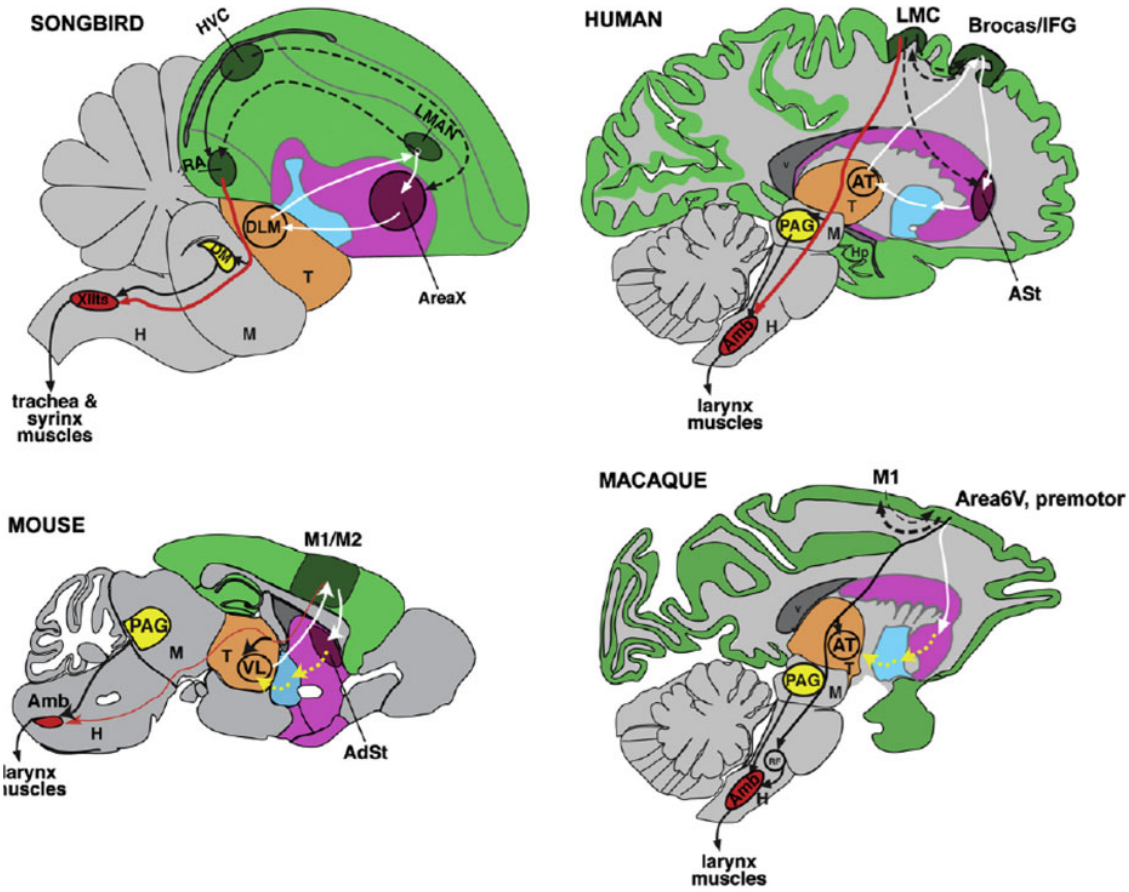

Figure 1 (Petkov & Jarvis). Summary diagrams of vocal systems in songbirds, humans, monkeys, and mice. Modified from Arriaga and Jarvis (2013). Cortico-striatal-thalamic loops are schematized from data in humans and songbirds. Yellow dashed lines in macaque monkeys and mice show proposed cortico-striatal-thalamic connections for vocalization that need to be tested.

Notably, the more precise link that the authors are pursuing with regard to the origins of spoken language and basal ganglia function, already has an evolutionary counterpart in vocal-learning and vocal-non-learning birds. The avian striatal vocal nucleus (called Area X in songbirds) sits within a cortico-striatal-thalamic loop, which is important for song learning (Jarvis 2004b; 2006; Jarvis et al. 2000), including covert-skill song learning (Charlesworth et al. 2012). Moreover, Feenders et al. (2008), by comparing the anterior-forebrain pathway in vocal-learning birds to this pathway in vocal-non-learning birds, found evidence to develop a motor theory of vocal-learning origin.

This theory proposes that the anterior-forebrain song pathway (including Area X) independently arose multiple times in vocal-learning birds from a set of regions that in vocal-non-learning birds control non-vocal motor actions. The discrete striatal Area X that sits within the cortico-striatal-thalamic vocal-learning loop (Fig. 1) is not present in vocal-non-learning birds. Motor striatal regions outside of Area X, or the comparable forebrain regions in vocal-non-learning birds, are more diffuse and relate to these animals' non-vocal motor learning abilities. Thus, considerable insights on the cortico-striatal-thalamic system have already been provided by avian models. These are only briefly alluded to but not meaningfully used to inform the current proposal.

In summary, Ackermann et al.'s proposal is an interesting review of the literature with an emphasis on the basal ganglia as an evolutionary substrate for spoken language. However, we found it heavy on conjecture and light on empirical hypotheses, which, as we have suggested, can be strengthened by (1) taking a broader evolutionary perspective that allows integrating data from birds and mammals, and (2) delineating more carefully how the current proposal can be integrated within or distinguished from other theories on spoken language origins.

The sensorimotor and social sides of the architecture of speech

doi:10.1017/S0140525X13004172

Giovanni Pezzulo,[a] Laura Barca,[a] and Alessando D'Ausilio[b]

[a]*Institute of Cognitive Sciences and Technologies, National Research Council, 00185 Rome, Italy;* [b]*Robotics, Brain and Cognitive Sciences Department, Italian Institute of Technology, 16163 Genova, Italy.*

giovanni.pezzulo@istc.cnr.it laura.barca@istc.cnr.it
alessandro.dausilio@iit.it
https://sites.google.com/site/giovannipezzulo/
https://sites.google.com/site/laurabarcahomepage/
http://www.iit.it/people/robotics-brain-and-cognitive-sciences-mirror-neurons-and-interaction-lab/researcher/alessandro-dausilio.html

Abstract: Speech is a complex skill to master. In addition to sophisticated phono-articulatory abilities, speech acquisition requires neuronal systems configured for vocal learning, with adaptable sensorimotor maps that couple heard speech sounds with motor programs for speech production; imitation and self-imitation mechanisms that can train the sensorimotor maps to reproduce heard speech sounds; and a "pedagogical" learning environment that supports tutor learning.

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