Signatures of speech and song: "Universal" links despite cultural diversity

Daniela Sammler[1,2]*

Equitable collaboration between culturally diverse scientists reveals that acoustic fingerprints of human speech and song share parallel relationships across the globe.

Humans all across the globe use their voices to talk and sing, although the exact forms of speech and song differ immensely between cultures. Curiously, no other species exhibits such a dual functional use of the vocal apparatus—although many species vocalize or even "sing." Which evolutionary pressures made speech and song emerge as two distinct vocal categories in humans? Is the divide a cultural construction or biologically grounded? And what makes speech and song distinguishable at all, in Europe, Asia, or Africa? In this issue of *Science Advances*, Ozaki *et al.* (*1*) find that people across a large range of societies, six continents and 55 languages, modify their voices in similar ways when they speak or sing. This cross-cultural consistency provides an important argument for speech and song being globally similar solutions of humans to similar communicative pressures, flavored by cultural practice.

The question why humans have two modes of vocal communication has led to countless speculations. Did speech and song diverge from a common precursor? Or did one mode evolve from the other? And what is their survival value in the first place? While it is not hard to imagine why verbal eloquence increased fitness and was, hence, passed on to new generations, the answer is far from clear for music. Why did humans or their ancient ancestors start to sing? To charm the opposite sex as proposed by Darwin (*2*)? To tie social bonds with offspring or peers (*3*)? Or to credibly signal group cohesion to strangers (*4*)? It is not easy to turn these theories from plausible tales into scientifically tenable facts. Speech and song do not fossilize, and the Neanderthals, who might have given us answers, died out long ago.

One way to address these questions is to study the products—speech and song—as we use them today and to do so across many (ideally all) societies and cultures. The idea is that elements of speech and song that can be found "universally" bring us closest to the prehistoric essence of these behaviors. These "universals" may be seen as "living fossils" buried under layers of cultural makeup that have been added over the course of evolutionary history.

Notably, this quest for "universality" does not mean to search for similar expressions or melodies across cultures. Rather, it means digging for deeper biocognitive principles and natural functions of human linguisticality and musicality that could explain global patterns of speech and song. These principles are thought to be firmly anchored in our species-specific genome, vocal anatomy, or brain structure and function. Take the volume of the lung, the degrees of freedom of the tongue, or the computational capacity of the brain as examples; they all provide natural constraints to how speech and song around the world can sound. The discovery of the constraints that make speech and song sound differently may bring us closer to answering the question why humans communicate in these two ways.

That speech and song sound differently seems clear to everyone. This clarity shrivels, however, when thinking of genres such as rap or sprechgesang (speech song) or when experiencing how looped speech turns magically into song in Diana Deutsch's famous speech-to-song illusion (https://deutsch.ucsd.edu/psychology/pages.php?i=212). In fact, many consider speech and song as end points of a continuum of vocal behaviors rather than two fixed categories (*5*). The distinction becomes even more blurred when looking at the large variety of languages spoken, and musics sung around the world: They are so diverse that comparing them has long been considered a waste of time. Fortunately, a few universals have crystallized recently, such as the predominance of simple integer ratios in musical rhythms (*6*) and a comparable information rate of speech across societies (*7*). However, this does not yet answer the question what distinguishes "speech" and "song" cross-culturally. Ozaki *et al.* took up that challenge in a large-scale registered report.

They started with two difficult decisions: Which sound features to compare between speech and song, and how to extract them? There are thousands of features to look at in sounds. They range from basics such as pitch height, brightness, and intensity that can be extracted directly from the sound's spectrogram to more structural features such as rhythmic regularity, contour shape, and pitch stability that rely on the segmentation of the continuous acoustic stream into perceptual units such as syllables and notes. Segmenting sounds is not trivial, and machine algorithms still lag behind human annotations. This leaves researchers with three choices: to either focus exclusively on basic features (*8*), to accept machine imprecisions as noise that might be mitigated by sampling from a gamut of recordings (*9*), or to recruit an army of human annotators who do the tedious work manually.

Ozaki *et al.* went for option three but with a twist: They shared the segmentation burden with 75 researchers, all coming from different corners of the world. In this global collaboration, each researcher provided a set of four manually annotated recordings of themselves, singing a traditional song of their culture, reciting the lyrics, talking about that song, and

[1]Max Planck Institute for Empirical Aesthetics, Research Group Neurocognition of Music and Language, Grüneburgweg 14, D-60322 Frankfurt am Main, Germany. [2]Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neuropsychology, Stephanstr. 1a, D-04103 Leipzig, Germany.
*Corresponding author. Email: daniela.sammler@ae.mpg.de

playing its melody on a musical instrument. Critics will immediately object that singing and speaking scientists are not a representative sample, and the authors do their best to rule this point out by comparing their findings to machine-analyzed recordings acquired in the field (*9*). What reinforced their choice was that it solved yet another problem of cross-cultural work: Judging behaviors of foreign cultures bears the risk of misinterpretations tinted by one's own cultural background. Ozaki *et al.*'s approach left the judgment to the experts of their own culture, who were—on top of that—also familiar with scientific rigor (not meaning that they were necessarily experts in musicology or linguistics).

Testing six preregistered hypotheses in their total of 300 hand-polished recordings, Ozaki *et al.* find that features praised in Western textbooks as particularly distinctive between speech and song indeed show up across cultures: Songs are globally slower and use higher and more stable pitches than speech. For some of these features, this distinction gets more pronounced when comparing instrumental music with speech, while recited lyrics take an intermediate position between speech and song. Have the authors put their finger on a global musi-linguistic continuum? Future studies will have to substantiate this idea with more diverse and nonacademic participants, more songs and more societies, more genres, more features, and other approaches. Still, Ozaki *et al.*'s results are reason enough to return to the question of which global drivers may explain the distinctive sound signatures of speech and song in the human species.

Bioacoustic theories provide a hot lead: They suggest that the acoustic form of a vocalization is systematically shaped by the social function of this vocal behavior (*10*). Auditory, vocal, and cognitive systems optimally tuned to perceiving and producing these forms offer a major selection advantage. Take human infant cries, heartfelt laughter, or aggressive roars as examples: They all exhibit characteristic sound signatures tailored to naturally tickle specific auditory and affective brain regions to elicit care, companionship, or rivalry, respectively. Speech and song are thought to also follow such form-function relationships (*9*) to optimally meet their opposite communicative goals (*5*). High pitched notes outside one's normal vocal profile sound louder and burn energy, suitable for signaling one's effort, arousal, and affect in songs; a speech given this way would be just tiring (although possible when emotional or infant-directed). Long notes sung on stable pitches are easier to mimic, and this is fundamental for establishing synchrony and harmony in social groups. Messaging at such slow speed would simply lack efficiency. While the present data cannot tell which functions exactly are shaping the acoustic forms of speech and song, knowing the acoustic features resolves one important variable in the bioacoustic equation.

Ways forward are twofold. First, instead of measuring sound features in speech and song recordings, these features could be explicitly manipulated, one by one, to test their effects on listeners in different cultures. This will allow gaining deeper, causal insights into putative evolved functions. Fortunately, the field of sound engineering is showing growing interest in musical and vocal signals, and promising software is becoming available (*10*). Second, instead of inferring functions from forms, these functions could be explicitly engineered, one by one, to test their effects on a communication system. Innovative transmission chain experiments with computers or humans can test whether and how initially artificial sound systems diverge into vocalizations with speech- and song-like properties depending on the functions ascribed to these systems, e.g., to refer to objects in the world or to express affective meaning (*11*). Moreover, tracking the emergence of these properties in the brain will allow identifying how neurobiological constraints may have shaped the divide of speech and song over the course of human evolution (*12*).

Much remains to be done, and only global and multidisciplinary collaborations will make it possible to reconstruct the evolutionary past of human vocal behavior. With such joint efforts, the quest for the origins of speech and song will continue to be an exciting and deeply insightful ride.

REFERENCES

1. Y. Ozaki, A. Tierney, P. Q. Pfordresher, J. McBride, E. Benetos, P. Proutskova, G. Chiba, F. Liu, N. Jacoby, S. C. Purdy, P. Opondo, W. T. Fitch, S. Hegde, M. Rocamora, R. Thorne, F. Nweke, D. P. Sadaphal, P. M. Sadaphal, S. Hadavi, S. Fujii, S. Choo, M. Naruse, U. Ehara, L. Sy, M. Lenini Parselelo, M. Anglada-Tort, N. C. Hansen, F. Haiduk, U. Færøvik, V. Magalhães, W. Krzyżanowski, O. Shcherbakova, D. Hereld, B. Suyanne Barbosa, M. A. Correa Varella, M. van Tongeren, P. Dessiatnitchenko, S. Zar Zar, I. El Kahla, O. Muslu, J. Troy, T. Lomsadze, D. Kurdova, C. Tsope, D. Fredriksson, A. Arabadjiev, J. P. Sarbah, A. Arhine, T. Ó Meachair, J. Silva-Zurita, I. Soto-Silva, N. E. Muñoz Millalonco, R. Ambrazevičius, P. Loui, A. Ravignani, Y. Jadoul, P. Larrouy-Maestri, C. Bruder, T. Puri Teyxokawa, U. Kuikuro, R. Natsitsabui, N. Bello Sagarzazu, L. Raviv, M. Zeng, S. Dabaghi Varnosfaderani, J. S. Gómez-Cañón, K. Kolff, C. Vanden Bosch der Nederlanden, M. Chhatwal, R. M. David, I. P. G. Setiawan, G. Lekakul, V. N. Borsan, N. Nguqu, P. E. Savage, Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech: A registered report. *Sci. Adv.* **10**, adm9797 (2024).
2. C. Darwin, *The Descent of Man, and Selection in Relation to Sex* (John Murray, 1871).
3. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. Mithen, W. T. Fitch, Music as a coevolved system for social bonding. *Behav. Brain Sci.* **44**, e59 (2021).
4. S. A. Mehr, M. W. Krasnow, G. A. Bryant, E. H. Hagen, Origins of music in credible signaling. *Behav. Brain Sci.* **44**, e60 (2021).
5. F. Haiduk, W. T. Fitch, Understanding design features of music and language: The choric/dialogic distinction. *Front. Psychol.* **13**, 786899 (2022).
6. N. Jacoby, R. Polak, J. A. Grahn, D. J. Cameron, K. M. Lee, R. Godoy, E. A. Undurraga, T. Huanca, T. Thalwitzer, N. Doumbia, D. Goldberg, E. H. Margulis, P. C. M. Wong, L. Jure, M. Rocamora, S. Fujii, P. E. Savage, J. Ajimi, R. Konno, S. Oishi, K. Jakubowski, A. Holzapfel, E. Mungan, E. Kaya, P. Rao, M. A. Rohit, S. Alladi, B. Tarr, M. Anglada-Tort, P. M. C. Harrison, M. J. McPherson, S. Dolan, A. Durango, J. H. McDermott, Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries. *Nat. Hum. Behav.* **4**, 10.1038/s41562-023-01800-9 (2024).
7. C. Coupé, Y. M. Oh, D. Dediu, F. Pellegrino, Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.* **5**, eaaw2594 (2019).
8. P. Albouy, S. A. Mehr, R. S. Hoyer, J. Ginzburg, R. J. Zatorre, Spectro-temporal acoustic markers differentiate speech from song across cultures. *bioRxiv* 2023.01.29.526133 [Preprint] (2023); https://doi.org/10.1101/2023.01.29.526133.
9. C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, A. Galbarczyk, G. Jasienska, C. T. Ross, M. Beth Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, A. Schachner, T. A. Vardy, Q. D. Atkinson, A. Salenius, J. Andelin, J. Antfolk, P. Madhivanan, A. Siddaiah, C. D. Placek, G. Deniz Salali, S. Keestra, M. Singh, S. A. Collins, J. Q. Patton, C. Scaff, J. Stieglitz, S. Ccari Cutipa, C. Moya, R. R. Sagar, M. Anyawire, A. Mabulla, B. M. Wood, M. M. Krasnow,S. A. Mehr, Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **6**, 1545–1556 (2022).
10. K. Pisanski, G. A. Bryant, C. Cornec, A. Anikin, D. Reby, Form follows function in human nonverbal vocalisations. *Ethol. Ecol. Evol.* **34**, 303–321 (2022).
11. W. Ma, A. Fiveash, W. F. Thompson, Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage. *Semiotica* **229**, 1–23 (2019).
12. M. Lumaca, L. Bonetti, E. Brattico, G. Baggio, A. Ravignani, P. Vuust, High-fidelity transmission of auditory symbolic material is associated with reduced right-left neuroanatomical asymmetry between primary auditory regions. *Cereb. Cortex* **33**, 6902–6916 (2023).

10.1126/sciadv.adp9620