Acoustic information about upper limb movement in voicing

Wim Pouw[a,b,c,1], Alexandra Paxton[a,d], Steven J. Harrison[a,e], and James A. Dixon[a,d]

[a]Center for the Ecological Study of Perception and Action, University of Connecticut, Storrs, CT 06269; [b]Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, Nijmegen 6525 HR, The Netherlands; [c]Max Planck Institute for Psycholinguistics, Max Planck Institute Nijmegen, Nijmegen 6525 XD, The Netherlands; [d]Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269; and [e]Department of Kinesiology, University of Connecticut, Storrs, CT 06269

Edited by Asif A. Ghazanfar, Princeton University, Princeton, NJ, and accepted by Editorial Board Member Peter L. Strick March 23, 2020 (received for review March 5, 2020)

We show that the human voice has complex acoustic qualities that are directly coupled to peripheral musculoskeletal tensioning of the body, such as subtle wrist movements. In this study, human vocalizers produced a steady-state vocalization while rhythmically moving the wrist or the arm at different tempos. Although listeners could only hear and not see the vocalizer, they were able to completely synchronize their own rhythmic wrist or arm movement with the movement of the vocalizer which they perceived in the voice acoustics. This study corroborates recent evidence suggesting that the human voice is constrained by bodily tensioning affecting the respiratory–vocal system. The current results show that the human voice contains a bodily imprint that is directly informative for the interpersonal perception of another's dynamic physical states.

vocalization acoustics | hand gesture | interpersonal synchrony | motion tracking

Human speech is a marvelously rich acoustic signal, carrying communicatively meaningful information on multiple levels and timescales (1–4). Human vocal ability is held to be much more advanced compared to our closest living primate relatives (5). Yet despite all its richness and dexterity, human speech is often complemented with hand movements known as co-speech gesture (6). Current theories hold that co-speech gestures occur because they visually enhance speech by depicting or pointing to communicative referents (7, 8). However, speakers do not just gesture to visually enrich speech: Humans gesture on the phone when their interlocutor cannot see them (9), and congenitally blind children even gesture to one another in ways indistinguishable from gestures produced by sighted persons (10).

Co-speech gestures, no matter what they depict, further closely coordinate with the melodic aspects of speech known as prosody (11). Specifically, gesture's salient expressions (e.g., sudden increases in acceleration or deceleration) tend to align with moments of emphasis in speech (12–17). Recent computational models trained on associations of gesture and speech acoustics from an individual have succeeded in producing very natural-looking synthetic gestures based on novel speech acoustics from that same individual (18), suggesting a very tight (but person-specific) relation between prosodic–acoustic information in speech and gestural movement. Such research dovetails with remarkable findings that speakers in conversation who cannot see and only hear each other tend to synchronize their postural sway (i.e., the slight and nearly imperceptible movement needed to keep a person upright) (19, 20).

Recent research suggests that there might indeed be a fundamental link between body movements and speech acoustics: Vocalizations were found to be acoustically patterned by peripheral upper limb movements due to these movements also affecting tensioning of respiratory-related muscles that modulate vocal acoustics (21). This suggests that the human voice has a further complexity to it, carrying information about movements (i.e., tensioning) of the musculoskeletal system. In the current study we investigate whether listeners are able to perceive upper limb movement information in human voicing.

Methods and Materials

To assess whether listeners can detect movement from vocal acoustics, we assessed whether listeners could synchronize their arm or wrist movement by listening to vocalizers who were instructed to move their arm or wrist at different tempos. We first collected naturalistic data from six prestudy participants (vocalizers; three each cisgender males and females) who phonated the vowel /ə/ (as in cinema) with one breath while moving the wrist or arm in rhythmic fashion at different tempos (slow vs. medium vs. fast). Prestudy participants were asked to keep their vocal output as stable and monotonic as possible while moving their upper limbs.

Movement tempo feedback was provided by a green bar that visually represented the duration of the participant's immediately prior movement cycle (as measured through the motion tracking system) relative to that specified by the target tempo (Fig. 1*A*). Participants were asked to keep the bar within a particular region (i.e., within 10% of the target tempo). The green bar therefore provided information about their immediately previous movement tempo relative to the prescribed tempo without the visual representation moving at that tempo itself. It is important to note that vocalizers were thus not exposed to an external rhythmic signal, such as a (visual) metronome. Further note that we found in an earlier study that when vocalizers move at their own preferred tempo—with no visual feedback about movement tempo—acoustic modulations are also obtained that are tightly synchronized with movement cycles (22). If participants vocalize without

Significance

We show that the human voice carries an acoustic signature of muscle tensioning during upper limb movements which can be detected by listeners. Specifically, we find that human listeners can synchronize their own movements to very subtle wrist movements of a vocalizer only by listening to their vocalizations and without any visual contact. This study shows that the human voice contains information about dynamic bodily states, breaking ground for our understanding of the evolution of spoken language and nonverbal communication. The current findings are in line with other research on nonhuman animals, showing that vocalizations carry information about bodily states and capacities.

Author contributions: W.P., A.P., S.J.H., and J.A.D. designed research; W.P. performed research; W.P. analyzed data; and W.P. wrote the paper with critical revisions by A.P., S.J.H., and J.A.D.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.A.G. is a guest editor invited by the Editorial Board.

Data deposition: The hypotheses and methodology have been preregistered on the Open Science Framework (OSF; https://osf.io/ygbw5/). The data and analysis scripts supporting this study can be found on OSF (https://osf.io/9843h/).

[1]To whom correspondence may be addressed. Email: w.pouw@psych.ru.nl.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2004163117/-/DCSupplemental.

First published May 11, 2020.

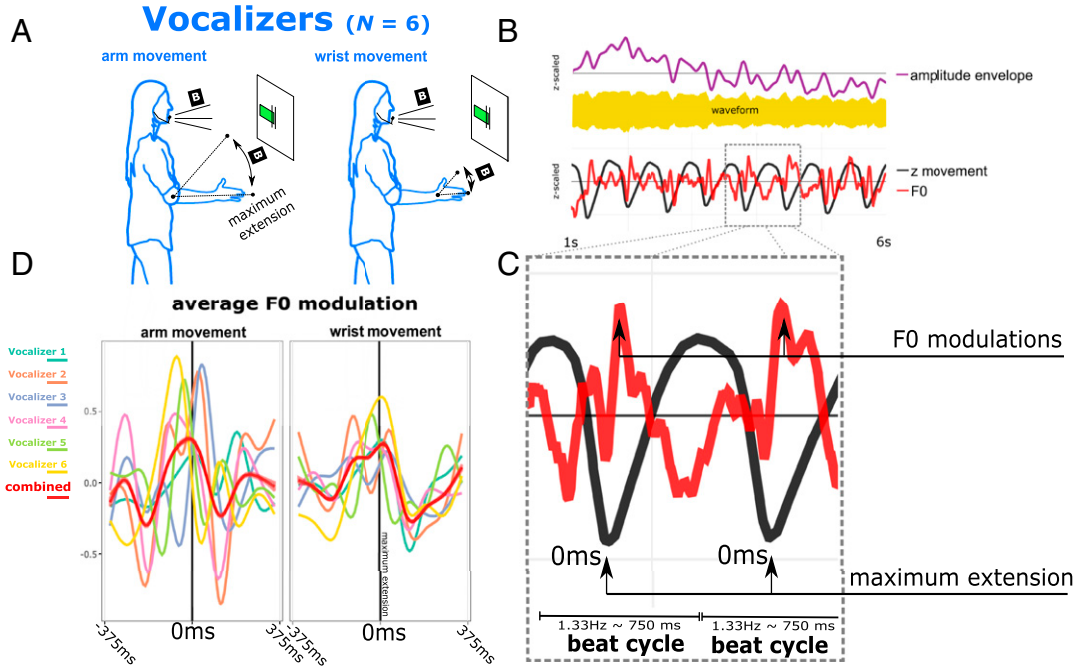

Fig. 1. Vocalizer movements (*A*) and resultant acoustic patterning caused by movement (*B*). (*A*) Six vocalizers moved their wrist and arm in rhythmic fashion at different tempos (slow = 1.06 Hz; medium = 1.33 Hz; fast = 1.6 Hz) that was guided via a green bar digitally connected to a motion-tracking system, which represented their movement frequency relative to the target tempo. Human postures modified from ref. 23. (*B*) The resultant movement and acoustic data were collected. Preanalysis indeed showed that acoustics were affected by movement, with sharp peaks in the fundamental frequency (perceived as pitch; *C*) and the smoothed amplitude envelope of the vocalization (in purple, *B*) when movements reached peaks in deceleration during the stopping motion at maximum extension. Peaks in deceleration of the movement lead to counteracting muscular adjustments throughout the body recruited to keep postural integrity, which also cascade into vocalization acoustics. (*D*) Here we assessed how the fundamental frequency of voicing (in the human range: 75 to 450 Hz) was modulated around the maximum extension for each vocalizer and combined for all vocalizers (red line). *D* shows that smoothed-average-normalized F0 (also linearly detrended and z-scaled per vocalization trial) peaked around the moment of the maximum extension, when a sudden deceleration and acceleration occurred; normalized F0 dipped at steady-state low-physical-impetus moments of the movement phase (when velocity was constant), rising again for a maximum flexion (~300 to 375 ms before and after the maximum extension), replicating previous work (21, 24). Vocalizer wrist movement showed a less pronounced F0 modulation compared to the vocalizer arm movement trials. For individual vocalizer differences for each tempo condition, see our interactive graph provided in the *SI Appendix*).

movements, however, acoustic modulations are absent (21). Similar to previous research (21, 22), in the current study hand movements inadvertently affected voice acoustics of these prestudy vocalizer participants (Fig. 1*D*), thereby providing a possible information source for listeners in the main study.

In the main study, 30 participants (listeners; 15 each cisgender males and females) were instructed to synchronize their own movements with the vocalizer's wrist and arm movements while only having access to the vocalizations of these prestudy participants presented via a headphone (for detailed materials and method, see *SI Appendix*). Thirty-six vocalizations (6 different vocalizers × 3 tempos × 2 vocalizer wrist vs. arm movements) were presented twice to listeners, once when they were instructed to synchronize with the vocalizer with their own wrist movement and once with their own arm movement. If listeners can synchronize the tempo and phasing of their movements to those of the vocalizers, this would provide evidence that voice acoustics may inform about bodily tensioned states—even when the vocalizer does not have an explicit goal of interpersonal communication.

The ethical review committee of the University of Connecticut approved this study (approval H18-260). All participants signed an informed consent, and vocalizer prestudy participants also signed an audio release form.

Data and Materials Availability. The hypotheses and methodology have been preregistered on the Open Science Framework (OSF; https://osf.io/ygbw5/). The data and analysis scripts supporting this study can be found on OSF (https://osf.io/9843h/).

Results

In keeping with our hypotheses, we found that listeners were able to detect and synchronize with movement from vocalizations (for these results, see Fig. 2; for detailed results, see *SI Appendix*). Listeners reliably adjusted their wrist and arm movement tempo to the slow, medium, and fast tempos performed by the vocalizers. Furthermore, listeners' circular means of the relative phases (Φ) were densely distributed around 0° (i.e., close to perfect synchrony), with an overall negative mean asynchrony of 45°, indicating that the listener slightly anticipated the vocalizer. Surprisingly—and against our original expectations—we even found that this held for the harder-to-detect vocalizer wrist movement. The variability of relative phase (as measured by circular SD Φ) was, however, slightly increased for wrist vs. arm vocalizations, with 0.28 increase in circular SD Φ; this indicated that listeners had greater difficulty synchronizing in phase with the vocalizer's wrist versus arm movements.

Discussion

We conclude that vocalizations carry information about upper limb movements of the vocalizer, given that listeners can adjust and synchronize to the movements by audition of vocalization alone. Importantly, this tempo and phase synchronization was not an artifact of chance since three different movement tempos were presented in random order. Nor are these effects reducible to idiosyncrasies in the vocalizers, as these patterns were observed across six different vocalizers with different voice acoustic qualities (e.g., cisgender male and female vocalizers). Further, vocalizers were not deliberately coupling vocal output with movement and were actually likely to try to inhibit these effects, as they had been instructed to keep their vocal output as stable

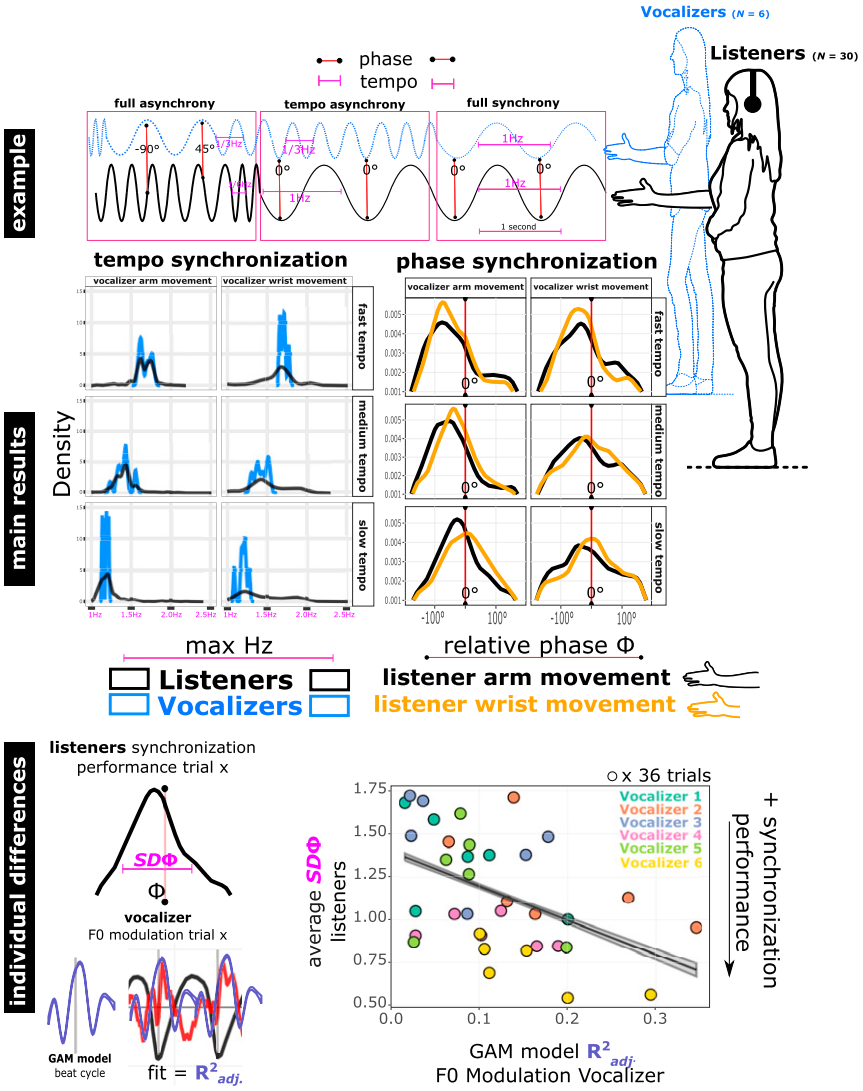

Fig. 2. Synchrony results. The example shows different ways movements can synchronize between the listener and the vocalizer. Fully asynchronous movement would entail a mismatch of movement tempo and a random variation of relative phases. Synchronization of phases may occur without exact matching of movement tempos. Full synchronization entails tempo matching and 0° relative phasing between vocalizer and listener movement. Main results show clear tempo synchronization, as the observed frequencies for each vocalization trial were well matched to the observed movement frequencies of listeners moving to that trial. Similarly, phase synchronization was clearly apparent, as phasing distributions are all pronouncedly peaked rather than having flat distributions, with a negative mean asynchrony regardless of vocalizer movement or movement tempo. Individual differences in vocalizer F0 modulations for each vocalizer trial were modeled using a nonlinear regression method, generalized additive modeling (GAM), providing a model fit (R^2 adjusted) for each trial, indicating the degree of variability of normalized F0 modulations around moments of the maximum extension (also see Fig. 1*D*). The variance explained for each vocalizer trial then was regressed against the average synchronization performance (average circular SD relative phase, SD Φ) of that trial by the listeners. It can be seen that more structural F0 modulations around the maximum extensions of upper limb movement (higher R^2 adjusted) predict better synchronization performance (lower SD Φ), $r = -0.48$, $P < 0.003$. This means that more reliable acoustic patterning in vocalizer's voicing predicts higher listener synchronization performance. Human postures modified from ref. 23.

as possible. Thus, the type of acoustic patterning affecting voice acoustics is a pervasive phenomenon and difficult to counteract.

Our understanding of the coupling between acoustic and motor domains is enriched by the present findings that the information about bodily movement is present in acoustics. Previous research has shown, for example, that a smoothed envelope of the speech amplitude closely correlates to mouth-articulatory movements (25). Indeed, seeing (26) or even manually feeling (27) articulatory movements may resolve auditorily ambiguous sounds that are artificially morphed by experimenters, leading listeners to hear a "pa" rather than a "da" depending on the visual or haptic information of the speaker's lips. The current results add another member to the family of acoustic–motor couplings by showing both that human voice contains acoustic signatures of hand movements and that human listeners are keenly sensitive to it.

Hand gestural movements may thus have evolved as an embodied innovation for vocal control, much like other bodily constraints on acoustic properties of human vocalization (27, 28). It is well established that information about bodies from vocalizations is exploited in the wild by nonhuman species (29). For example, rhesus monkeys associate age-related body size differences of conspecifics from acoustic qualities of "coos" (30). Orangutans even try to actively exploit this relation: They cup their hands in front of their mouths when vocalizing, changing the sound quality, presumably so as to acoustically appear more threatening in size (31). Humans, too, can predict with some

success the upper body strength of male vocalizers (32), especially from roaring as opposed to, for example, screaming vocalizations (33). The current results add to this literature that peripheral upper limb movements imprint their presence on the human voice as well, providing an information source about dynamically changing bodily states. An implication of the current findings is that speech recognition systems may be improved when becoming sensitive to these acoustic–bodily relations.

With the current results in hand, it becomes thus possible that hearing the excitement of a friend on the phone is, in part and at times, perceived by us through the gesture-induced acoustics that are directly perceived as bodily tensions. Gestures, then, are not merely seen—they may be heard, too.

ACKNOWLEDGMENTS. This research has been funded by The Netherlands Organisation of Scientific Research (NWO; Rubicon Grant "Acting on Enacted Kinematics," Grant no. 446-16-012; PI W.P.). In writing the research report, W.P. has further been supported by a DCC fellowship awarded by the Donders Institute for Brain, Cognition and Behaviour, and a postdoctoral position within the Language in Interaction Consortium (Gravitation Grant 024.001.006 funded by the NWO).

1. A. Ravignani *et al.*, Rhythm in speech and animal vocalizations: A cross-species perspective. *Ann. N. Y. Acad. Sci.* **1453**, 79–98 (2019).
2. D. H. Abney, A. Paxton, R. Dale, C. T. Kello, Complexity matching in dyadic conversation. *J. Exp. Psychol. Gen.* **143**, 2304–2315 (2014).
3. E. D. Jarvis, Evolution of vocal learning and spoken language. *Science* **366**, 50–54 (2019).
4. P. Hagoort, The neurobiology of language beyond single-word processing. *Science* **366**, 55–58 (2019).
5. A. A. Ghazanfar, Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behav. Ecol. Sociobiol.* **67**, 1441–1448 (2013).
6. J. Holler, S. C. Levinson, Multimodal language processing in human communication. *Trends Cogn. Sci.* **23**, 639–652 (2019).
7. D. McNeill, *Hand and Mind: What Gestures Reveal about Thought* (University of Chicago Press, Chicago, IL, 1992).
8. A. Kendon, Reflections on the "gesture-first" hypothesis of language origins. *Psychon. Bull. Rev.* **24**, 163–170 (2017).
9. J. Bavelas, J. Gerwing, C. Sutton, D. Prevost, Gesturing on the telephone: Independent effects of dialogue and visibility. *J. Mem. Lang.* **58**, 495–520 (2008).
10. J. M. Iverson, S. Goldin-Meadow, The resilience of gesture in talk: Gesture in blind speakers and listeners. *Dev. Sci.* **4**, 416–422 (2001).
11. P. Wagner, Z. Malisz, S. Kopp, Gesture and speech in interaction: An overview. *Speech Commun.* **57**, 209–232 (2014).
12. A. Rochet-Capellan, S. Fuchs, Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130399 (2014).
13. W. Pouw, J. A. Dixon, Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cogn. Sci.* **43**, e12721 (2019).
14. D. P. Loehr, Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab. Phonol.* **3**, 71–89 (2012).
15. N. Esteve-Gibert, P. Prieto, Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *J. Speech Lang. Hear. Res.* **56**, 850–864 (2013).
16. M. Chu, P. Hagoort, Synchronization of speech and gesture: Evidence for interaction in action. *J. Exp. Psychol. Gen.* **143**, 1726–1741 (2014).
17. B. Parrell, L. Goldstein, S. Lee, D. Byrd, Spatiotemporal coupling between speech and manual motor actions. *J. Phonetics* **42**, 1–11 (2014).
18. S. Ginosar *et al.*, "Learning individual styles of conversational gesture" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, L. Davis, P. Torr, and S.-C. Zhu, Eds. (IEEE Xplore, Long Beach, CA, 2019), pp. 3497–3506.
19. K. Shockley, M.-V. Santana, C. A. Fowler, Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* **29**, 326–332 (2003).
20. K. Shockley, A. A. Baker, M. J. Richardson, C. A. Fowler, Articulatory constraints on interpersonal postural coordination. *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 201–208 (2007).
21. W. Pouw, S. H. Harrison, J. A. Dixon, Gesture-speech physics: The biomechanical basis of the emergence of gesture-speech synchrony. *J. Exp. Psychol. Gen.* 149, 391–404 (2020).
22. W. Pouw, S. A. Harrison, J. A. Dixon, The physical basis of gesture-speech synchrony: Exploratory study and pre-registration. https://doi.org/10.31234/osf.io/9fzsv (20 August 2018).
23. Dimensions.Guide, Standing - Female (Side) Dimensions & Drawings | Dimensions.Guide. Retrieved from https://www.dimensions.guide/element/standing-female-side. Accessed 19 April 2020.
24. W. Pouw, S. A. Harrison, N. E. Gibert, J. A. Dixon, Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. https://doi.org/10.31234/osf.io/rnpav (27 November 2019).
25. C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, A. A. Ghazanfar, The natural statistics of audiovisual speech. *PLoS Comput. Biol.* **5**, e1000436 (2009).
26. H. McGurk, J. MacDonald, Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976).
27. C. A. Fowler, D. J. Dekle, Listening with eye and hand: Cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 816–828 (1991).
28. D. E. Blasi *et al.*, Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* **363**, eaav3218 (2019).
29. K. Pisanski, V. Cartei, C. McGettigan, J. Raine, D. Reby, Voice modulation: A window into the origins of human vocal control? *Trends Cogn. Sci.* **20**, 304–318 (2016).
30. A. A. Ghazanfar *et al.*, Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr. Biol.* **17**, 425–430 (2007).
31. M. E. Hardus, A. R. Lameira, C. P. Van Schaik, S. A. Wich, Tool use in wild orang-utans modifies sound production: A functionally deceptive innovation? *Proc. Biol. Sci.* **276**, 3689–3694 (2009).
32. K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. M. O'Connor, D. R. Feinberg, Return to Oz: Voice pitch facilitates assessments of men's body size. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1316–1331 (2014).
33. J. Raine, K. Pisanski, R. Bond, J. Simner, D. Reby, Human roars communicate upper-body strength more effectively than do screams or aggressive and distressed speech. *PLoS One* **14**, e0213034 (2019).