Andrea Ravignani* and Bart de Boer

# Joint origins of speech and music: testing evolutionary hypotheses on modern humans

https://doi.org/10.1515/sem-2019-0048
Received April 29, 2019; accepted October 5, 2019;
published online February 3, 2021

**Abstract:** How music and speech evolved is a mystery. Several hypotheses on their origins, including one on their joint origins, have been put forward but rarely tested. Here we report and comment on the first experiment testing the hypothesis that speech and music bifurcated from a common system. We highlight strengths of the reported experiment, point out its relatedness to animal work, and suggest three alternative interpretations of its results. We conclude by sketching a future empirical programme extending this work.

**Keywords:** cognitive semiotics; evolution of music; evolution of language; evolution of speech; zoosemiotics; animal communication

Why do humans have language, speech, and music? These traits are extremely common in our species, though their ultimate evolutionary functions are still hotly debated (Christiansen and Kirby 2003; Honing 2018). If they had a strong adaptive function, one would expect to find them in several other species (e.g., birds, bats, and insects all independently evolved wings to fly). However, speech-related and music-related abilities are rare in non-human animals and scattered across taxonomic groups. Biological outliers are interesting per se, especially in evolutionary terms, and even more so if humans are one of the few species to qualify as outliers.

Human language, speech, and music are particularly difficult to study via paleo-anthropology. For instance, speech and song rely on vocal folds and brain circuits that do not fossilize well. Overcoming these methodological issues, recent work in paleo-anthropology has nonetheless provided solid inference on human cognitive evolution (e.g., Blasi et al. 2019; Gunz et al. 2019). However, in many cases, the best available data for inference on human cognitive evolution comes from modern humans. This scientific methodology, however, assumes that cognitive processes have remained unchanged re from the time in the past one is interested in, until today's humans.

***Corresponding author: Andrea Ravignani**, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, E-mail: andrea.ravignani@mpi.nl. https://orcid.org/0000-0002-1058-0024
**Bart de Boer,** Vrije Universiteit Brussel, Brussel, Belgium, E-mail: bart@ai.vub.ac.be

A recent experiment made use of this idea to trigger a referential-emotional communicative distinction in human participants (Ma et al. 2019). The basic idea of the experiment was to test the musical protolanguage hypothesis (indirectly originated from Darwin's ideas; Darwin 1871): Speech and music should be sister systems, both resulting from a bifurcation of a "proto-musilanguage" (Brown 2017; Fitch 2005, 2013; Kirby 2011; Mithen 2005). Although Darwin (1871) did not formulate it exactly in these terms, this hypothesis has come to be called the "*Darwin's musical protolanguage hypothesis*" (Fitch 2013; Kirby 2011), and we will stick to this term for the rest of this paper. In the experiment by Ma and colleagues (2019), participants had to imitate vocalizations they heard, namely, nonsense syllable strings, associated with pictures showing either object-like referents or emotional states. To constrain the space of vocalizations produced and simulate the effect of cultural transmission, participants did not create their vocalizations from scratch. Instead, as in the game of telephone, each participant heard nonsense syllable streams recorded from previous participants, who were themselves recorded while imitating previous participants, etc. Over these experimental "generations," Ma and colleagues (2019) measured prosodic cues in the syllable streams. Additional participants also rated how music-like or speech-like the resulting systems were. In both cases, the main prediction was that the pictures with referential meaning would trigger more speech-like acoustic features and ratings, while those with emotional meaning would trigger more music-like acoustic features and ratings. This was indeed the case: both methods showed that emotional meaning triggers more music qualities than referential meaning. For instance, Ma and colleagues' (2019) emotional vocalizations have a larger pitch variation and broader sound intensity range than referential vocalizations. In brief, emotional vocalizations appear to trigger enhanced prosodic modulation, which dovetails with recent frameworks linking prosody and protolanguage (Filippi 2016; Filippi et al. 2017; Ma et al. 2019). In addition, the results on perceptual rating in English speakers replicate in Mandarin speakers (Ma et al. 2019).

The authors justly interpret their results as the first empirical test of the notion that a single system, a musical protolanguage, can bifurcate into two sub-systems, music and language (Ma et al. 2019). This is a more specific version of the general *Darwin's musical protolanguage hypothesis*, which has been discussed or tested before (either directly or indirectly, e.g., Brown 2017; Fenk-Oczlon and Fenk 2009; Fitch 2005, 2013; Kirby 2011; Lumaca and Baggio 2017; Mithen 2005; Nordström and Laukka 2019; Rauschecker 2018; Reybrouck and Podlipniak 2019; Thompson et al. 2012). Indeed, these results are a first step towards probing the validity of the bifurcation hypothesis. In particular, the authors aptly summarize their results as showing "that when a single system is used for both emotional and referential communication, it will tend to bifurcate into two systems with distinct characteristics" (Ma et al. 2019: 21). The experiment is a proof of concept that this bifurcation may have happened during human evolution. It is a proof of concept because, like many other studies on modern humans, it cannot rule out alternative explanations of three kinds.

First, Ma and colleagues' (2019) proof of concept tests one possible logical antecedent to the music-speech segregation in modern humans, our status quo. They show that if a particular condition occurred in our past, then the status quo in modern humans *could* follow. As we cannot be sure that the particular condition actually occurred, the human status quo may have originated via alternative paths. This is a common issue in language evolution research (Martins et al. 2014). A practical example of this reasoning involves gestural hypotheses for the origins of language. If human language originated in the gestural modality whereas music in the auditory modality, to then both converge to the same modality (Christiansen and Kirby 2003; Wallin et al. 2001), modern humans would still show the cognitive mapping seen in the current experiment (Ma et al. 2019). If this scenario were true, the current experiment would not be proof of *Darwin's musical protolanguage hypothesis*. It would provide an equally interesting insight, however: there is a (still unspecified) cognitive mechanism segregating music-like from speech-like signal-meaning mappings. In other words, even if speech and music had two distinct evolutionary paths, Ma and colleagues (2019) show that some cognitive mechanism must have *kept them apart*.

Second, participants are modern humans, for whom speech and music are already two distinct, cognitively-segregated systems. Hence, this experiment (like many similar ones, e.g., Kirby et al. 2008; Lumaca and Baggio 2017; Ravignani et al. 2016; Verhoef et al. 2014) may trigger biases already present in humans for other reasons. In other words, iterated learning experiments rely on the assumption that changes occurring to a cultural artifact in the minute-to-hours timescale mimic those that happened at much longer timescales in human evolution. While the latter cultural changes were mutually reinforced by cognitive adaptations, the former cultural changes developed on a fixed cognitive substrate, that of modern humans.

Third, both music and language include a strongly learned component. If participants were primed by their cultural medium to preferentially associate speech with referentiality, and music with emotion, this experiment would mainly generalize to Australian English speakers, rather than humans. It is unclear whether this mapping is universal across humans (Filippi 2016; Filippi et al. 2017; Savage et al. 2015), but detection of vocal (Juslin and Laukka 2003) and musical expression of emotions seems equally accurate across cultures. In addition, Ma and colleagues (2019) aptly test participants from two linguistic (and cultural) groups, finding no perceptual differences between groups. Finding no differences between linguistic groups alleviates concerns about cultural priming, and strengthens the link suggested by the authors. The logical next step would entail testing for potential *production* differences across cultures (not tested here), as these differences have been detected in other iterated learning experiments (e.g., Jacoby and McDermott 2017).

Ma and colleagues' (2019) experiment is sound, intriguing, and thought-provoking. However, it is only a first step towards testing joint evolutionary hypotheses of speech and music. This experiment focused on general spectro-temporal properties of imitated vocal signals. Obvious follow-up experiments should tackle the temporal and rhythmic dimensions of the transmitted signals. The authors' analyses focused on syllables' pitch, intensity, and rate. In particular, different communicative conditions significantly affected the syllable rate. In turn, syllable rate is a proxy for rhythmic structures potentially present in the signals, but it is not fine-grained enough to be a direct measure of these structures. Additional analyses of the available data could test whether the emotional condition triggers more isochronous vocalizations: isochrony is, in fact, a characteristic of several musical cultures, but is not typical of speech (Brown and Weishaar 2010; Jadoul et al. 2016; Savage et al. 2015). Collection of a new dataset would make it possible to test for the emergence of rhythmic structures in the music-like versus speech-like signals. In particular, new data should be collected, trying to obtain longer sequences of nonsense syllables (de Castro-Arrazola and Kirby 2019). In this case, one would expect the emergence of (1) a regular beat, (2) interonset intervals related by small integer ratios, and (3) metrical structures, like 2/4 (March) and 3/4 (Waltz) in the music/emotional condition but not in the speech/referential condition (Savage et al. 2015). The presence of a regular beat could be tested indirectly via data analysis (Norton and Scharff 2016; Ravignani et al. 2016) or directly by asking other participants to tap to the vocalizations (Van Dantzig 1940). Whether interonset intervals are related by small integer ratios could be tested analytically on the timing data between one syllable onset and the next (Jacoby and McDermott 2017; Ravignani et al. 2016). The presence of metrical structures could be inferred by combining durational data and intensity data, testing whether high-intensity peaks occur every two or three syllables (Ravignani et al. 2016).

A possible alternative experiment would be to test this hypothesis in a modality different from audition. Sign languages have many features analogous to classical phonetics and phonology (e.g., Nespor and Sandler 1999). In addition, choirs of signers exist, where members silently sing by signing. So, here we have a parallel test bench of the musical protolanguage hypothesis for visuo-motor speech and visuo-motor song. Repeating a similar experiment in signers deaf from birth would provide an even stronger test of *Darwin's musical protolanguage hypothesis*, pushing the burden of proof to the domain of cognitive neuroscience (e.g., Thompson et al. 2012). In fact, if signers showed a similar branching pattern to the speakers of the current experiment, *Darwin's musical protolanguage hypothesis* would be supported at a more domain-general, modular level. Functional neuroimaging data on overlapping networks would be needed (e.g., Belyk et al. 2018), ideally in a 2 × 2 design: speech versus song, and voice versus sign. The caveats on interpretability discussed above would still apply though. If deaf signers did not show a similar branching pattern to the speakers of the current experiment, we would not need to throw away *Darwin's musical protolanguage hypothesis*, but we would obtain evidence of its domain-specificity.

Zooming out from individual to group vocalizations, the acoustic features measured in this experiment could have either been themselves targets of bio-cultural evolution, or byproducts of bio-cultural evolutionary processes targeting other behavioral features. For instance, interactive turn-taking has been hypothesized as a key process in the evolution of human speech and language (Pika et al. 2018). It may be that syllables' pitch, intensity, and rate, which differ here between experimental conditions, are only proxies for interactive communication schemas differing between speech and music. Future experiments testing communicative exchanges on a short time scale (e.g., dialogue-like vs. jazz-improv duet) will enlighten this causality relationship. Quantitative measures to assess these similarities in solo displays versus interactive communication are readily available (Kello et al. 2017).

Apart from human experiments, the comparative cross-species approach is a powerful tool to answer questions about human traits that do not fossilize. How do Ma and colleagues' (2019) results connect to the comparative study of the evolution of animal communication? In other words, can one interpret Ma and colleagues' (2019) experiment in a comparative perspective? In a comparative framework, animal communication can highlight similarities across species, including our own (e.g., Stansbury and Janik 2019). Within animal acoustic communication, a classic distinction exists among alarm calls, contact calls and songs. Calls often convey referential meaning. Alarm calls are often employed to signal danger to other conspecifics (Manser et al. 2002), while contact calls are often used to signal one's presence and location (Snowdon and Hodun 1981). Songs are vocal displays often used to compete, entice a potential mate or defend a territory (Okobi et al. 2019). It would be tempting to establish a parallel between Ma and colleagues' (2019) vocalization types and types of animal signals: contact calls may map to human (speech-like) referential vocalizations and songs to human (music-like) emotional vocalizations. However, the mapping is not straightforward. In fact, the acoustic combinatorial structure of calls is usually quite simple, often entailing emission of one call or of a combination of two calls. Songs are more combinatorially complex, like both human music and speech. The complex "syntax" of songs, however, is not used to convey complex referential meanings as in human language; instead, songs are complex for the sake of "advertising complexity," a strategy often used to attract mates. In brief, it is far from clear how experiments like those from Ma and colleagues' (2019) will be situated within the broader comparative communication literature, but we suggest they should be.

To conclude, the study of the origins and evolution of music and speech is rapidly progressing. The last two decades have seen an empirical turn, going from armchair speculation to theoretically-driven hypothesis testing (Fisher 2017; Fitch 2017; Fröhlich et al. 2019; Honing 2018; Kirby et al. 2008; Ravignani et al. 2018). The intriguing work by Ma and colleagues (2019) should be source of much follow-up work and inspiration for design of experiments.

## References

Belyk, Michel, Joseph F. Johnson & Sonja A. Kotz. 2018. Poor neuro-motor tuning of the human larynx: A comparison of sung and whistled pitch imitation. *Royal Society Open Science* 5(4). 171544.

Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). eaav3218.

Brown, Steven. 2017. A joint prosodic origin of language and music. *Frontiers in Psychology* 8. 1894.

Brown, Steven & Kyle Weishaar. 2010. Speech is "heterometric": The changing rhythms of speech. *Paper presented at Speech Prosody 2010, Chicago*.

Christiansen, Morten H. & Simon Kirby. 2003. Language evolution: The hardest problem in science? In Morten H. Christiansen & Simon Kirby (eds.), *Language evolution*, 1–15. Oxford: Oxford University Press.

Darwin, Charles. 1871. *The descent of man and selection in relation to sex*, Vol. 1. London: John Murray.

de Castro-Arrazola, Varun & Simon Kirby. 2019. The emergence of verse templates through iterated learning. *Journal of Language Evolution* 4(1). 28–43.

Fenk-Oczlon, Gertraud & August Fenk. 2009. Some parallels between language and music from a cognitive and evolutionary perspective. *Musicae Scientiae* 13(2_suppl). 201–226.

Filippi, Piera. 2016. Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology* 7. 1393.

Filippi, Piera, Sebastian Ocklenburg, Daniel L. Bowling, Larissa Heege, Onur Güntürkün, Albert Newen & Bart de Boer. 2017. More than words (and faces): Evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion* 31(5). 879–891.

Fisher, Simon E. 2017. Evolution of language: Lessons from the genome. *Psychonomic Bulletin & Review* 24(1). 34–40.

Fitch, W. Tecumseh. 2005. Dancing to Darwin's tune. *Nature* 438(7066). 288.

Fitch, W. Tecumseh. 2013. Musical protolanguage: Darwin's theory of language evolution revisited. In Johan J. Bolhuis & Martin B. H. Everaert (eds.), *Birdsong, speech, and language: Exploring the evolution of mind and brain*, 489–503. Cambridge, MA: MIT Press.

Fitch, W. Tecumseh. 2017. Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review* 24(1). 3–33.

Fröhlich, Marlen, Christine Sievers, Simon W. Townsend, Townsend Gruber & Carel P. van Schaik. 2019. Multimodal communication and language origins: Integrating gestures and vocalizations. *Biological Reviews* 94(5). 1809–1829.

Gunz, Philipp, Amanda K. Tilot, Katharina Wittfeld, Alexander Teumer, Chin Yang Shapland, Theo G. Van Erp, Benjamin Vernot, Simon Neubauer, Tulio Guadalupe & Guillén Fernández. 2019. Neandertal introgression sheds light on modern human endocranial globularity. *Current Biology* 29(1). 120–127.

Honing, Henkjan. 2018. *The origins of musicality*. Cambridge: MIT Press.

Jacoby, Nori & Josh H. McDermott. 2017. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology* 27. 359–370.

Jadoul, Yannick, Andrea Ravignani, Bill Thompson, Piera Filippi & Bart de Boer. 2016. Seeking temporal predictability in speech: Comparing statistical approaches on 18 world languages. *Frontiers in Human Neuroscience* 10. 586.

Juslin, Patrik N. & Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129(5). 770–814.

Kello, Christopher T., Simone Dalla Bella, Butovens Médé & Ramesh Balasubramaniam. 2017. Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of the Royal Society Interface* 14(135). 20170231.

Kirby, Simon. 2011. Darwin's musical protolanguage: An increasingly compelling picture. In Patrick Rebuschat, Martin Rohmeier, John A. Hawkins & Ian Cross (eds.), *Language and music as cognitive systems*, 96–102. Oxford: Oxford University Press.

Kirby, Simon, Hannah Cornish & Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences* 105, 10681–10686.

Lumaca, Massimo & Giosué Baggio. 2017. Cultural transmission and evolution of melodic structures in multi-generational signaling games. *Artificial Life* 23(3). 406–423.

Ma, Weiyi, Anna Fiveash & William Forde Thompson. 2019. Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage. *Semiotica* 229(1/4). 1–23.

Manser, Marta B., Robert M. Seyfarth & Dorothy L. Cheney. 2002. Suricate alarm calls signal predator class and urgency. *Trends in Cognitive Sciences* 6(2). 55–57.

Martins, Mauricio, Archishman Raju & Andrea Ravignani. 2014. Evaluating the role of quantitative modelling in language evolution. In Luke McCrohon, Bill Thompson, Tessa Verhoef & Hajime Yamauchi (eds.), *The past, present and future of language evolution research*, 84–93. Kyoto: EvoLang9.

Mithen, Steven. 2005. *The singing neanderthals: The origins of music, language, mind and body*. London: Weidenfeld & Nicolson.

Nespor, Marina & Wendy Sandler. 1999. Prosody in Israeli sign language. *Language and Speech* 42(2–3). 143–176.

Nordström, Henrik & Petri Laukka. 2019. The time course of emotion recognition in speech and music. *Journal of the Acoustical Society of America* 145(5). 3058–3074.

Norton, Philipp & Constance Scharff. 2016. "Bird song metronomics": Isochronous organization of zebra finch song rhythm. *Frontiers in Neuroscience* 10. 309.

Okobi, Daniel E., Arkarup Banerjee, Andrew M. Matheson, Steven M. Phelps & Michael A. Long. 2019. Motor cortical control of vocal interaction in neotropical singing mice. *Science* 363(6430). 983–988.

Pika, Simone, Ray Wilkinson, Kobin H. Kendrick & Sonja C. Vernes. 2018. Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B: Biological Sciences* 285(1880). 0598.

Rauschecker, Josef P. 2018. Where did language come from? Precursor mechanisms in nonhuman primates. *Current Opinion in Behavioral Sciences* 21. 195–204.

Ravignani, Andrea, Tania Delgado & Simon Kirby. 2016. Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour* 1. 0007.

Ravignani, Andrea, Bill Thompson & Piera Filippi. 2018. The evolution of musicality: What can be learned from language evolution research? *Frontiers in Neuroscience* 12. 20.

Reybrouck, Mark & Piotr Podlipniak. 2019. Preconceptual spectral and temporal cues as a source of meaning in speech and music. *Brain Sciences* 9(3). 53.

Savage, Patrick E., Steven Brown, Emi Sakai & Thomas E. Currie. 2015. Statistical universals of human music. *Proceedings of the National Academy of Sciences* 112(29). 8987–8992.

Snowdon, Charles T. & Alexandra Hodun. 1981. Acoustic adaptation in pygmy marmoset contact calls: Locational cues vary with distances between conspecifics. *Behavioral Ecology and Sociobiology* 9(4). 295–300.

Stansbury, Amanda L. & Vincent M. Janik. 2019. Formant modification through vocal production learning in gray seals. *Current Biology* 29(13). 2244–2249.

Thompson, William Forde, Manuela M. Marin & Lauren Stewart. 2012. Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences* 109(46). 19027–19032.

Van Dantzig, Marijn. 1940. Syllable-tapping, a new method for the help of stammerers. *Journal of Speech Disorders* 5(2). 127–131.

Verhoef, Tessa, Simon Kirby & Bart De Boer. 2014. Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics* 43. 57–68.

Nils Lennart Wallin, Björn Merker & Steven Brown (eds.). 2001. *The origins of music*. Cambridge: MIT Press.