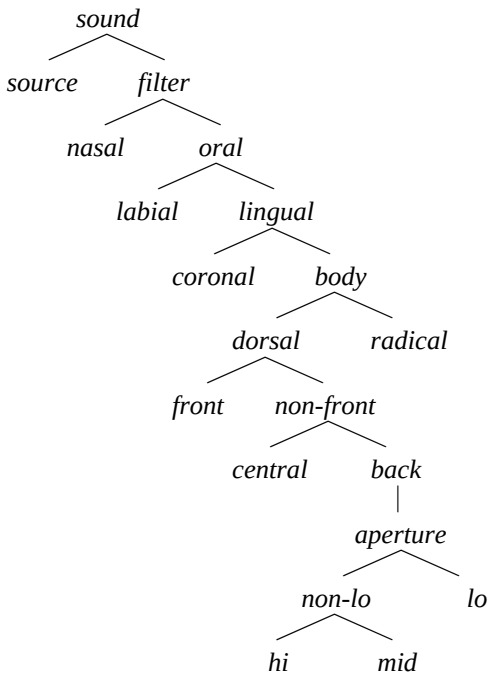
(27) English ‘long’ and ‘short’ vowels

ATR	i ‘beat’
	e ‘bait’
	o ‘boat’
	u ‘boot’
RTR	ɪ ‘bit’
	ɛ ‘bet’
	ɔ ‘bought’
	ʊ ‘put’

The great connect comes next, with the issue of aperture. Again, aperture is relevant mainly to the tongue body. Although some languages use only two degrees of aperture (high-ish and low), most languages use three (hi,

mid, and low). The following tree shows these height differences linked to back (u, o, a), but they can also be linked to front (i, e, æ) and central (i, ɨ, u):

(28)



the great connect

With the ability to round the lips already there, humans would now have been able to produce 17 distinct simple vowels (when two vowels occupy the same cell, the rightmost vowel is rounded).

(29) basic human vowel space

	front	central	back
high	i y	ɨ ʉ	ɯ u
mid	e ø	ə ɵ	ɤ o
Low	æ œ	ɐ	ɑ ɒ

(Lip-rounding is difficult with the mouth wide open, so there aren't as many low rounded vowels as we would expect.)

Crucially, each of these vowels could in principle be produced with any of the 5 tones of Hmong, giving us 17 x 5 = 85 vowel sounds. And any of these 85 sounds could be contrastively breathy or creaky: 85 plain + 85 creaky = 255 vowel sounds. (Hmong doesn't allow breathy and creaky to cross-classify with H, L, M, rising and

falling; but other languages do). And these 255 sounds can be oral or nasal, giving us over 500 possible vowel sounds. More are possible once we consider combinations of vowels (diphthongs and triphthongs), and vowel length (short, long, overlong). In short, there are lots of vowels.

Not all the cells above are equally distinct acoustically. The most distinct vowels form a triangle of sorts: i, a, u. These 'quantal vowels' are located at acoustic hot spots (Stevens 1972) and they are the most common vowels on the planet, followed by e and o. Most languages of the world have a five vowel system, but all systems (aside from the rare cases of two-vowel systems) exploit the basic i/a/u triangle (Lindblom 1986). So from Hawaiian to Spanish to the keyboards we type on we find the five vowels a, e, i, o, u.

We have no idea which of these many oral distinctions arose first, but there are robust asymmetries in the vowel systems of the languages of the world that are highly suggestive (Maddieson 1984; Ladefoged & Maddieson 1996). Every language has oral vowels, but only some have nasal vowels (probably for acoustic reasons, since nasal vowels are harder to distinguish from one another). Almost all languages use height and backness, but distinctive rounding isn't nearly as common and is usually just used to exaggerate the acoustic properties of non-low back vowels. And tongue-root distinctions are the least common of all. So if we assume that the commonness of a given vowel distinction is a function of how easy it is to perceive or produce, we are probably safe in assuming that the most salient distinctions are the oldest ones. We would speculate that the first vowels people produced well were i, a, u; that e and o were added soon after that; and that things like creaky and breathy, nasal, ATR/RTR and so on were developed some time after that. This is roughly how children acquire vowels, too.

4. Conclusion

We began our discussion with the simplest communication device, a vibrating larynx, essentially the same type of communication device found in anurans and birds. For early hominids the larynx emptied directly into the nasal cavity, excluding the mouth from speech production unless the larynx was yanked down (as dogs do when they bark). But once the larynx was permanently lowered enough to let egressive lung air into the mouth all sorts of possibilities opened up and the embedded treelet expanded into a bush full of branching structures, each capable of signaling a change in meaning (*pad, bad, mad; bid, bed, bad, bud*; etc.). This is distinctly human and we're off on the evolution of language, having left our primate relatives in the dust.

The mapping of system bodies also seems to be highly relevant to language learning by artificial devices. For instance, Bailly et al. (1997) conducted an experiment whereby an artificial device gradually learned to produce acoustic signals that matched articulatory gestures, and which coordinated these gestures to produce strings of sounds that could be meaningful. They characterize the first stage in the learning process as "a babbling phase, where the device builds up a model of the forward kinematics, i.e. the articulatory-to-audio-visual mapping" (Bailly et al. 1997, 593). The findings of Bailly et al. highly suggest that such a physiological-acoustics mapping would have been likely to have occurred in non-human primates (as their implications are more directly for child acquisition).

We have seen that a number of important phonological distinctions are readily described in terms of nested categories, often involving a basic distinction between two categories, one of which is further divided into two categories. We assume that this partitioning of the vocal tract and of the distinctions that can be made with it was originally based on physiology and acoustics and had nothing to do with grammar or language *per se*. However it came about, there is ample evidence for a basic embedded treelet (quite unadorned) in phonological representation below the level of the syllable. That the treelet evolved out of something non-linguistic (an internal map of the vocal tract) provides an external footing for the initial adaptation of the structure.

Again, this works well with the assumption that there is a special internal mapping that characterizes the evolution of a reactive system into a cognitive system and that such a mapping was facilitated by a 'mirror system'. What we are not suggesting is that most life forms (including pre-linguistic 'humans') are simply reactive systems and that speech marked the conversion into a true cognitive system; rather, what we are claiming is that the *communicative* system itself can be modeled in this way. There is no doubt that pre-linguistic individuals from the genus *Homo* had a wide range of very specialized cognitive abilities; it is only the module of speech and communication that we think entered into such a late transition.

References

Andruski, J.E. & M. Ratliff. (2000). Phonation types in production of phonological tone: the case of Green Mong. *Journal of the International Phonetic Association, 30*, 63-82.

Archangeli, D. & D. Pulleyblank. (1994). *Grounded phonology*. Cambridge (MA): MIT Press.

Bailly, G., Laboissière, R., & A. Galván. (1997). Learning to speak: speech production and sensori-motor representations. In P. Morasso & V. Sanguineti (eds.), *Self-organization, computational maps, and motor control* (pp. 593-615). Amsterdam: Elsevier Science.

Bamgbose, A. (1967). *A short Yoruba grammar*. Lagos: Heinemann Educational Books.

Brown, J.C. & C. Golston. (2002). Generalized x-bar theory and the evolution of grammar. Paper presented at the Fourth International Evolution of Language Conference, Harvard University.

Carstairs-McCarthy, A. (1999). *The origins of complex language: an inquiry into the evolutionary beginnings of sentences, syllables, and truth*. New York: Oxford.

Catford, J.C. (1964). Phonation types: the classification of some laryngeal components of speech production. In D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott & J.L.M. Trim (eds.), *In honor of Daniel Jones* (pp. 26-37). London: Longmans.

Catford, J.C. (1977). *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press.

Chomsky, N. (1959). Review of B.F. Skinner: Verbal Behavior. *Language, 35*, 26-58.

Cruse, H. (2003). The evolution of cognition: a hypothesis. *Cognitive Science, 27*, 135-155.

Darwin, C. R. (1859). *The origin of species*. London: John Murray.

Esling, J.H. (2002). The laryngeal sphincter as an articulator: how register and phonation interact with vowel quality and tone. Paper presented at the Western Conference on Linguistics, University of British Columbia.

Fisher-Jørgensen, E. (1985). Some basic vowel features, their articulatory correlates, and their explanatory power in phonology. In V.A. Fromkin (Ed.), *Essays in honor of Peter Ladefoged*, (pp. 79-100) Orlando, Florida: Academic Press.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). 1996. Action recognition in the premotor cortex. *Brain,*

199, 593-609.

Gildersleeve-Neumann, C.E., Davis, B.L. & P.F. MacNeilage. (2000). Contingencies governing the production of fricatives, affricates, and liquids in babbling. *Applied Psycholinguistics, 21*, 341-363.

Hauser, M. (1997). *The evolution of communication*. Cambridge: MIT Press.

Herbert, R.K. (1990). The relative markedness of click sounds: Evidence from language change, acquisition, and avoidance. *Anthropological Linguistics, 32*, 120-138.

Iverson, G.K. & J.C. Salmons. (1995). Aspiration and laryngeal representation in Germanic. *Phonology, 12*, 369-96.

Jakobson, R. (1962). *Selected writings*. The Hague: Mouton.

Kehrein, W. (2002). *Phonological representation and phonetic phasing: affricates and laryngeals*. Max Niemeyer Verlag.

Kirk, P., Ladefoged, J. & P. Ladefoged. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In A. Mattina & T. Montler (Eds.), *American Indian linguistics and ethnography in honor of Laurence C. Thompson*, (pp. 435-450). Missoula: University of Montana.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

Ladefoged, P. & I. Maddieson. (1996). *Sounds of the world's languages*. Oxford: Blackwell.

Ladefoged, P. & A. Trail. (1984). Linguistic phonetic description of clicks. *Language, 60*, 1-20.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Laver, J. (1994). *Principles of phonetics*. Cambridge: Cambridge University Press.

Lieberman, P. (1984). *The biology and evolution of language*. Cambridge (MA): Harvard University Press.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology* (pp. 13-44). Orlando: Academic Press.

Locke, J.L. (1983). *Phonological acquisition and change*. New York: Academic Press.

Lombardi, L. (1995). Laryngeal neutralization and syllable wellformedness. *Natural Language and Linguistic Theory, 13*, 39-74.

MaKay, I.R.A. (1977). Tenseness in vowels: an ultrasonic study. *Phonetica, 34*, 325-315.

MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences, 21*, 499-511.

MacNeilage, P.F. & B.L. Davis. (1990). Acquisition of speech production: frames, then content. In M. Jeannerod (Ed.), *Attention and performance XIII* (pp. 453-475). Hillside, NJ: Lawrence Erlbaum Associates.

MacNeilage, P.F. & B.L. Davis. (1993). Motor explanations of babbling and early speech patterns. In B. De Boysson-Bardies, S. De Schonen, P. Jusczyk, P. MacNeilage & J. Morton (Eds.), *Developmental neurocognition: speech and face processing in the first year of life* (pp. 341-352). Dordrecht: Kluwer.

MacNeilage, P.F. & B.L. Davis. (1999). Evolution of the form of spoken words. *Evolution of Communication, 3,* 3-20.

MacNeilage P.F. & B.L. Davis. (2000). Evolution of speech: The relation between ontogeny and phylogeny. In C. Knight, M. Studdert-Kennedy, & J.R. Hurford, (Eds.), *The evolutionary emergence of language: social function and the origins of linguistic form,* (pp. 146-160). Cambridge: Cambridge University Press.

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

Maddieson, I. (1987a). Revision of the IPA: lingua-labials as a test case. *Journal of the International Phonetic Association, 17,* 26-30.

Maddieson, I. (1987b). Lingua-labials. Paper presented the 113th Meeting of the Acoustical Society of America. Indianpolis, May 1987. Abstract in *Journal of the Acoustical Society of America* 81/S1, S65.

Merrifield, W.R. (1963). Palantla Chinantec syllable types. *Anthropological Linguistics, 5*, 1-16.

Nowicki, S., Westneat, M. & W. Hoese. (1992). Birdsong: motor function and the evolution of communication. *Seminars in Neuroscience, 4,* 385-390.

Rand, A.S. (1988). An overview of anuran acoustic communication. In B. Fritzsch, M.J. Ryan, W. Wilczynski, T.E. Hetherington, & W. Walkowiak (Eds.), *The evolution of the amphibian auditory system* (pp. 415-432). New York: John Wiley & Sons.

Rizzolatti, G. & M.A. Arbib. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188-194.

Rizzolatti, G., Craighero, G., & L. Fadiga. (2002). The mirror system in humans. In Stamenov & Gallese (2002), (pp. 37-59).

Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3,* 131-141.

Saussure, F. de. (1916/1960). *Cours de linguistique générale* [Course in General Linguistics]. Ed. C. Bally, A. Sechehaye & A. Riedlinger, translated by W. Baskin. New York: McGraw-Hill.

Stamenov, M.I. & V. Gallese (eds.). (2002). *Mirror neurons and the evolution of brain and language*. Amsterdam: John Benjamins Publishing Company.

Steriade, D. (1994). Complex onsets as single segments: the Mazateco pattern. In J. Cole & C. Kisseberth (Eds.), *Perspectives in phonology*, (pp. 203-91). Stanford, CA: Center for the Study of Language and Information.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. I .B. Denes & E.E. David Jr. (Eds.), *Human communication: A unified view*, (pp. 51-66) New York: McGraw-Hill.

Suthers, R.A. & D.H. Hector. (1988). Individual variation in vocal tract resonance may assist oilbirds in recognizing echoes of their own sonar clicks. In P.E. Noichtigall & P.W.B. Moore (Eds.), *Animal sonar: processes and performances* (pp. 87-91). New York: Plenum Press.

Svantesson, J.-O. (1983). *Kammu phonology and morphology*. Lund, Sweden: Liber Forlug.

Westneat, M., Long, H., Hoese, W. & S. Nowicki. (1993). Kinematics of birdsong: functional correlation of cranial movements and acoustic features in sparrows. *Journal of Experimental Biology, 182*, 147-171.

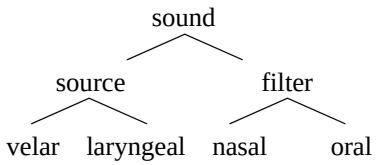
Yang, P. (2000). A reanalysis of tones in White Hmong. Paper presented at the *Western Conference on Linguistics*, California State University, Fresno.

Notes

[*] The authors wish to thank Brian Agbayani, Andrew Carstairs-McCarthy, Will Lewis, and Douglas Pulleyblank for discussion and comments on the present work. Special thanks to Will Lewis for generating the text count discussed in section 1. All errors, however, remain the author's.

[i] For evidence that humans are endowed with a mirror system, see Rizzolatti et al. (2002), who also speculate on new functions that the system may have acquired in humans.

[ii] As mentioned earlier, the dichotomies in phonetic and phonological distinctions we are proposing are much more common than ternary distinctions with no sub-grouping, or quaternary distinctions with elaborations on both sides of the initial split. However, these types of categorizations do indeed arise. For example, the category of 'sound' seems to break down into both 'source' and 'filter', and each of these can break down into 'velar'/'laryngeal' and 'nasal'/'oral' respectively.



[iii] Catford (1964) has mentioned problematic phonation types such as 'whispery voiced creak' which makes use of both spreading and constriction of the glottis, as well as 'ligamental voice' whereby the "arytenoid cartilages are tightly occluded" (pg. 32). 'Whispery voiced creak' stands as a combination of gestures not allowable under this model, and 'ligamental voice' would not be adequately considered as a distinctive gesture or setting. As Laver (1980:139) points out, "There are physiologically possible phonation types quite outside the descriptive system presented here, omitted because they seem never to be used in normal speech, whose possibilities of occurrence in compound phonations are not yet well analysed." We are left to agree with Laver on these points.

[iv] The tongue can also contact the top lip, making for a linguo-labial sound, found in a very small number of languages (Maddieson 1987ab).

[v] 'Tense/lax' distinctions also fall in this category. Although tense/lax is not straightforwardly correlated with ATR/RTR (for example in Yi, where tense correlates more closely with RTR and lax with ATR (Esling 2002; see MacKay 1977 for general discussion), either way the distinction is made is consistent with our model.