frontiers in PSYCHOLOGY

IMPACT 2.8 FACTOR

Language Sciences

Attention mechanisms and the mosaic evolution of speech

Pedro Tiago Martins and Cedric Boeckx

Journal Name:	Frontiers in Psychology
ISSN:	1664-1078
Article type:	Perspective Article
First received on:	28 Jul 2014
Revised on:	23 Nov 2014
Frontiers website link:	www.frontiersin.org

Attention mechanisms and the mosaic evolution of speech

Pedro Tiago Martins[1,2,5]* and Cedric Boeckx[3,4,5]

[1]Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain
[2]Center of Linguistics of the University of Porto, Porto, Portugal
[3]Department of General Linguistics, University of Barcelona, Barcelona, Spain
[4]ICREA (Catalan Institute for Research and Advanced Studies), Barcelona, Spain
[5]Biolinguistics Initiative Barcelona, Barcelona, Spain

***Correspondence:**
Pedro Tiago Martins
Biolinguistics Initiative Barcelona
Universitat de Barcelona
Department of General Linguistics
Gran Via de les Corts Catalanes 585
08007, Barcelona, Spain
pedrotiago.martins01@estudiant.upf.edu;

Word-count (sans abstract): 2540

Abstract: There is still no categorical answer as to why humans, and no other species, have speech, or why speech is the way it is. Several purely anatomical arguments have been put forward, but they have been shown to be false, biologically implausible, or of limited scope. This perspective paper supports the idea that evolutionary theories of speech could benefit from a focus on the cognitive mechanisms that make speech possible, for which antecedents in evolutionary history and brain correlates can be found. This type of approach is part of a very recent but rapidly growing trend that has already provided crucial insights on the nature of human speech by focusing on the biological bases of vocal learning. Here we contend that a general mechanism of attention, which manifests itself not only in the visual but also in the auditory modality, might be one of the key ingredients of human speech, in addition to the mechanisms underlying vocal learning, and the pairing of facial gestures with vocalic units.

Keywords: evolution of speech, attention mechanisms, consonants and vowels, evolution, oscillatory cycles.

The mechanics of speech have been thoroughly studied. Various techniques and methodologies have been developed that allow us to know with great precision what goes on anatomically when human and non-human primates vocalize, from the lungs to the lips (Harcastle et al. 1989; Fitch and Hauser 1995; Fishman 2003; Ghazanfar and Rendall 2008). However, the question of *why* (only) humans have speech in the first place remains to be categorically answered. Different purely anatomical arguments have been put forward, such as the uniqueness of the descended human larynx (Fant 1960; Lieberman and Crelin 1971) or the loss of air sacs in humans (de Boer 2012), but both arguments have been seriously questioned (Fitch and Reby 2001; Nishimura et al. 2006; Littauer 2012). In fact arguments of this type all share a general problem: they fail to grasp the mosaic nature of cognitive faculties that evolution has tinkered with. Modern evolutionary biology shows that complex traits—and surely speech or indeed language as a whole falls within that category—require complex and multi-dimensional explanations (West-Eberhardt 2003; Pigliucci & Müller 2010).

In addition, there are cases of other, non-human, even non vocal-learning species that are capable of producing human-like vowels (Vs) and consonants (Cs), such as the Gelada Baboons (*Theropithecus gelada*), which seem to possess an extremely rich sound repertoire, comparable to that of humans. More specifically, it has been shown that this species is able to produce vocalizations that not only employ what we would perceive as consonants and vowels, but also are structured in a way that resembles human sound systems, with different vowel qualities and consonants distinguished by manner and place of articulation, as well as duration similar to that of human speech (Richman 1976, et seq; Bergman 2013). There are, of course, different ways of articulating sounds with the same acoustic effect, even among humans, but the very fact that there are indeed other species that are able to produce consonants and vowels in a dynamic manner and yet lack human-like speech, shows that merely having that inventory is not a diagnosis for neither speech nor language. So why is it, then, that we humans have it and species like Gelada Baboons don't? We agree with a growing trend in the study of human speech (Deacon 1997; MacNeilage 2008; Fitch 2010) that the answer surely has to do with the presence of vocal learning mechanisms in humans, for which a biological basis is emerging (a robust direct laryngeal connection from the motor cortex seems to be key; Fitch 2010). But we contend that answers currently entertained in the literature are insufficient to account for a foundational property of speech: its Consonant-Vowel-based *organization*. Other species, including vocal learning ones, do not organize their vocal behavior the way we do.

It is generally agreed upon that vowel and consonant sounds definitely exist. There are mechanical reasons for this (see e.g. Fant (1960), MacNeilage (1998) among others). In linguistics , especially since the work of Chomsky and Halle (1968), it has been generally assumed that speech-sounds are abstractly represented as bundles of features, which must be somehow encoded in the brain (see Bouchard et al. (2013) and Mesgarani et al. (2014) for recent brain work regarding the latter point. Whether features are innately specified or not is an issue that does not bear on what follows, and on which we take no stand; see Clements and Ridouane (2011) for discussion of different perspectives). Whatever basic units of phonological analysis one chooses,

they generally boil down to articulatory characteristics (e.g. “bilabial” or “voiced”). However, since just being able to perceive or produce vowels and consonants is not a diagnostic for speech, as the vocal behavior of the Geladas illustrates, the productive use of Cs and Vs in something like speech (and language) must be explained not only by their encoding in the brain, but also, and crucially, by their distinct *functional/cognitive values*.

It has been shown that vowels and consonants are not treated on a par by human brains. When presented with speech, humans not only pick out the segments that comprise the sound continuum, but also ascribe different functional/cognitive weight to different kinds of segments. Vowels and consonants indeed have different roles, with consonants providing lexical cues and vowels providing cues about syntactic structures (Nespor et al 2003). Following Toro et al. (2008), we will refer to this as the CV hypothesis. A common initial objection to the CV hypothesis as stated is that the reason why humans process consonants and vowels in this differentiated manner is their asymmetric statistical distribution. Languages usually have more consonants than vowels, and they are distributed in such a way as to facilitate the extraction of transitional probabilities of consonants and their lexical information, with the subtler alternations of vowels providing the cues for structural information. If this were true, humans would be able to extract lexical and structural information based on the statistical distribution of segments alone, regardless of their being vowels or consonants. Building on previous work (e.g. Bonatti et al. 2005), Toro et al. (2008) tested just that, and found that the bias is deeper than “mere” statistics: they inverted the roles of consonants and vowels in the data and presented it to several subjects, who were simply unable to extract the same rules from the signal. Another objection to the CV hypotheses would be that the acoustic differences between consonants and vowels are responsible for their differentiated processing. But if physical aspects of speech sounds were the sole responsible for rule extraction one should not expect variation in their functional roles based on whether the same sounds are interpreted as speech sounds or as noise. However, research points the other way. For example, language-related areas are modulated differently by identical sounds depending on whether they are perceived as speech or non-speech (Möttönen et al. 2006), and audio-visual speech perception is triggered by acoustic stimuli perceived as speech, and not triggered when the exact same stimuli are perceived as something else (Toumainen et al. 2005).

These results show that the claim that language acquisition is made possible by general-purpose learning mechanisms (e. g. Elman et al. 1996) must be qualified: surely, if this were the case, humans would have no problem extracting different kinds of information from any sound system with asymmetrically distributed segments. It seems instead that there is a more basic biological bias for extracting lexical information for consonants and structural information from vowels, by virtue of their functional—and not statistical—differences.

More generally, it is not the case that the human brain processes different information equally. Not unlike speech, vision is a good example of selective processing of noisy input. In a recent study (Fiebelkorn et al 2013), researchers draw a very important

connection between endogenous oscillatory rhythms and space-based and object-based selection mechanisms. They suggest that the problem of retrieving the right information despite the abundance of signal is achieved through "rhythmic patterns of visual-target detection both within (8Hz) and between (4Hz) objects" (p. 2553). Compatible results are reported by Landau and Fries (2012).

These frequencies fall right within the range reported in Giraud and Poeppel (2012): the articulatory and the auditory systems structure their outputs in agreement with one another, that is, they are mediated by something which allows them to be in sync. This infrastructure provided by neuronal oscillations might be the key in explaining how the brain decodes continuous speech. Crucially, there is a robust relation between the time scales associated with speech cues (phonemes, syllables, and intonational phrases) and the time constants underlying neuronal oscillations (low-gamma, theta and delta oscillations). These same oscillatory cycles have been linked to various "putative precursors" of speech, such as monkey lip-smacking (Ghazanfar et al. 2012; Fitch 2013). We contend that in fact these entrainment patterns, despite being manifested upon contact with different kind of stimuli, point to one very general attention mechanism, which has been put to new use in humans, and which has given speech one of its most distinctive signatures. Specifically, the fact that two different patterns (~4Hz and ~8Hz) are associated, respectively, with within- and between-object attention (Fiebelkorn et al 2013), plausibly reveals that Cs and Vs are specifically targeted by these different frequencies, or at least by a low-frequency/high-frequency dichotomy within the ranges reported and reviewed by the studies cited above. This would help explain why Cs are associated with lexical properties (*between*-word) and Vs with syntactic/structural properties (*within*-word) in the sound continuum. This would represent a very important ingredient of human speech, absent in other species, including other vocal-learners. A central question for our proposal is whether this attention mechanism is confined to a single domain or, instead, much more general.

Close relationships have been drawn between the underlying mechanisms behind visual and auditory attention, which all revolve around the recognition, selection and processing of information in space and/or time. De Freitas et al. (2013) ran attention experiments on which they tested the so-called *same-object advantage* (when the same physical distance is considered, responses are faster when probes occur within the same object than when in other objects) in the sound domain. Indeed, they show that responses are also faster within the same rhythmic phrase (a tone of a single frequency) than across different rhythmic phrases, with duration being the analogue of distance. These results strongly suggest that human object-based attention is not exclusive to vision, and most likely not fundamentally spatial, but rather shared across domains.

There is remarkable coherence between the acoustic and visual cues of speech, such as temporal correspondence between mouth opening and acoustic envelope, area of mouth opening and formants, and temporal modulation of mouth movements and voice envelope (2–7Hz) (Chandrasekaran et al. 2009). A popular topic that comes up when discussing the relation between visual and auditory speech cues is the McGurk

effect (McGurk & MacDonald 1976), but here we are referring to something different: while the McGurk effect refers to the interference of (discrepant) visual cues in acoustic perception (see Tiippana (2014) for a clarification of some misconceptions in this regard), we instead refer to the shared history and interdependence of auditory and speech cues at the neural level. Indeed, speech rhythm and facial expressions in humans are both rhythmic (3-8Hz) and very much correlated (Ghazanfar and Takahashi 2014a, 2014b; Golumbic et al. 2013). Such a correlation has moreover been deemed crucial for the social interaction required for speech to prosper, that is, the coordination of individuals of a group through the syncing of neural processes across brains, in what has been called "brain-to-brain coupling" (Hasson et al. 2012). As these authors argue, if cognitive processes underlying complex behavior depended solely on the processing within the individual's brain, it would hard—if not impossible—to reach a set or rules for interactive behavior to follow and sync. By sending cyclic, brain-generated signals through the physical environment to another brain, which decodes and accommodates them, brains really do sync through oscillatory activity.

On the basis of the findings we have pointed out so far, it is plausible that this mechanism is indeed general, thus representing a good example of an already existing capacity put to new use. Presumably, the recruitment of this domain-general attention mechanism in the domain of speech was the solution to the externalization of the complex syntactic/semantic component that other species lack (Berwick et al. 2011).

A prediction of our hypothesis is that non-human animals—crucially those which have been shown to distinguish between vowels and consonants, which might or might not be able to produce them—, will display no functional difference between these two kinds of segments, and whatever rules they extract from auditory input will therefore not depend on the cues being vowels or consonants. This is also a prediction of de la Mora & Toro (2013), who performed experiments on Long-Evans rats (*Rattus norvegicus*) and concluded that they actually surpass humans in rule extraction from auditory input tasks: whereas rats had no problem generalizing rules in CVCVCV words both over vowels and consonants, humans could only do it for the vowels, with the same stimuli. Further rule-extracting experiments with non-human animals will surely strengthen the import of these results.

Though related, de la Mora & Toro's (2013) prediction and ours, however, are not equivalent. For them, whatever makes us worse than mice at rule-extraction from auditory data and thus better than them at inferring lexical and structural information must be unique to language, and to humans (by definition, this would fall under Hauser et al. (2002) 'Faculty of Language in the Narrow Sense'). They leave the exact nature of this constraint up for grabs, but they assume that only "if the observed differences in how humans process speech are a result of language-specific constraints, we should not observe functional differences in other species." (p. 308) Our prediction makes no claim of uniqueness to language; we contend that the functional difference humans attribute to consonants and vowels is due to non-linguistic aspects of our neurology, namely a general mechanism of attention, and which along with vocal learning and the ability to produce a large enough sound

inventory formed the basis of human speech. We agree with Gervain & Mehler (2010: 196), that "[i]f humans and non-human animals share cognitive and/or learning abilities, these cannot be language specific since only our species have language. However, they may have been precursors bringing humans closer to language."

Our perspective may also benefit from a closer examination of the relation between the attention mechanism appealed to here and the general issue of working memory. The literature on working memory often relates to the one on attention. Furthermore, it is known that the storage of vowels and consonants differ, both in terms of stability (storing of vowels being more stable than that of consonants), and in terms of sensitivity to order information (vowels being more related to order information in the phonological sequence than consonants; Drewnowski 1980, Baddeley 2007). This seems to converge with the proposal of Nespor, Peña and Mehler (2003), and with our emphasis on consonants and vowels having distinct cognitive imports.

A rapprochement between our proposal and working memory could shed light on the neuroanatomical basis of the C/V distinction, given recent progress in the characterization of human-specific connectivity patterns (see Aboitiz 2012, Scott et al. 2012, Neubert et 2014, among others). We leave a detailed exploration of this issue for future research.

What is clear to us already is that mechanisms of the sort we have appealed to here are very much line with what de Waal & Ferrari (2010) call the bottom-up perspective on human and animal cognition: looking for wide-ranging, basic mechanisms across species and domains, instead of asking what is special and unique about any one trait and species.

As put by Fitch & Zuberbühler (2013: 27), "Although language, as a composite system, is clearly unique to our species, substantial empirical work is still required before any of the mechanisms involved in language can be conclusively labeled unique". We think this true also of speech, and believe that the most fruitful way of unveiling its nature is to study the structure and evolution of each of the mechanisms involved, such as the one we put forth.

Conflict of Interest Statement
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Authors and Contributors
Both authors contributed to the writing of this paper.

Acknowledgements: The present work was made possible through a Marie Curie International Reintegration Grant from the European Union (PIRG-GA-2009-256413), research funds from the Fundació Bosch i Gimpera, and a grant from the Spanish Ministry of Economy and Competitiveness (FFI-2010-20634).

References

Aboitiz F (2012) Gestures, vocalizations, and memory in language origins. Front. Evol. Neurosci. 4:2. doi: 10.3389/fnevo.2012.00002

Baddeley AD (2007) Working memory, thought and action. Oxford University Press, Oxford.

Bergman TJ (2013) Speech-like vocalized lip-smacking in geladas. Curr. Biol. 23:R268–R269. doi:10.1016/j.cub.2013.02.038

Berwick RC, Okanoya K, Beckers GJ, Bolhuis JJ (2011) Songs to syntax: the linguistics of birdsong. Trends Cogn. Sci. 15:113–121. doi:10.1016/j.tics.2011.01.002
de Boer B (2012) Loss of air sacs improved hominin speech abilities. J. Hum. Evol. 62:1–6. doi:10.1016/j.jhevol.2011.07.007

Bonatti LL, Peña M, Nespor M, Mehler J (2005) Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychol. Sci.* 16:451–459

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495:327–332. doi:10.1038/nature11911

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput. Biol. 5:e1000436. doi:10.1371/journal.pcbi.1000436

Chomsky N, Halle M (1968) The sound pattern of english. Harper & Row, New York.

Clements GN, Ridouane R (eds.) (2011) Where do phonological features come from? Cognitive, physical and developmental bases of distinctive speech categories. John Benjamins, Amsterdam/Philadelphia

de Freitas J, Liverence BM, Scholl BJ (2013) Attentional rhythm: a temporal analogue of object-based attention? J. Exp. Psychol.-Gen. 143:71–76. doi: 10.1037/a0032296

de la Mora DM, Toro JM (2013) Rule learning over consonants and vowels in a non-human animal. Cognition 126:307–312.

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

Drewnowski A (1980) Attributes and priorites in short-term recall: a new model of memory span. J. Exp. Psychol. 109:208-250

Elman JL, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K (1996) Rethinking innateness: a connectionist perspective on development. MIT Press, Cambridge, MA

Fant G (1960) Acoustic theory of speech production, Number 2. Walter de Gruyter, The Hague

Fiebelkorn IC, Saalmann YB, Kastner S (2013) Rhythmic sampling withim and between objects despite sustained attention at a cued location. Curr. Biol. 23:2553–2558. doi:10.1016/j.cub.2013.10.063

Fishman EK (2003) A brief overview of CT angiography. Appl. Radiol. 32:9–11

Fitch WT (2010) The evolution of language. Cambridge University Press

Fitch WT (2013) Tuned to the rhythm. Nature 494:434–435. doi:10.1038/494434a

Fitch WT, Hauser MD (1995) Vocal production in nonhuman primates: acoustics, physiology and functional constraints on "honest" advertisement. Am. J. Primatol. 37:191–219. doi:10.1002/ajp.1350370303

Fitch WT, Reby D (2001) The descended larynx is not uniquely human. P. Roy. Soci. Lond. B. Bio. 268:1669–1675. doi:10.1098/rspb.2001.1704

Fitch WT, Zuberbühler K (2013) Primate precursors to human language: beyond discontinuity. In Altenmüller E, Schmidt S, Zimmermann E (eds) The evolution of emotional communication: from sounds in nonhuman mammals to speech and music in man. Oxford University Press, Oxford, pp 26–48

Gervain J & Mehler J (2010). Speech perception and language acquisition in the first year of life. Annu. Rev. Psychol. 61:191–218

Ghazanfar AA, Rendall D (2008) Evolution of human vocal production. Curr. Biol. 18:R457. doi:10.1016/j.cub.2008.03.030

Ghazanfar AA, Takahashi DY (2014a) Facial expressions and the evolution of speech rhythm. J. of Cognitive Neurosci. 26:1196–1207. doi:10.1162/jocn_a_00575

Ghazanfar AA, Takahashi DY (2014b) The Evolution of speech vision, rhythm, cooperation. Trends. Cog. Sci. 18:543-553 doi: 10.1016/j.tics.2014.06.

Ghazanfar AA, Takasahi DY, Mathur N, Fitch WT (2012) Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Curr. Biol. 22:1176–1182. doi:10.1016/j.cub.2012.04.055

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging

computational principles and operations. Nature Neurosci. 15:511–517. doi:10.1038/nn.3063

Golumbic EZ, Cogan GB, Schroeder CE, Poeppel D (2013) Visual Input Enhances Speech Envelope Tracking in Auditory Cortex at a 'Cocktail Party'. J. Neurosci. 33: 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

Hardcastle W, Jones W, Knight C, Trudgeon A, Calder G (1989) New developments in electropalatography: A state-of-the-art report. Clin. Linguist. Phon. 3:1–38

Hasson U, Ghazanfar AA, Galantucci B, Garrod S, Keysers C (2012) Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends Cogn. Sci. 16:114–121. doi:10.1016/j.tics.2011.12.007

Hauser MD, Chomsky N, Fitch WT (2002) The Faculty of Language: What is it, Who has it, and How did it Evolve? Science 298:1569–1579.

Landau AN, Fries P (2012) Attention samples stimuli rhythmically. Curr. Biol. 22:1000–1004. doi:10.1016/j.cub.2012.03.054

Lieberman P, Crelin ES (1971) On the speech of the Neanderthal man. Linguist. Inq. 2:203–222

Littauer R (2012) Re-dating rhe loss of laryngeal air sacs in homo sapiens. In: Scott-Phillips TC, Tamariz, M, Cartmill EA, Hurford JR (eds) The evolution of language: proceedings of the 9th international conference (EVOLANG9). World Scientific, Singapore, pp 486–487

MacNeilage PF (1998) The frame/content theory of evolution of speech production. Behav. Brain. Sci. 21:499–546

MacNeilage PF (2008) The origin of speech. Cambridge University Press

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748. doi: 10.1038/264746a0

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in Human Superior Temporal Gyrus. Science 343:1000-1010. doi: 10.1126/science.1245994

Möttönen R, Calvert GA, Jääskeläinen IP, Matthews PM, Thesen T, Tuomainen J, Sams M (2006) Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage 30: 563-569. Doi: 10.1016/j.neuroimage.2005.10.002

Nespor M, Peña M, Mehler J (2003) On the different roles of vowels and consonants

in speech processing and language acquisition. Lingue e Linguaggio ii:207-227.

Neubert F, Mars RB, Thomas AG, Sallet J, Rushworth MFS (2014) Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. Neuron 8:700-713. doi: 10.1016/j.neuron.2013.11.012

Nishimura T, Mikami A, Suzuki J, Matsuzawa T (2006) Descent of the hyoid bone in chimpanzees: evolution of face flattening and speech. J. Hum. Evol. 51:244–254 Pigliucci M, Müller G (eds.) (2010). Evolution – The Extended Synthesis, Cambridge MA, MIT Press.

Richman B (1976) Some vocal distinctive features used by gelada monkeys. J. Acoust. Soc. Am. 60:718–724

Scott BH, Mishkin M, Yina P (2012) Monkeys have a limited form of short-term memory in audtion. Proc. Natl. Acad. Sci. 109:12237-12241. doi: 10.1073/pnas.1209685109

Tiippana K (2014) What is the McGurk effect? Front. Psychol 5:725. doi: 10.3389/fpsyg.2014.00725

Toumainen J, Andersen T, Tiippana K, Sams M (2005) Audio-visual speech perception is special. Cognition 96:B13-B22. doi:10.1016/j.cognition.2004.10.004

Toro JM, Nespor M, Mehler J, Bonatti L (2008) Finding words and rules in a speech stream: functional differences between vowels and consonants. Physch. Sci. 9:13–7144. doi:10.1111/j.1467-9280.2008.02059.x

de Waal F, Ferrari PF (2010) Towards a bottom-up perspective on animal and human cognition. Trends Cogn. Sci. 14:201-207. doi:10.1016/j.tics.2010.03.003

West-Eberhard MJ (2003) Developmental Plasticity and Evolution. Oxford, Oxford University Press