Besides sophisticated phono-articulatory abilities, the architecture of speech has key computational, neuronal, and social prerequisites that can shed light on its phylogenetic and ontogenetic origins.

As a first important requirement, the architecture of speech has to be configured for vocal learning, with adaptable sensorimotor circuits that couple heard speech sounds with motor programs for speech production. From a computational perspective, mastering speech in naturalistic environments plagued by uncertainty and noise is hard; this fact has long motivated control-theoretic views of speech emphasizing error-correction mechanisms and internal modeling (Guenther & Perkell 2004; Moore 2007).

Computational considerations also suggest that speech processing (and learning, see below) might benefit from a close interaction of perception and production systems. For example, production systems might support perceptual processes by predicting and "synthesizing" auditory candidates (as in *analysis by synthesis*), while perceptual systems might support the self-monitoring and error-correction of vocal production by affording an advance auditory analysis of the produced speech sounds. Neurobiological experiments support this idea by showing that the neuronal mechanisms for speech production and perception are not segregated in the brain; for example, specific motor circuits are recruited for the analysis of speech sound features (D'Ausilio et al. 2012). An organic proposal on the architecture of speech can be formulated within the framework of *generative systems*, in which perception and action systems share computational (and neuronal) resources and are both guided by a common prediction-error minimization process (Dindo et al. 2011; Friston 2010; Kiebel et al. 2008; Pezzulo 2012a; 2013; Yildiz et al. 2013).

A second important requirement is a learning method powerful enough to train the aforementioned sensorimotor architecture to perceive and (re)produce sounds and speech. This problem has been studied particularly in songbirds that, while not speaking, have sophisticated vocal learning abilities. Most theories assume that songbird learning is a staged process (Brainard & Doupe 2002). An initial period of auditory learning is needed to tune sensory maps to represent sensory "prototypes" of heard speech sounds (e.g., memorize learned song patterns heard by conspecifics). These prototypes are then used as "reference signals" for imitation learning; by learning to reproduce the stored template, an animal can acquire equivalent vocal sound production skills. In control-theoretic terms, this process uses (auditory and articulatory) feedback error-correction mechanisms to produce a sound (sing or speech) that closely matches the stored template (Guenther & Perkell 2004). During the learning process, internal (inverse and forward) models are trained, too, that successively afford skilled sing or speech processing.

To speed up learning, learners benefit from using self-imitation, too. Covert rather than overt singing (or speaking) might reproduce frequently heard speech sounds in the same way they are encoded in their sensory maps (note that generative architectures afford this form of learning quite naturally; Hinton 2007). Using both overt and covert processes, animals (including humans) might reproduce their stored prototypes with high fidelity, including the local *accents* of their communities.

The brain architecture supporting the aforementioned learning processes is incompletely known. Indeed, speech is a computationally challenging skill as it requires sensorimotor circuits to be sensitive enough to discriminate subtle changes in speech sounds, and accurate enough to afford extremely precise control (e.g., of the timing of speech). The brain could finesse these problems by recruiting cortico-subcortical loops (especially those involving the basal ganglia and the cerebellum) especially during learning. The role of these loops is seldom recognized in "cortico-centric" theories of motor skills (including speech), but the evidence indicates that they could play an important role in skill learning and mastery (Ackermann 2008; Caligiore et al. 2013). For example, vocal learning in the swamp sparrow might involve a loop between forebrain neurons that establish auditory-vocal correspondences and striatal structures important for song learning (Prather et al. 2008).

The high-fidelity reproduction of sounds could be key to cultural transmission and the evolutionary value of singing in songbirds (Merker 2012). However, human communities have richer social structures than other animals, which might have favored an open-ended instrumental use of vocal production besides ritualized display. The importance of this skill might have led to a greater investment of parental time in teaching and, we propose, to advanced forms of "tutor learning" (Canevari et al. 2013). Of note, a so-called pedagogical learning environment (Csibra & Gergely 2011) might have afforded specialized teaching strategies that could be uniquely human and that greatly improve on imitation and self-teaching learning methods. One example is "motherese": Mothers modify their speech when speaking to young children in order to simplify their auditory processing and learning (see Pezzulo et al. 2013). This example suggests that social and interactive aspects of the learning environment are important prerequisites – or at least a useful scaffold – for speech acquisition and cultural transmission.

In sum, speech processing requires a sophisticated neuro-computational architecture in which physiologic, motoric, sensory, and social aspects mutually constrain each other and plausibly co-evolve. In addition to studying genetic determinants, it is important to recognize that speech could have found a suitable "neuronal niche" (Dehaene & Cohen 2007) in existing brain structures (cortical and subcortical) supporting skilled action. For example, speech could have re-used "generative" dynamics of such structures for imitation and self-imitation, and re-deployed existing computational resources for combinatorial processing (Chersi et al. 2014; Fadiga et al. 2009).

In parallel, speech could have found a suitable "socio-cultural niche": It could have been incubated within the sophisticated interactive and social dynamics of our species. The social context in which human speech is acquired is extremely rich, and human speech learning operates on top of the sophisticated interactive, joint action, mutual emulation, and pedagogical abilities, most of which are unique or at least much more developed in our species (Pickering & Garrod 2013; Sebanz et al. 2006). The demands of sophisticated social interactions might have contributed to transform vocalization from an initially quite limited sensorimotor feat to a powerful, open-ended instrumental tool that permits conveying rich communicative intentions and forming extremely varied cultures (Pezzulo 2012b). In turn, we should not neglect how the intertwined sensorimotor and social sides of speech had a transformative impact on the destiny of our species.

Vocal learning, prosody, and basal ganglia: Don't underestimate their complexity[1]

doi:10.1017/S0140525X13004184

Andrea Ravignani,[a] Mauricio Martins,[a,b] and W. Tecumseh Fitch[a]

[a]*Department of Cognitive Biology, University of Vienna, A-1090 Vienna, Austria;* [b]*Language Research Laboratory, Lisbon Faculty of Medicine, 1649-028 Lisbon, Portugal.*

andrea.ravignani@univie.ac.at mauricio.martins@univie.ac.at
tecumseh.fitch@univie.ac.at
http://homepage.univie.ac.at/andrea.ravignani/
www.researchgate.net/profile/Mauricio_Martins4/
http://homepage.univie.ac.at/tecumseh.fitch/

Abstract: Ackermann et al.'s arguments in the target article need sharpening and rethinking at both mechanistic and evolutionary levels. First, the authors' evolutionary arguments are inconsistent with recent evidence concerning nonhuman animal rhythmic abilities. Second, prosodic intonation conveys much more complex linguistic information

than mere emotional expression. Finally, human adults' basal ganglia have a considerably wider role in speech modulation than Ackermann et al. surmise.

While Ackermann et al.'s theory is interesting, seems plausible, and may initially appear tempting, it is based on incomplete readings of several literatures. First, it is unclear why some of their arguments should only apply to the specific instances of rhythmic and prosodic control the authors discuss or why they *fail* to apply in other animal species. Their model assumes that enhancement of in-group cooperation and cohesion was the main driving force for the evolution of speech via the intermediate step where vocal control and rhythm production would serve as chorusing and bonding tools. A key assumption is that speech would produce rhythmic abilities as an evolutionary by-product. This scenario is in line with some empirical observations (for reviews, see Fitch 2012; Geissmann 2000) and previous theoretical frameworks for the *origins of music* (Hagen & Bryant 2003; Hagen & Hammerstein 2009; Merker 2000; Merker et al. 2009). However, when applied to *language*, Ackermann et al.'s evolutionary model does not withstand cross-species validation: Many nonhuman animals exhibit rhythmic behaviors while lacking speech. Before primate rhythmic abilities can be compared with humans' at all, more evidence regarding flexibility in vocalizations' temporal patterning (Fedurek et al. 2013) and motor synchronization (Hattori et al. 2013) is needed in apes (cf. (Ravignani et al. 2013).

Evidence from non-primate species also seems to undermine Ackermann et al.'s model. Two bird species, both vocal learners, have been shown to entrain to steady pulses (Hasegawa et al. 2011; Patel et al. 2009a), supporting Ackermann et al.'s model and Patel's hypothesis, whereby auditory-motor entrainment skills would be evolutionary by-products of vocal learning abilities (Patel 2006). However, recent evidence suggests that vocal learning and rhythmic abilities might be dissociated. Sea lions, unlike seals, show no evidence of vocal learning (Janik & Slater 1997) but nonetheless can reliably synchronize their movements to a range of musical stimuli at different tempi (Cook et al. 2013). Humans and sea lions are both rhythmically skilled, but only humans evolved vocal learning and speech. Therefore, sea lions constitute outliers inconsistent with the prediction of Ackermann et al.'s model. This species evolved cognitive rhythmic abilities, without evolving speech. Invoking additional evolutionary forces and physiological mechanisms thus appears necessary: How can Ackermann et al.'s model be modified to avoid incorrectly predicting vocal learning in rhythmic-skilled species?

Second, Ackermann et al.'s model assumes that prosodic modulation of speech conveys mainly simple motivational-emotional information, and thus, that prosody and complex speech production had separate evolutionary histories. But evidence showing a tight connection between prosody and complex linguistic functions argues against this "double pathway" theory. Prosodic contour is influenced by syntactic constituent structure, semantic relations, phonological rhythm, pragmatic considerations, as well as by the length, complexity, and predictability of linguistic material (Wagner & Watson 2010). Furthermore, prosodic cues are used in childhood during acquisition of words (Christophe et al. 2008) and grammatical constructions (Männel et al. 2013), and in adulthood for syntactic processing (Christophe et al. 2008; Kjelgaard & Speer 1999; Langus et al. 2012; Wagner 2010) and word recognition (Cutler et al. 1997).

Contra Ackermann et al., such complex linguistic modulation of prosody seems to be a prerequisite for the acquisition and use of language, and this process is likely to be influenced by cognitive mechanisms specially modified in the human lineage. Comparative research on syntax precursors favors this hypothesis: The ability to assemble sequences of sounds into *hierarchical* patterns might be either human-specific, or very poorly developed in other species (Conway & Christiansen 2001; ten Cate & Okanoya 2012). Hence, developmental and comparative evidence point to a more complex cognitive integration of prosody and speech than allowed by the dual-pathway proposal of Ackermann et al. The challenge for Ackermann et al.'s theory is, therefore, to account for the modulation of prosody by human-specific cognitive functions (e.g., syntax), which are clearly not evolutionary homologues of primate emotional vocalizations controlled by the anterior cingulate cortex.

Finally, Ackermann et al. propose an ontogenetic pathway in which: (1) basal ganglia (BG) are important to generate integrated templates of orofacial and laryngeal movements during childhood, but (2) in adulthood can be retrieved from cortical areas because these motor templates become well-trained. Later in ontogeny, BG would mostly subserve the modulation of emotional prosody, and not the coordination of speech production. These claims are not supported by currently available empirical data. For instance, Ackermann et al. cite Parkinson's Disease (PD) data to support their claims that, in adults, BG lesions only impair emotional prosody. In fact, PD patients with normal cognitive functioning are more impaired in semantic fluency tasks than in phonetic fluency (Henry & Crawford 2004). Additionally, contra Ackermann et al., BG subserve complex syntactic and semantic processing in adults, with empirical findings consistent across PD (Dominey & Inui 2009; Henry & Crawford 2004; Lewis et al. 1998), BG lesion (Kotz et al. 2003; Teichmann et al. 2008; Ullman et al. 1997), and neuroimaging research (Friederici & Kotz 2003). These data suggest that in adults the BG support multiple functions relevant to spoken language, not just simple emotional prosodic modulation.

Furthermore, contrary to the developmental pathway proposed by Ackermann et al., the acquisition of novel syntactic structures in adults depends on the medial temporal cortex, and the retrieval of syntactic templates *after* thorough learning mostly recruits the BG and perisylvian structures (Ullman 2004). This evidence shows that, contra Ackermann et al., BG are active in the retrieval of over-learnt procedures. Ackermann et al. therefore need to propose alternative explanations to reconcile child and adult data concerning the function of BG.

In conclusion, to make their model robust, Ackermann et al. must modify and refine their evolutionary and mechanistic explanations, and clarify which assumptions are necessary, and which are sufficient, for their explanatory framework to hold. Is their model robust enough to stand up to the clear, strong relationship between prosody and complex linguistic functions? How can Ackermann et al.'s model account for the complex functions of BG in adulthood? If in-group cohesion had to be achieved, why was precise vocal control specifically selected for, rather than general non-vocal rhythmic abilities? These and other questions need to be addressed if Ackermann et al.'s model is to become convincing.

NOTE

1. Andrea Ravignani and Mauricio Martins contributed equally to this commentary as joint first authors.

ACKNOWLEDGMENTS

This work was supported by Fundação para a Ciência e Tecnologia grant SFRH/BD/64206/2009 (to Mauricio Martins) and European Research Council Advanced Grant 230604 SOMACCA (to Andrea Ravignani and W. Tecumseh Fitch).

Perceptual elements in brain mechanisms of acoustic communication in humans and nonhuman primates

doi:10.1017/S0140525X13004196

David H. Reser[a] and Marcello Rosa[a,b]

[a]*Department of Physiology, Monash University, Melbourne, VIC3800, Australia;* [b]*Australian Research Council Centre of Excellence for Integrative Brain Function, Monash University Node, Melbourne, VIC 3800.*