Vocal size exaggeration may have contributed to the origins of vocalic complexity

Katarzyna Kasia Pisanski, Andrey Anikin, David Reby

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Pisanski K, Anikin A, Reby D. 2021 Vocal size exaggeration may have contributed to the origins of vocalic complexity. *Phil. Trans. R. Soc. B* **376**: 20200401.
https://doi.org/10.1098/rstb.2020.0401

Accepted: 14 June 2021

One contribution of 12 to a theme issue 'Voice modulation: from origin and mechanism to social impact (Part II)'.

Subject Areas:
behaviour, cognition, evolution

Keywords:
body size, vocal tract length, voice modulation, acoustic communication, formants, speech articulation

Authors for correspondence:
Katarzyna Pisanski
e-mail: katarzyna.pisanski@cnrs.fr
David Reby
e-mail: d.reby@me.com

†Present address: CNRS, French National Centre for Scientific Research | DDL Lab, University of Lyon, 69007 Lyon, France.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5660017.

THE ROYAL SOCIETY PUBLISHING

Vocal size exaggeration may have contributed to the origins of vocalic complexity

Katarzyna Pisanski[1,†], Andrey Anikin[1,2] and David Reby[1]

[1]Equipe de Neuro-Ethologie Sensorielle, CNRS and Jean Monnet University of Saint Étienne, UMR 5293, 42023, St-Étienne, France
[2]Division of Cognitive Science, Lund University, 22100 Lund, Sweden

KP, 0000-0003-0992-2477; AA, 0000-0002-1250-8261; DR, 0000-0001-9261-1711

Vocal tract elongation, which uniformly lowers vocal tract resonances (formant frequencies) in animal vocalizations, has evolved independently in several vertebrate groups as a means for vocalizers to exaggerate their apparent body size. Here, we propose that smaller speech-like articulatory movements that alter only individual formants can serve a similar yet less energetically costly size-exaggerating function. To test this, we examine whether uneven formant spacing alters the perceived body size of vocalizers in synthesized human vowels and animal calls. Among six synthetic vowel patterns, those characterized by the lowest first and second formant (the vowel /u/ as in 'boot') are consistently perceived as produced by the largest vocalizer. Crucially, lowering only one or two formants in animal-like calls also conveys the impression of a larger body size, and lowering the second and third formants simultaneously exaggerates perceived size to a similar extent as rescaling all formants. As the articulatory movements required for individual formant shifts are minor compared to full vocal tract extension, they represent a rapid and energetically efficient mechanism for acoustic size exaggeration. We suggest that, by favouring the evolution of uneven formant patterns in vocal communication, this deceptive strategy may have contributed to the origins of the phonemic diversification required for articulated speech.

This article is part of the theme issue 'Voice modulation: from origin and mechanism to social impact (Part II)'.

1. Introduction

The vibrating vocal folds of a vocalizing animal produce an acoustic signal, the frequency of which (fundamental frequency, f_o) is perceived as pitch. This sound then passes through the vocal tract where the amplification of specific frequency components leads to the formation of formant frequencies. Formants are broadbands of acoustic energy whose centre frequencies depend on vocal tract length (VTL) and vocal tract shape [1,2]. While f_o tends to be lower in large animals, there are numerous exceptions to this rule among mammals, particularly within species and sexes [3,4]. By contrast, because formant frequencies are inversely related to VTL, formant spacing provides the most reliable cue to body size in terrestrial mammals, including humans [5–7]. Indeed, taller men and women will tend to have longer vocal tracts than shorter individuals [8] and will produce speech with uniformly lower formant frequencies [5]. Numerous studies have further shown that human listeners strongly associate voices with uniformly low formant frequencies and thus narrow formant spacing with a relatively large body size, and interestingly, also associate a low-pitched voice (lowered f_o) with largeness despite the lack of a robust pitch–size relationship between human adults of the same sex [9–13].

Because body size is a crucial factor affecting the outcome of social interactions in animals, notably in the context of sexual selection [14], strong evolutionary

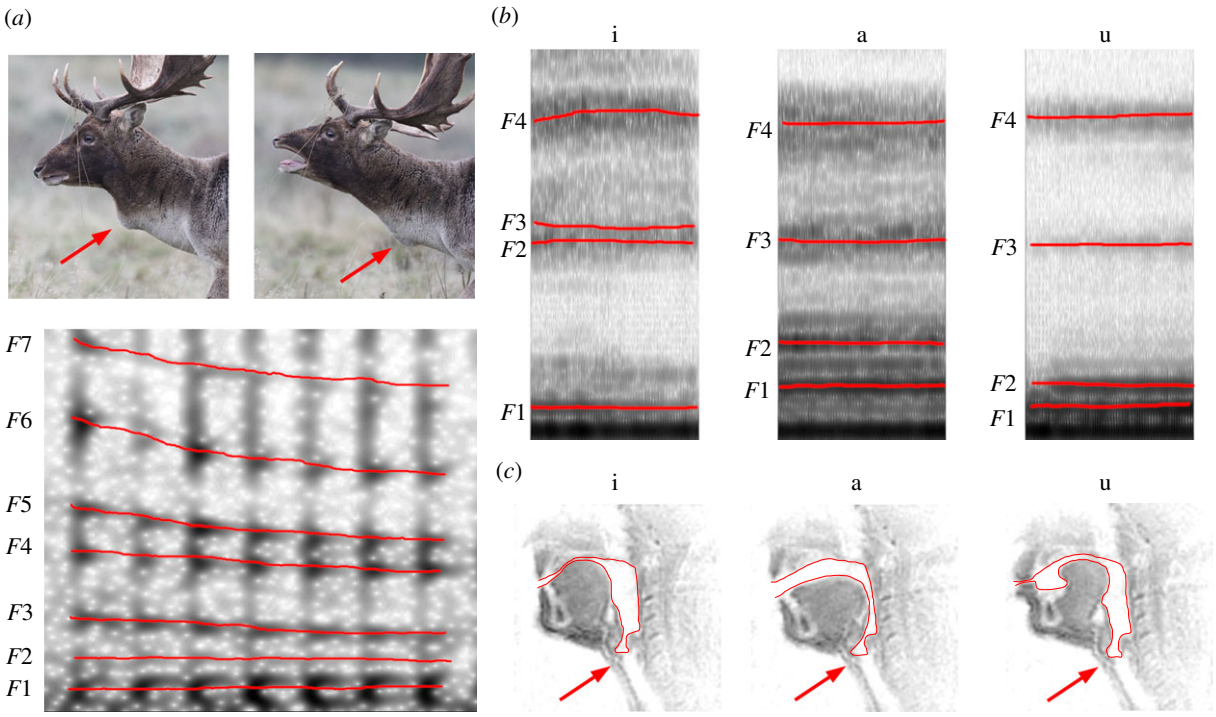

Figure 1. The effects of vocal tract elongation and vowel articulation on formant frequencies. (*a*) Scaling all formants (labelled *F*1 to *F*7) downwards via laryngeal lowering and full vocal tract elongation in the roar of a fallow deer stag. Laryngeal position is indicated with red arrows. Spectrogram range 0–3 kHz. Source: Reby *et al.* [19] re-used with permission. (*b*) Individual formant shifts in human vowels [i a u]. Observe that *F*1 and *F*2 vary across vowel sounds, whereas *F*3 and *F*4 remain relatively stable. Spectrogram range 0–5 kHz. Source: demonstration by A.A. (*c*) Sagittal MRIs of the human supralaryngeal vocal tract during the production of three cardinal vowels [i a u]. Observe that the larynx does not change its position (red arrows). Images were extracted with permission from the real-time MRI IPA chart (Span, USC [20]). Electronic supplementary material, Audio files SA1 and SA2.

pressure often operates on signallers to permanently or transiently exaggerate their perceived size, including via mechanisms that lower otherwise reliable indices of size such as formant frequencies [13]. For instance, highly elongated tracheas in birds as well as vocal sacs and descended or movable larynges in mammals are well-documented examples of anatomical adaptations that serve to enlarge the acoustic resonator [4,15], thereby making the animal sound larger than it actually is [9,13,16].

Lowering all formants to sound big, for instance by lowering the larynx to extend the vocal tract, may be a cheap trick relative to actually being big, but physically expanding the vocal resonator still imposes costs on the signaller. Permanently lengthened airways interfere with respiration [17], while repeatedly pulling down the larynx and extending the neck while vocalizing involves energetically costly postural and muscular efforts. For instance, roaring contests between male red deer are gruelling marathon events that demonstrate the stamina of individual stags, who may conserve energy (and avoid retaliation costs [18]) by reserving roars with the lowest laryngeal position and thus the lowest formant spacing for the most formidable opponents [16]. A metabolically efficient way to convey the impression of a large resonator could thus confer a significant advantage.

Here, we propose a novel hypothesis that lowering only one or two individual formants may elicit a perceptual effect similar to that of scaling all formants for size exaggeration. Lowering the larynx elongates the vocal tract and has the effect of shifting all formant frequencies down by the same proportion (figure 1*a*). This is known to make animals, including humans, sound larger [13,15]. By contrast, independent shifts in only the lower formants (particularly *F*2) are caused by the movements of supralaryngeal articulators, namely the lips, tongue and jaw. In human speech, this gives rise to different vowel sounds [21,22]. For example, relative to even formant spacing, both *F*1 and *F*2 are positioned low in [u] (as in '*boo*t'), a rounded back vowel. To produce this formant configuration and thus the 'oo' sound during speech production, the tongue is positioned high and central in the mouth, and the lips protrude into a rounded shape (figure 1*b*,*c*). By contrast, *F*2 is positioned much higher in the vowel sound [i] (as in '*bee*t'), an unrounded front vowel, compared to even formant spacing. When producing the 'ee' sound, the tongue is positioned high and frontward in the mouth, and the lips are retracted into a smile-like shape (figure 1*b*,*c*). The ability to voluntarily control the vocal articulators is a critical prerequisite for speech because it allows for individual formant shifts and non-uniform formant spacing, and in turn, different vowel sounds (figure 1*b*,*c*). Our key research question here is whether individual formant shifts, such as those observed in different vowels, can influence the perceived body size of the vocalizer.

Critically, the capacity to produce non-uniform formant configurations is phylogenetically older than human speech. Indeed, while non-human primates lack spoken language and a human-like descended larynx, their vocal tract anatomy appears adequate for producing a range of human-like speech sounds [23], with compelling evidence for the production of formant contrasts and proto-vowels in the vocal repertoires of living primate species ([24] for review). Current behavioural and neural evidence further suggests that other mammals, including non-human primates, may have more control over their vocal anatomy than was previously believed, allowing them to produce formants that are non-

uniformly spaced and/or that vary dynamically [24,25]. However, the origins and functions of such non-uniform formant configurations in non-human species that lack speech and language, including in our ancestors, remain unclear. We propose that, if individual formant shifts (i.e. lowering only one or two formants) elicit perceptions of largeness in a manner similar to full formant rescaling during vocal tract elongation, the adaptive advantage associated with such 'fast-and-frugal' articulatory manoeuvers could have provided an evolutionary route for the emergence of vocalic complexity in our ancestors.

Although vocal tract length changes only slightly across human vowel sounds (mostly due to lip rounding or retraction) and formants above *F*3 remain relatively stationary (figure 1*b*,*c*), early studies found that human listeners nevertheless perceive certain vowels as 'larger' than others. Nearly a century ago, Sapir [26] reported that 80% of human listeners guessed that *mil* referred to a small table and *mal* to a large table. Subsequent studies of sound symbolism and iconicity in language conducted in the mid-twentieth century suggested that high-*F*2 vowels [i] and [e] are often associated with small size, while [a] and low-*F*2 back vowels [u] and [o] are associated with large size [27–29]. Why are specific vowels perceptually associated with largeness and others with smallness? The answer almost certainly lies in their formant configuration. Indeed, more recent psycholinguistic experiments further suggest that listeners associate vowels that have low average formant frequencies, such as [u], with largeness [11,30], while [i] is associated with smallness [31], even by 4-month-old infants [32]. As such, human listeners, who are known to robustly associate uniformly low resonant frequencies with largeness [9,10,12,33], may also associate specific vowels that have low-positioned lower formants (such as a relatively low *F*2) with largeness, supporting a general *low is large* perceptual bias in human and animal size perception [15,33,34]. Nevertheless, previously reported differences in perceived size across vowels are unexpectedly small compared to the effect of a lower pitched voice (low f_o) or full VTL manipulations [9], and are not consistently replicated. This could be due to experimental procedures and/or the effects of vocal tract or speaker normalization, whereby listeners unconsciously adjust formant frequencies for speaker size in order to perceive the same speech sounds regardless of the VTL and thus formant spacing of the person speaking [35]. In addition, earlier studies did not systematically and directly test the causal effects of individual formant shifts in size perception.

In this study, we used state-of-the-art voice resynthesis technology to test the hypothesis that lowering only one or two individual formants will project an impression of a larger body size similar to that of full formant rescaling. We created synthesized vocal stimuli with the R package *soundgen* [36], an open-access parametric voice synthesizer that allows us to produce carefully controlled yet realistic 'human' and 'animal' vocal stimuli (stimuli and custom R code freely available online: https://osf.io/z6tuv/). Using these stimuli, we conducted three psychoacoustic playback experiments involving 511 adult human listeners:

— *Vowel-manipulation experiment* ($n = 291$), in which we test the prediction that listeners will associate vowels with relatively low lower formants (e.g. rounded vowels, back vowels with low *F*2 such as [u]) with large body size and vowels with relatively high lower formants (e.g. unrounded vowels, front vowels with high *F*2 such as [i]) with small body size;
— *Single-formant manipulation experiment* ($n = 58$), in which we systematically manipulate the relative positions of individual lower formants to further corroborate the results of the vowel-manipulation experiment, and to directly test the prediction that lowering only one or two formants will cause listeners to perceive a larger vocalizer;
— *Implicit Associations Task*, IAT ($n = 162$), in which we retest our key predictions using an established implicit behavioural paradigm in order to confirm the robustness and ecological validity of our observations.

2. Results

All data analyses were performed in R v. 4.0.2 using Bayesian mixed models that were written either directly in *Stan* [37] or via the *brms* package [38]. In all cases, we modelled individual trial-level responses, without aggregating any data, and specified moderately informative regularizing priors on regression coefficients so as to reduce overfitting and improve convergence. All reported estimates are medians of the posterior distribution with Bayesian 95% credible intervals (CIs) [39]. See Methods for a full description of each experiment including sound synthesis, participant details, playback procedures and statistical analyses.

To first test whether listeners systematically associate vowel sounds with perceived size (vowel-manipulation experiment), we synthesized the human vowels [u ɑ a ɛ i] and the central vowel schwa [ə], using *soundgen* [36] (*human* condition). These six vowels represent the extreme positions in the *F*1–*F*2 vowel space (figure 2*a*). In order to attenuate the expected adverse effect of the speaker or vocal tract normalization on size perception in human vowels, we also synthesized vocalizations resembling animal calls in which formants followed patterns corresponding to the same six vowels (*animal* condition) (electronic supplementary material, audio SA3–SA6). In psychoacoustic playback experiments using these synthesized vowels, we then asked listeners to judge which of two human speakers (human condition, $n = 92$ listeners) or two animals (animal condition, presented along with synthetic mammal vocalizations, $n = 198$ different listeners) sounded larger, or to indicate no perceivable difference. All 15 possible pairs of six vowels were tested in both the human and animal conditions (see Methods for full experimental protocols). We then fitted two ordinal logistic regression Bayesian mixed models in R v. 4.0.2 [40] to estimate the most credible ranking of vowels by size (figure 2*b*,*c*) and the differences in perceived size between vowel pairs (figure 2*d*; see Methods for further details).

In both the human and animal conditions, the synthesized vowel [u] as in 'boot' consistently conveyed the largest size. Although formant positions of each vowel were identical in both conditions, rankings of other vowels by the perceived size of the vocalizer differed across conditions (figure 2*b*,*c*), suggesting that the perceived source of the vocalization (human versus animal) may have modulated the association between formant patterns and size. Contrary to the classic *mil-mal* study [26], [a] was not perceived as larger than [i] in either the human ([i] > [a] by 0.3%, 95% Bayesian CI −22, 24) or animal ([i] > [a] by 5.2%, CI −23, 59) condition, with large individual variation in responses.

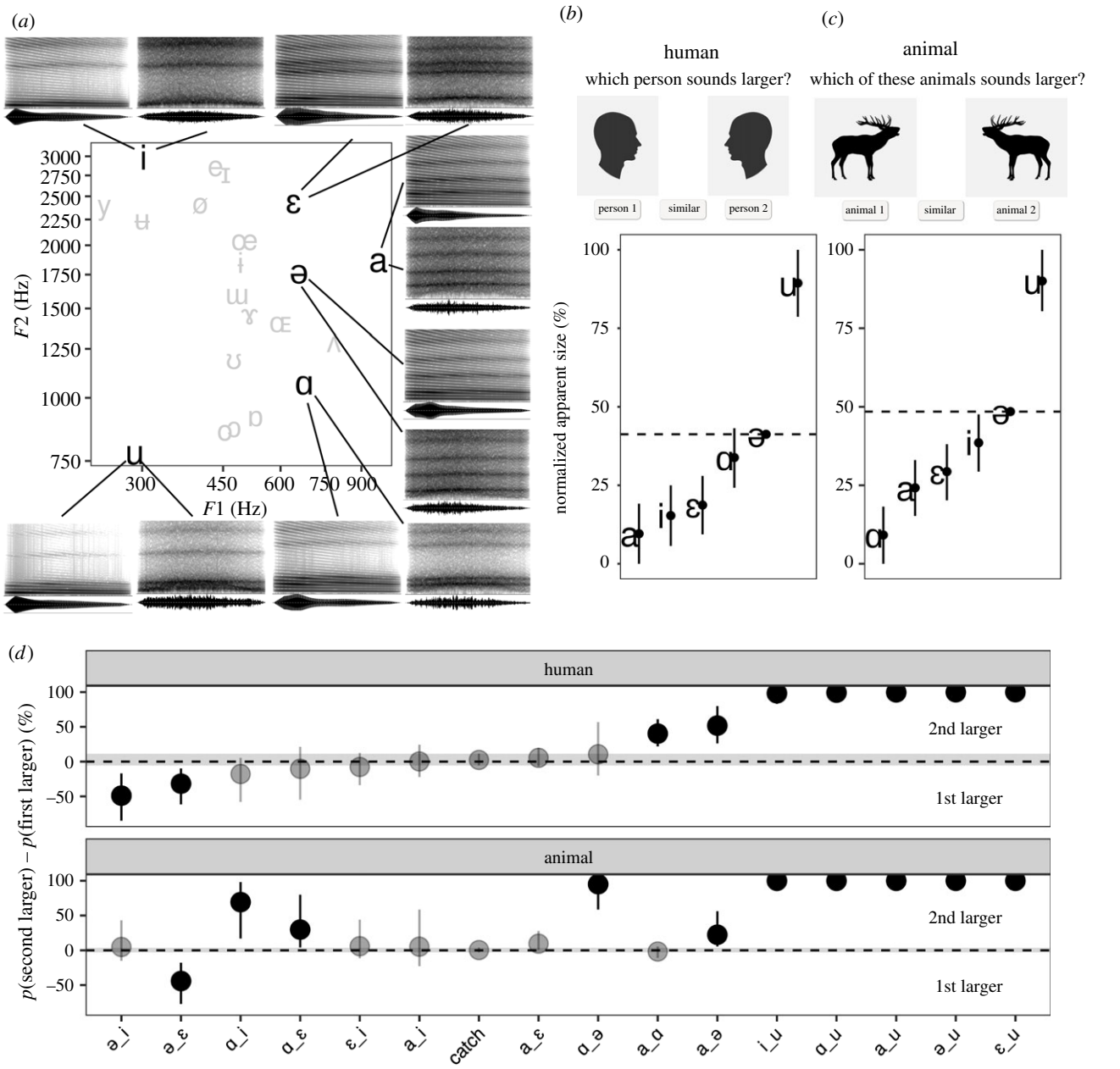

Figure 2. Vowels and perceived body size. (*a*) Relative positions of six synthesized vowels in *F*1–*F*2 space. Spectrograms show human and animal versions of each vowel, with formants *F*1 to *F*4 visible as dark bands of high spectral energy. (*b*,*c*) The ranking of vowels by size. In both human and animal conditions, vowel [u] was associated with the largest size. The choice between two stimuli was assumed to reflect the distance between them on a latent 'largeness' variable, mapped onto responses via ordinal logistic regression. The model was written in *Stan* [37]. (*d*) Fitted value of the probability of perceiving the second vowel in a pair as larger minus the probability of perceiving it as smaller (ignoring ties). Ordinal logistic regression with subject-specific slopes fitted with R package *brms* [38]. The greyed-out points have CIs that fail to clear the Region of Practical Equivalence (ROPE) corresponding to the CI for catch trials with two identical sounds. Electronic supplementary material, audio files SA3–SA6.

In the single-formant manipulation experiment (see Methods), we then tested our key prediction that experimentally lowering only one or two individual formants would cause a vocalizer to sound physically larger. Synthetic vocalizations were created in *soundgen* [36] with an intonation contour and voice quality patterned to resemble the call of a large mammal (electronic supplementary material, audio SA7–SA9). Like the vowel-manipulation experiment, we used animal-like calls to increase ecological validity and importantly, to reduce the probability that listeners would associate any of the formant manipulated stimuli with a specific vowel sound, thus reducing any potential effect of speaker or vocal tract normalization on size perception. Vocalizations were synthesized at low, medium and high levels of pitch, f_o (figure 3*a*), respectively, appropriate for a large mammal such as a red deer stag (mean $f_o = 94$ Hz, range 65–120 Hz), a human male (141 Hz, 97–180 Hz) or a human female (235 Hz, 162–300 Hz). This allowed us to test and control for any interaction effects between pitch and formant manipulations. The schwa vowel [ə] contained equidistant formants corresponding to a vocal tract shaped as a closed-open tube 16 cm long (human-range VTL), 28 cm long (intermediate VTL) or 40 cm long (deer-range VTL). We then modified each of nine schwa prototypes (3 VTL × 3 f_o levels) by shifting one of three formants (*F*1, *F*2 or *F*3) or both *F*2 and *F*3 either up or down by 3.15 semitones (figure 3*b*). Finally, we scaled all formants simultaneously to simulate a VTL change known to affect perceived size in

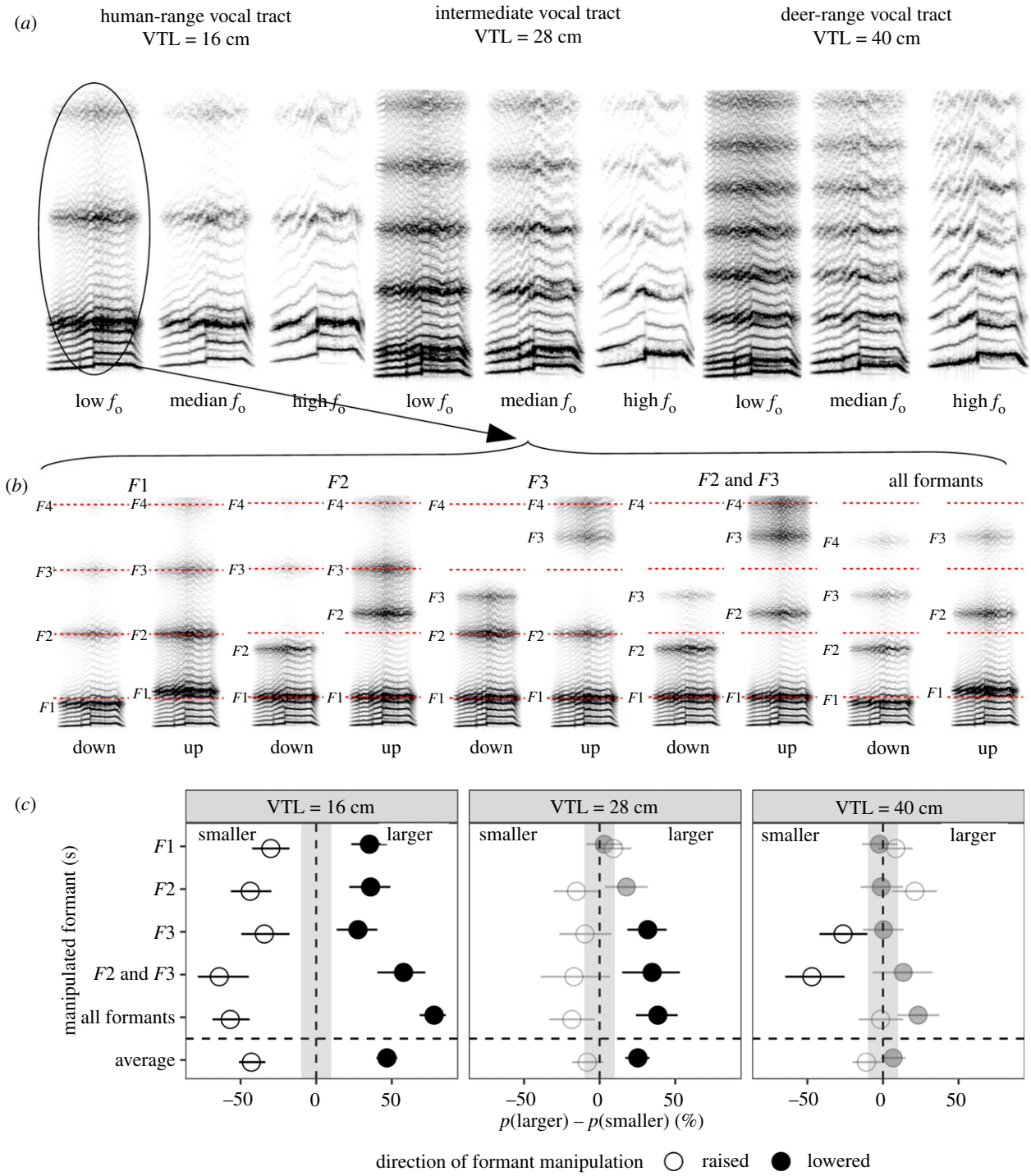

Figure 3. Shifting individual formants alters the perceived body size of the vocalizer. (*a*) Synthetic calls patterned after the call of a large mammal with uniformly spaced formants (schwa) at three levels of VTL and f_o. (*b*) Manipulations of one schwa configuration (VTL = 16 cm, low f_o), with the original schwa formants shown by dotted lines. Frequency range 0–3 kHz, duration 1500 ms. (*c*) The effect of formant manipulation on perceived size: fitted values from ordinal logistic regression. Markers indicate the probability of perceiving the manipulated vocalization as larger than schwa minus the probability of perceiving it as smaller (median of posterior distribution and 95% CI), ignoring ties and averaging across three f_o levels. The greyed-out markers have CIs that fail to clear the ROPE of ±10%. Electronic supplementary material, audio files SA7–SA9.

humans and non-human listeners [5,7]. In this experiment, human listeners ($N = 58$) were asked to indicate which of two animals sounded larger, comparing a schwa vowel [ə] that contained equally spaced formants to a manipulated version of that same vowel, in which one, two, or all formants were re-scaled. Both vocal stimuli in each pair were matched on VTL and pitch (f_o), wherein we tested all three VTLs and all three f_o levels, for a total of 90 unique stimulus pairs (see Methods). As in the vowel experiment, we ran Bayesian mixed models (ordinal logistic regression) in R [40] to estimate the most credible effect of each manipulation on perceived body size (see Methods).

Supporting our prediction, lowering one or more individual formants increased the perceived size of a vocalizer, whereas raising individual formants decreased perceived size for vocalizations representing a human-range VTL of 16 cm (figure 3*c*; where 100% = always perceived as larger than schwa, −100% = always smaller, 0% = random response).

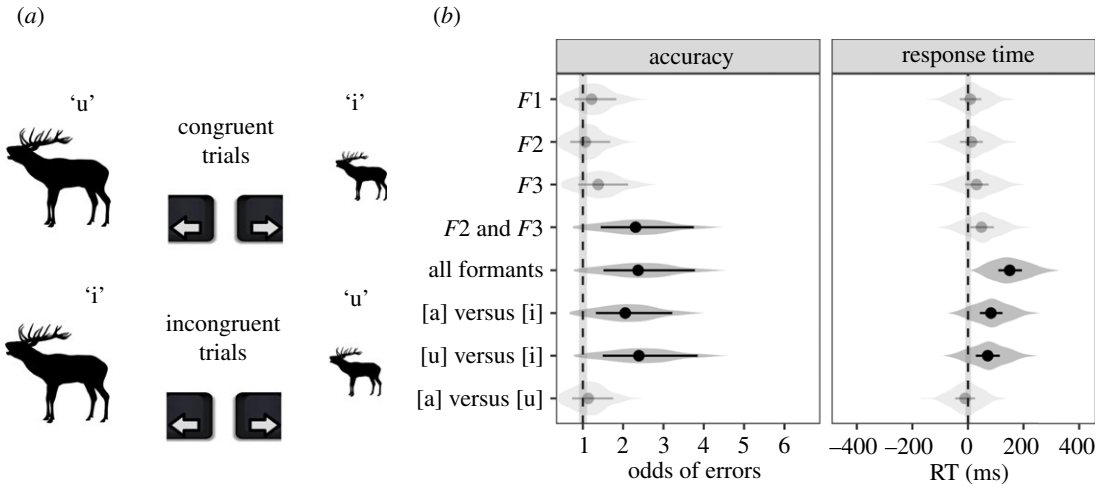

Figure 4. Implicit Associations Task. (*a*) Participants pressed two keys in response to two visual and two auditory cues. Accuracy and response times (RT, in ms) were compared in blocks in which two congruent or incongruent stimuli were assigned to the same key. Eight IAT experiments were performed (16 blocks of 16 trials, 20 listeners each, $N = 162$), including the [u]-[i] comparison shown above as an example. (*b*) We used logistic regression to predict accuracy and a lognormal model to predict RT in correct trials, both as a function of Congruence × Experiment, with a subject-specific congruence effect and a stimulus-specific intercept as random effects. Solid markers show the odds ratio of errors (left) and the difference in RT in ms (right) for incongruent versus congruent pairs (medians of posterior distributions and 95% CI); violin plots show the distribution of fitted values per listener. Greyed-out effects failed to clear a ROPE of ±10% (for accuracy) or ±10 ms (for RT). See main results for the direction of congruence effects.

On average, the effect of manipulating only one of the first three formants was half as strong as the effect of scaling all formants (e.g. 35%, CI 23, 47 for lowering *F*1, compared to 78%, CI 68, 86 for lowering all formants). Critically, however, the effect of lowering *F*2 and *F*3 simultaneously (58%, CI 40, 72) was comparable in magnitude to rescaling all formants.

While lowering all formants or lowering *F*2 and *F*3 simultaneously conveyed a large size at VTLs of 16 cm and 28 cm (effect size averaging across all manipulations 25%, CI 13, 33), the effects of individual formant manipulations on perceived size were attenuated or eliminated when VTL was increased beyond the human anatomical range (VTL = 28 or 40 cm; figure 3*c*). Indeed, at a VTL of 40 cm, even rescaling all formants downwards (7%, CI −1, 15) or upwards (11%, CI −1, 20) had little to no effect on the perceived size, suggesting that the well-established effect of apparent VTL on size perception in humans may not generalize to VTLs well outside the human range. The effects of formant manipulations were qualitatively similar across the three tested f_o levels, and including an interaction between manipulation and f_o failed to improve the model's predictive power (difference in the LOOIC information criterion = −4.3 in favour of the simple model, standard error = 10.0).

Finally, to ensure the robustness of these effects, we retested our key manipulations using an Implicit Associations Task (IAT) ($N = 162$; figure 4*a*). This test is designed to reveal perceptual associations that participants may wish to conceal or that may operate below the level of conscious awareness, including cross-modal associations [31,41]. Participants in the IAT were thus not required to make explicit size judgements. Instead, they were trained to press different keys in response to pairs of visual and auditory cues that were either congruent (e.g. icon of a large animal paired with [u]) or incongruent (e.g. icon of a large animal paired with [i]; see Methods and figure 4*a*). Using this implicit timed-response test, we would only expect to find congruence effects if listeners naturally and involuntarily associate specific sound stimuli with large or small size. As predicted, both accuracy and response times (RT) (figure 4*b*) showed a congruence effect for scaling all formants (odds ratio (OR) of errors = 2.3, Bayesian 95% CI 1.5, 3.5; difference in RT (dRT) = 146 ms, CI 108, 188) and for shifting *F*2 and *F*3 simultaneously (OR = 2.3, CI 1.4, 3.6; dRT = 47 ms, CI 8, 88), although not for shifting only one formant at a time (*F*1, *F*2 or *F*3). Thus, lowering all formants simultaneously, or lowering only *F*2 and *F*3 in unison, caused listeners to judge vocal stimuli as larger (figure 4*b*). For vowel comparisons, [u] was perceived as larger than [i] (OR = 2.4, CI 1.5, 3.8; dRT = 69 ms, CI 28, 113), but not larger than [a] (OR = 1.1, CI 0.7, 1.7; dRT = −9.5 ms, CI -45, 26). Interestingly, in accordance with *mil-mal* studies but contrary to the results of our explicit pairwise comparisons (figure 2), [a] was perceived as larger than [i] in the IAT (OR = 2.0, CI 1.3, 3.1; dRT = 83 ms, CI 46, 121; figure 4*b*). The IAT results thus confirm that acoustic changes corresponding merely to articulatory movements, without modifications to voice pitch or apparent VTL, can give rise to implicit associations with body size.

3. Discussion

Taken together, these experiments demonstrate that lowering only one or two formant frequencies in synthesized animal calls and human speech is sufficient to achieve the impression of a larger body size. The effect of simultaneously lowering the second and third formant frequencies of the voice, *F*2 and *F*3 (achieved by small and rapid articulatory movements of the tongue, jaw and lips), is indeed comparable in magnitude to rescaling all formants simultaneously, which typically requires slower and more effortful elongation of the entire vocal tract. Our results thus support the prediction that individual formant shifts can constitute a fast-and-frugal mechanism for acoustic size exaggeration.

The pressure to sound large is extremely common across the animal kingdom [3,18], eliciting a range of anatomical and/or behavioural adaptations [15] such as a descended or mobile larynx now documented in dozens of mammalian

species [4,15]. Several species also have highly specialized vocal organs, such as the enlarged laryngeal hyoid bones of howler monkeys that contain an air sac thought to enable them to produce some of the loudest calls of any terrestrial animal [42], or the additional set of velar vocal folds in koalas thought to enable them to produce vocalizations that are twenty times lower in pitch than expected for the animal's body size [43]. While spoken language is uniquely human, revolutionary advances in animal communication research are beginning to reveal surprisingly complex articulatory and vocal control abilities in non-human primates [24,25,44], including contrasting formant patterns [24]. Such vocal control abilities could, as suggested by the preliminary results of our study, have also originated as a means to exaggerate body size. Indeed our results show that formant contrasts can attenuate or exaggerate apparent body size, which may play an important role in aggressive interactions to communicate physical formidability and motivation (e.g. appeasement versus threat; [34,45]), particularly in the context of intrasexual competition between males in polygynous species.

Curiously, at some point in the evolution of hominins, our ancestors lost the air sacs still found in other great apes. In addition to ostensibly increasing call intensity and loudness [46], air sacs create additional low-frequency formants [47] and have thus been hypothesized to function to increase apparent body size [44]. Selection pressure to sound big after the loss of air sacs may have led to alternative vocal adaptations for size exaggeration, such as the descended larynx found in anatomically modern humans [44]. In parallel to a descended larynx, however, hominins acquired unusually fine neuromotor control over vocal production [48], without which humans would not be able to volitionally produce different vowel sounds. The conditions that enabled our ancestors to achieve this neurological control, and ultimately to acquire articulated speech, remain hotly debated [6,25,44,48]. Our results suggest that basic articulatory gestures supporting the manipulation of lower formants for size exaggeration may have been a contributing factor.

Interestingly, our results also show that the back vowel [u], produced by protruding and rounding the lips, most effectively exaggerates perceived body size. Lip protrusion effectively extends the vocal tract at the mouth end, whereas lip retraction, as observed in the smile-like lip configuration of the vowel [i], effectively shortens the vocal tract. It is thus possible that protruding the lips, as many animals do during call production [15,42], may constitute yet another fast-and-frugal 'trick' for size exaggeration, potentially also leading to the diversification of articulatory gestures during vocal production.

While our results confirm that the vowel [u], with an extremely low second formant *F*2, consistently exaggerates the perceived size of a vocalizer, the size-exaggerating effect of [a] was less consistent. Nevertheless, the fact that [a], a vowel with high *F*1 and neutral *F*2, still sounds relatively large, suggests that articulatory size exaggeration may involve a trade-off between lowering formant frequencies and maintaining a loud voice. Sonorous open vowels such as [a] may be more effective for size exaggeration than the closed vowel [u] under conditions when loud calling is critical, potentially explaining the mixed results of past studies in sound symbolism [11,27–29,32]. More research is needed on the physiological constraints of vowel-like vocalizations and their interplay with the cognitive processes responsible for sound-size associations in receivers.

Signallers can gain substantial benefits from exaggerating their size; however, listeners are also under evolutionary pressure to detect deceptive signals [13,18,49]. Because anatomical constraints typically ensure an allometric relationship between body size and VTL, formant frequency spacing generally provides a reliable and honest index of body size [5–7], even where anatomical innovations allow for permanent and/or behavioural vocal tract extension [4,13,15]. Moreover, human and non-human listeners have been shown to use formant frequencies to infer relative body size in exaggerated signals, suggesting that they adjust size judgements to the shifted baseline [13,16,50]. If individual formant shifts are efficient in exaggerating body size, as our results indicate, then one would expect selection on receivers to detect these articulatory manoeuvers in order to retrieve unbiased size information. Indeed, while our results confirm that listeners associate individual formant shifts with changes in physical size, we also show that listeners partly compensate for these perceptual biases when exposed to familiar vowels (human condition) rather than to formant configurations that do not occur in speech (animal condition). We suggest that such perceptual mechanisms, equating to 'articulatory normalization', could constitute a precursor for vowel perception. Our results also show that globally or individually lowered formants convey a larger body size regardless of the underlying pitch (f_o) of the vocal stimulus, which here was manipulated to reflect either a large mammal such as a red deer stag, a human male or a human female. This further corroborates the potential robustness of formant contrasts for size exaggeration.

Phylogenetically controlled investigations of the production and perception of vowel-like vocalizations in non-human animals are now needed. Recent phylogenetic studies have demonstrated that non-human mammals can produce and distinguish among calls with non-uniform formant distributions [24], regardless of whether these vocalizers possess a human-like descended larynx [23,51]. It remains to be seen whether the individual formant size exaggeration we report here is present in the communication systems of non-human animals. Indeed, if the *low is large* perceptual bias contributed to the evolution of vocalic complexity, we would also expect to find [u]-like proto-vowels in aggressive animal vocalizations, and [i]-like proto-vowels in submissive vocalizations, particularly in the context of male competition in sexually dimorphic species and during social interactions between conspecifics or groups in which dominance is established through acoustic communication [18].

4. Methods

(a) Summary of experimental materials and methods

We created synthesized vocal stimuli with the R package *soundgen* [36], an open-access parametric voice synthesizer that creates realistic yet highly controlled voice stimuli by synthesizing a mixed harmonic-noise excitation source filtered with a vocal tract transfer function based on manually provided global contours of control parameters such as f_o, individual formant frequencies, amplitude and other acoustic features [36,52]. The synthetic stimuli were patterned after actual voice recordings, and all pairs of synthesized stimuli in both vowel and single-formant experiments were fully matched for duration, root mean square amplitude, intonation and voice quality, differing only in the relative positions of formant frequencies, as described below for each experiment. The *soundgen* [36] custom R code for

creating all acoustic stimuli is freely available online in OSF (https://osf.io/z6tuv/). In total, 533 adult listeners took part in one of three psychoacoustic playback experiments: vowel-manipulation experiment ($n = 291$), single-formant manipulation experiment ($n = 58$), and the IAT ($n = 162$), as further described below. Data from 511 participants were retained for analysis (see exclusion criteria below). Sample sizes were determined to achieve adequate precision of effect sizes and narrow CIs in Bayesian models [39]. Listeners who took part in the vowel experiments and the IAT were recruited via *Prolific* (https://www.prolific.co/), a well-known online participant recruitment platform, and were paid 7.50 GBP per hour for their participation. Participants taking part in the single-formant manipulation experiment, which required high-quality audio playback equipment (see single-formant manipulation experiments below), were tested in person at the ENES lab (Equipe de Neuro-Ethologie Sensorielle) in the University of Saint-Etienne, France, or recruited via personal contacts. In all experiments, the inclusion criteria were fluency in English and the absence of self-reported hearing difficulties.

(b) Vowel-manipulation experiment

(i) Sound synthesis

Vowel stimuli consisted of six human and six animal synthetic vowel sounds: /i ɛ a ə u ɑ/ (International Phonetic Alphabet, IPA [21,22]). We manually measured the first six formant values in recordings of the corresponding vowels pronounced by the same female phonetician, obtained from *Seeing Speech* [53], and then synthesized sounds with the same formants using *soundgen* [36]. The first six formant frequencies were synthesized based on measurements obtained from the original vowels, and higher formants up to Nyquist frequency were added at approximately equal spaces calculated from the VTL, which was estimated from the lower formants. The advantage of creating fully synthetic stimuli was the ability to manipulate individual formant frequencies, while ensuring that all other acoustic characteristics remained constant; it also enabled us to create vocalizations that resembled animal calls with human-like formants (animal condition). However, to avoid unnaturally static stimuli, the target formant values varied slightly in the course of each vocalization, with parallel formant transitions around the target values created with the *mouth* parameter in *soundgen* [36]. Each synthetic vowel was 1200 ms long and had exactly the same intonation contour and voice quality in all six vowels of each type (human or animal). The human vowels were synthesized to sound human-like (e.g. electronic supplementary material, audio SA3 and SA4), while the animal vowels were synthesized to sound animal-like, with strong jitter and shimmer (rapid random variations of f_o and amplitude, respectively [54]), and intonation rapidly rising to 650 Hz and then falling to 210 Hz (e.g. electronic supplementary material, audio SA5 and SA6).[1] Despite having a vowel-like formant structure, the animal stimuli were thus engineered to resemble animal calls rather than human vowels, reducing the potential for vocal tract normalization by listeners [35].

(ii) Participants

Using the online platform *Prolific*, we recruited 98 adult listeners to rate the human vowels and 215 listeners to rate the animal vowels. Exclusion criteria were completing fewer than 10 trials or missing half or more of catch trials with two identical sounds (i.e. failing to rate them as similar). Based on these criteria, data from 6 of 98 participants were excluded from analyses in the human condition and data from 16 of 215 participants were excluded from analyses in the animal condition, for final samples of 92 and 199 listeners, respectively. Hence, each pair of vowels was rated at least 90 times.

(iii) Psychoacoustic playback

On each trial, participants were presented with two acoustic stimuli and instructed to indicate which of the two sounds was produced by a larger person/animal, or otherwise to indicate no perceivable difference. All 15 possible pairs of six vowels /i ɛ a ə u ɑ/ were tested. Participants heard either human or animal vowel stimuli, judged in two separate experiments by two independent samples of participants. The human vowels were presented among other manipulated versions of the same sounds (with all formants scaled by three semitones and f_o shifted by 1 semitone, always in incongruent directions), which were included as filler stimuli and intended to make the vowel contrasts less conspicuous. Participants clicked images of human faces to hear each acoustic stimulus and were asked to judge *Which person sounds larger?* By contrast, animal vowels were presented among mammal-like vocalizations analogous to those used for testing single-formant manipulations and here used as fillers. Participants clicked images of deer to hear the sounds and were asked *Which of these animals sounds larger?* This experimental set-up was designed to reinforce the impression that the vocalizations were produced by either humans or non-human animals, despite an identical formant structure.

(iv) Statistical analysis

For the purpose of ranking vowels by the perceived size of the vocalizer, we assumed that the choice between two stimuli reflected the distance between them on a latent size scale, so that one sound was chosen if it exceeded the other by some threshold, which could vary across participants. The latent size variable was mapped onto responses via ordered logistic regression. The corresponding Bayesian model was written in *Stan* [37]. The latent size variable was underspecified; to ensure convergence, the position of the first sound was therefore fixed at zero, and the overall scale was set by the normal prior with an arbitrarily chosen standard deviation. In addition, a standard ordinal logistic regression model was fit in *brms* [38] to estimate the most credible differences in perceived size between all possible pairs of vowels. The model was of the form $size \sim pair + (pair \mid subject)$, where *size* was encoded as '1' (first vowel judged as smaller), '2' (no difference) or '3' (first vowel judged as larger), and size preferences were allowed to vary across subjects. Both ranking and pairwise comparisons were modelled separately for human and animal vowel stimuli.

(c) Single-formant manipulation experiment

(i) Sound synthesis

The stimuli were synthetic vocalizations 1500 ms in duration, whose intonation contour and voice quality were patterned to resemble the vocalization of a large mammal such as a red deer stag (listen to electronic supplementary material, audio SA7–SA9). The schwa vowel stimulus [ə] contained equally spaced formants predicted for a closed-open vocal tract shaped like a regular tube 16, 28 or 40 cm long. The vocalizations were synthesized at three f_o levels: low (mean $f_o = 94$ Hz, range 65–120 Hz), median (141 Hz, 97–180 Hz) and high (235 Hz, 162–300 Hz). In addition to the schwa versions [ə], we modified the vocalizations by shifting one of the first three formants (*F*1, *F*2 or *F*3) up or down by 3.15 semitones (approximately 20%), shifting *F*2 and *F*3 up or down together, or scaling all formants simultaneously by the same amount. This produced 99 unique stimuli (3 VTL levels × 3 f_o levels × 11 manipulations). The magnitude of formant manipulations (3.15 semitones or 20%) was large enough to be clearly audible in all stimuli, exceeding the just-noticeable difference [9], but not so large as to fuse one formant with a neighbouring formant.

(ii) Participants

Similar to the vowel psychoacoustic experiments, we initially pilot-tested the single-formant experiment with an online sample recruited on *Prolific*. A comparison of the results with in-laboratory pilot-testing indicated that low-quality audio playback was interfering with the perception of low-frequency components, which in this experimental condition would be problematic for manipulated formant frequencies under 300 Hz. We therefore conducted the experiment ensuring that all participants used high-quality professional headphones with broad frequency resolution (14 to 20 kHz). We tested 29 participants in the laboratory (ENES, University of Saint-Etienne), and an additional 29 participants were recruited via personal contacts to conduct the study from home with the verified use of professional headphones (these results were qualitatively comparable to those obtained from in-laboratory testing). With this sample of 58 participants, each pair of stimuli was rated on average 48 times, providing sufficiently high precision of estimates for planned Bayesian modelling [39].

(iii) Psychoacoustic playback

On each trial, participants were presented with two acoustic stimuli and instructed to indicate which of the two sounds was produced by a larger animal, with an option to rate both sounds as similar (indicating no perceivable difference). In each stimulus pair, one of the vocalizations was a neutral schwa version [ə] containing the original equally spaced formants, and the other vocalization was manipulated to contain one or more shifted formants. Each participant completed 46 experimental trials and four catch trials with a pair of identical sounds, which were included as attention checks. To play the vocalizations, participants clicked on two same-sized icons showing deer profiles (figure 2*c*); deer icons were used to encourage participants to imagine animal calls rather than a human voice or synthetic sounds. The experiment was written in html/javascript and conducted in a web browser. The participants tested in the laboratory ($N = 29$) were provided with Sennheiser HD 205 II professional headphones to ensure optimal sound quality (frequency response 14–20 kHz, less than 0.5% THD, total harmonic distortion).

(iv) Statistical analysis

Statistical modelling for single-formant manipulations was similar to that applied to pairwise vowel comparisons, with the exception that here, the mid-central vowel schwa [ə] always constituted one stimulus, paired with a manipulated vocalization. The outcome variable was therefore encoded as '1' if the manipulated vocalization was judged to be smaller than schwa [ə], '2' if there was no difference and '3' if the manipulated vocalization was judged to be larger than schwa [ə]. This outcome variable was modelled with ordinal logistic regression. The simpler model without including f_o level was of the form:

$$\text{size} \sim \text{manipulation} * \text{VTL} + \text{manipPos} + (\text{manipulation} \mid \text{subject}),$$

where manipulation was a factor with 11 levels (catch trial with two identical sounds, one formant *F*1–*F*3 shifted up or down, *F*2 and *F*3 shifted up or down together, or all formants shifted up or down), and VTL was a factor with three levels (16, 28 or 40 cm). The model with f_o level included an additional triple interaction with f_o; the results were similar, but the simpler model was preferred by approximate leave-one-out cross-validation or LOOIC (elpd difference = 4.3, s.e. = 10.0). Because there was a marked tendency to choose the stimulus on the right-hand side as larger, potentially due to a right-large perceptual bias [55], we also corrected for stimulus position (*manipPos*). To allow for the nested nature of data and possible differences among participants, we included subject-specific random slopes for the effect of manipulation per subject. The models were fitted using *brms* [38].

(d) Implicit associations task

(i) Sound stimuli

The IAT [31,41] requires a much greater amount of data compared to explicit rating paradigms because response biases are deduced from performance, rather than explicitly stated by the participant. Indeed, this is the key strength of the IAT [31]. However, as a result of the increased number of trials per participant, we could not feasibly retest all stimulus combinations and focused instead on replicating key results from the vowel and single-formant experiments: vowels [u], [i] and [a] (chosen to represent the most extreme locations in vowel space), and those vocalizations patterned to reflect the intonation and voice quality of a large mammal but with a human-range VTL (VTL = 16 cm, high f_o level) and with either raised or lowered *F*1, *F*2, *F*3, both *F*2 and *F*3, or all formants fully rescaled. Participants were trained to press different keys on the keyboard or screen (depending on the device used) in response to pairs of visual and auditory cues, which could be congruent (e.g. the picture of a large deer paired with [u]) or incongruent (e.g. large deer paired with [i]). The rule specifying which two stimuli were assigned to the left and which to the right key (see §4d(iii) below) changed in every block, and we tested whether participants answered more accurately and rapidly in blocks in which two congruent stimuli were assigned to the same key. To make the stimuli easier to distinguish and to set a fast pace, as required by IAT [31,41], we resynthesized all stimuli with a shorter duration of 500 ms and compared stimuli with raised versus lowered formants against one other, rather than against the schwa vowel.

(ii) Participants

Following previous work [41], we recruited 20 participants for each of 10 experiment blocks, for a total of 162 participants (two extra submissions were received and included). Participants were recruited via *Prolific* and completed the IAT online (following [31,41]), wherein they needed to achieve an accuracy score of 75% or higher to demonstrate their understanding and compliance with the testing procedure [41]. All participants exceeded this target accuracy and therefore data from all participants were included in statistical analyses.

(iii) Psychoacoustic playback

We implemented a web-based version of the IAT (described by [31,41]). Pairs of acoustic stimuli varied across experiment blocks (e.g. [i] versus [u], schwa with original formants versus schwa with lowered *F*2, etc.); however, visual icons of two deer profiles differed only in size (the large deer icon was twice the size of the small icon in terms of linear dimensions, approximately 300×300 versus 150×150 pixels, and thus four times larger in total surface area). Listeners were required to learn a rule associating the left arrow on a keyboard or touchscreen with one image and sound, and the right arrow with another image and sound. For example, in one block of trials, the image of a small deer and the vowel sound [u] might be assigned to the left arrow key, and the large deer and vowel [i] to the right arrow key (figure 4*a*). In the next block, the rule would change, and all four possible combinations would recur in random order in multiple blocks throughout the experiment. Participants first performed two rounds of practice trials as many times as necessary (typically just once) to reach the target accuracy of 75%. Once the participant had understood the procedure and achieved an accuracy of 75% or better, they proceeded to complete 16 test blocks of 16 trials each. As each trial began, a

fixation cross was shown in the middle of the browser screen for a random period of 500–600 ms. After a delay of 300–400 ms, the stimuli were presented. Visual stimuli were shown for 400 ms in the same location as the fixation cross against a uniform white background; synthesized sounds were 500 ms in duration. If the response of the listener was correct, the next trial began immediately. If the response was incorrect, a red warning cross was flashed for 500 ms before proceeding to the next trial [31].

(iv) Statistical analysis

All training trials were discarded and only test trials were analysed. A single model was fitted to this unaggregated dataset to analyse accuracy across all 10 experiment blocks ($N = 41\,270$ trials), and another to analyse response times (RT) in trials with correct responses and RT under 5 s ($N = 39\,259$ trials). The models were as follows, in *brms*/*lme4* syntax:

Accuracy (logistic): correct ~ experimentBlock * congruent + (congruent | subject) + (1 | target)

RT (lognormal): responseTime ~ experimentBlock * congruent + (congruent | subject) + (1 | target).

The random intercept per target primarily captured the variance in accuracy or RT depending on the modality of the stimulus (e.g. responses to visual stimuli were faster than to acoustic stimuli). The random intercept per participant was included to account for individual differences in both accuracy and RT (taking into account the use of keyboard versus touchscreen). Finally, congruence effects were allowed to vary across participants. The models were fitted using *brms* [38].

Ethics. All participants provided informed consent. Ethical approval for performing perceptual experiments with human subjects was provided by the Comité d'Ethique du CHU de Saint-Etienne (IRBN692019/CHUSTE).

Data accessibility. All datasets, R scripts for data analysis, audio stimuli, R scripts for generating synthesized stimuli and html code for running psychoacoustic experiments can be downloaded from the Open Science Framework (https://osf.io/z6tuv/, doi:10.17605/OSF.IO/Z6TUV). These electronic supplementary materials enable full validation and replication of results.

Authors' contributions. D.R., A.A. and K.P. conceived and designed the studies. A.A. created the vocal stimuli, programmed the experiments, collected the data and performed the statistical analyses. K.P., A.A., and D.R. wrote the manuscript, and all authors approved the final version.

Competing interests. The authors declare no competing interests.

Funding. K.P., A.A. and D.R. were supported by the University of Lyon IDEXLYON project as part of the 'Programme Investissements d'Avenir' (grant no. ANR-16-IDEX-0005) to D.R.

Endnote

[1]Note: in the supplementary online materials (scripts, html code, wav files: https://osf.io/z6tuv/) the labels given to these two conditions are *natural* (for *human*) and *nonverbal* (for *animal*).

References

1. Fant G. 1960 *Acoustic theory of speech production*. The Hague, The Netherlands: Mouton.
2. Taylor AM, Reby D. 2010 The contribution of source-filter theory to mammal vocal communication research: advances in vocal communication research. *J. Zool.* **280**, 221–236. (doi:10.1111/j.1469-7998.2009.00661.x)
3. Charlton BD, Pisanski K, Raine J, Reby D. 2020 Coding of static information in terrestrial mammal vocal signals. In *Animal signals and communication* (eds T Aubin, N Mathevon), pp. 115–136. Berlin, Germany: Springer Nature.
4. Fitch WT, Hauser MD. 2003 Unpacking 'honesty': vertebrate vocal production and the evolution of acoustic signals. In *Acoustic communication*, Springer Handbook of Auditory Research 16 (eds AM Simmons, RR Fay, AN Popper), pp. 65–137. New York, NY: Springer. http://link.springer.com/chapter/10.1007/0-387-22762-8_3.
5. Pisanski K *et al.* 2014 Vocal indicators of body size in men and women: a meta-analysis. *Anim. Behav.* **95**, 89–99. (doi:10.1016/j.anbehav.2014.06.011)
6. Fitch WT. 2000 The evolution of speech: a comparative review. *Trends Cogn. Sci.* **4**, 258–267. (doi:10.1016/S1364-6613(00)01494-7)
7. Reby D, McComb K. 2003 Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Anim. Behav.* **65**, 519–530. (doi:10.1006/anbe.2003.2078)
8. Fitch WT, Giedd J. 1999 Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *J. Acoust. Soc. Am.* **106**, 1511–1522. (doi:10.1121/1.427148)
9. Pisanski K, Rendall D. 2011 The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *J. Acoust. Soc. Am.* **129**, 2201–2212. (doi:10.1121/1.3552866)
10. Smith DRR, Patterson RD, Turner R, Kawahara H, Irino T. 2005 The processing and perception of size information in speech sounds. *J. Acoust. Soc. Am.* **117**, 305–318. (doi:10.1121/1.1828637)
11. Barreda S. 2016 Investigating the use of formant frequencies in listener judgments of speaker size. *J. Phonet.* **55**, 1–18. (doi:10.1016/j.wocn.2015.11.004)
12. Charlton BD, Taylor AM, Reby D. 2013 Are men better than women at acoustic size judgements? *Biol. Lett.* **9**, 20130270. (doi:10.1098/rsbl.2013.0270)
13. Pisanski K, Reby D. 2021 Efficacy in deceptive vocal exaggeration of human body size. *Nat. Commun.* **12**, 968. (doi:10.1038/s41467-021-21008-7)
14. Andersson MB. 1994 *Sexual selection*. Princeton, NJ: Princeton University Press.
15. Charlton BD, Reby D. 2016 The evolution of acoustic size exaggeration in terrestrial mammals. *Nat. Commun.* **7**, 12739. (doi:10.1038/ncomms12739)
16. Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton-Brock T. 2005 Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proc. R. Soc. B* **272**, 941–947. (doi:10.1098/rspb.2004.2954)
17. Clench MH. 1978 Tracheal elongation in birds-of-paradise. *The Condor* **80**, 423–430. (doi:10.2307/1367193)
18. Searcy WA, Nowicki S. 2005 *The evolution of animal communication: reliability and deception in signaling systems*. Princeton, NJ: Princeton University Press.
19. Reby D, Wyman M, Frey R, Charlton B, Dalmont JP, Gilbert J. 2018 Vocal tract modelling in fallow deer: are male groans nasalized? *J. Exp. Biol.* **221**, jeb179416. (doi:10.1242/jeb.179416)
20. Toutios A *et al.* 2016 Illustrating the production of the International Phonetic Alphabet sounds using fast real-time Magnetic Resonance Imaging. *Interspeech* **605**. (doi:10.21437/Interspeech.2016-605)
21. Fernández EM, Cairns HS. 2010 *Fundamentals of psycholinguistics*. New York, NY: John Wiley & Sons.
22. Ladefoged P, Johnson K. 2014 *A course in phonetics*. Boston, MA: Nelson Education.
23. Fitch WT, Boer Bd, Mathur N, Ghazanfar AA. 2016 Monkey vocal tracts are speech-ready. *Sci. Adv.* **2**, e1600723. (doi:10.1126/sciadv.1600723)
24. Boë LJ, Sawallis TR, Fagot J, Badin P, Barbier G, Captier G, Ménard L, Heim JL, Schwartz JL. 2019 Which way to the dawn of speech?: reanalyzing half a century of debates and data in light of speech science. *Sci. Adv.* **5**, eaaw3916. (doi:10.1126/sciadv.aaw3916)
25. Pisanski K, Cartei V, McGettigan C, Raine J, Reby D. 2016 Voice modulation: a window into the origins of human vocal control? *Trends Cogn. Sci.* **20**, 304–318. (doi:10.1016/j.tics.2016.01.002)

26. Sapir E. 1929 A study in phonetic symbolism. *J. Exp. Psychol.* **12**, 225. (doi:10.1037/h0070931)
27. Bentley M, Varon EJ. 1933 An accessory study of phonetic symbolism. *Am. J. Psychol.* **45**, 76–86. (doi:10.2307/1414187)
28. Johnson RC. 1967 Magnitude symbolism of English words. *J. Verbal Learn. Verbal Behav.* **6**, 508–511. (doi:10.1016/S0022-5371(67)80008-2)
29. Newman SS. 1933 Further experiments in phonetic symbolism. *Am. J. Psychol.* **45**, 53–75. (doi:10.2307/1414186)
30. Barreda S. 2017 An investigation of the systematic use of spectral information in the determination of apparent-talker height. *J. Acoust. Soc. Am.* **141**, 4781–4792. (doi:10.1121/1.4985192)
31. Parise CV, Spence C. 2012 Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Exp. Brain Res.* **220**, 319–333. (doi:10.1007/s00221-012-3140-6)
32. Peña M, Mehler J, Nespor M. 2011 The role of audiovisual processing in early conceptual development. *Psychol. Sci.* **22**, 1419–1421. (doi:10.1177/0956797611421791)
33. Pisanski K, Bryant GA. 2019 The evolution of voice perception. In *The Oxford handbook of voice studies* (eds NS Eidsheim, KL Meizel). New York, NY: Oxford University Press.
34. Morton ES. 1977 On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *Am. Nat.* **111**, 855–869. (doi:10.1086/283219)
35. Barreda S. 2017 Listeners respond to phoneme-specific spectral information when assessing speaker size from speech. *J. Phon.* **63**, 1–18. (doi:10.1016/j.wocn.2017.03.002)
36. Anikin A. 2019 Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behav. Res. Methods* **51**, 778–792. (doi:10.3758/s13428-018-1095-7)
37. Carpenter B *et al.* 2017 Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32. (doi:10.18637/jss.v076.i01)
38. Bürkner P-C. 2017 brms: an R package for Bayesian generalized linear mixed models using Stan. *J. Stat. Softw.* **80**, 1–28. (doi:10.18637/jss.v080.i01)
39. Kelley K, Maxwell SE, Rausch JR. 2003 Obtaining power or obtaining precision: delineating methods of sample-size planning. *Eval. Health Prof.* **26**, 258–287. (doi:10.1177/0163278703255242)
40. R Core Team. 2020 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.
41. Anikin A, Johansson N. 2019 Implicit associations between individual properties of color and sound. *Attent. Percept. Psychophys.* **81**, 764–777. (doi:10.3758/s13414-018-01639-7)
42. Dunn JC, Halenar LB, Davies TG, Cristobal-Azkarate J, Reby D, Sykes D, Dengg S, Fitch WT, Knapp LA. 2015 Evolutionary trade-off between vocal tract and testes dimensions in howler monkeys. *Curr. Biol.* **25**, 2839–2844. (doi:10.1016/j.cub.2015.09.029)
43. Charlton BD, Frey R, McKinnon AJ, Fritsch G, Fitch WT, Reby D. 2013 Koalas use a novel vocal organ to produce unusually low-pitched mating calls. *Curr. Biol.* **23**, R1035–R1036. (doi:10.1016/j.cub.2013.10.069)
44. Fitch WT. 2018 The biology and evolution of speech: a comparative analysis. *Annu. Rev. Linguist.* **4**, 255–279. (doi:10.1146/annurev-linguistics-011817-045748)
45. Ohala JJ. 1984 An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* **41**, 1–16. (doi:10.1159/000261706)
46. Schoen MA. 1971 The anatomy of the resonating mechanism in howling monkeys. *Folia Primatol.* **15**, 117–132. (doi:10.1159/000155371)
47. de Boer B. 2009 Acoustic analysis of primate air sacs and their effect on vocalization. *J. Acoust. Soc. Am.* **126**, 3329–3343. (doi:10.1121/1.3257544)
48. Ackermann H, Hage SR, Ziegler W. 2014 Brain mechanisms of acoustic communication in humans and nonhuman primates: an evolutionary perspective. *Behav. Brain Sci.* **37**, 529–546. (doi:10.1017/S0140525X13003099)
49. Dawkins R, Krebs JR. 1978 Animal signals: information or manipulation. *Behav. Ecol.* **2**, 282–309.
50. Charlton BD, Reby D, McComb K. 2007 Female red deer prefer the roars of larger males. *Biol. Lett.* **3**, 382–385. (doi:10.1098/rsbl.2007.0244)
51. Boë LJ, Berthommier F, Legou T, Captier G, Kemp C, Sawallis TR, Becker Y, Rey A, Fagot J. 2017 Evidence of a vocalic proto-system in the baboon (*Papio papio*) suggests pre-hominin speech precursors. *PLoS ONE* **12**, e0169321. (doi:10.1371/journal.pone.0169321)
52. Anikin A. 2020 The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics* **29**, 226–247. (doi:10.1080/09524622.2019.1581839)
53. Lawson E *et al.* 2015 *Seeing speech: an articulatory web resource for the study of phonetics* [website]. http://www.seeingspeech.ac.uk/.
54. Kreiman J, Sidtis D. 2011 *Foundations of voice studies: an interdisciplinary approach to voice production and perception*. Chichester, UK: Wiley-Blackwell.
55. Pisanski K, Isenstein SGE, Montano KJ, O'Connor JJM, Feinberg DR. 2017 Low is large: spatial location and pitch interact in voice-based body size estimation. *Attent. Percept. Psychophys.* **19**, 1239–1251. (doi:10.3758/s13414-016-1273-6)