Supplementary material for:

Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size

Andrey Anikin[1,2†*], Katarzyna Pisanski[2,3], Mathilde Massenet[2], David Reby[2]

[1] Division of Cognitive Science, Lund University, 22100, Lund, Sweden
[2] Equipe de Neuro-Ethologie Sensorielle, CNRS & University of Saint Étienne, UMR 5293, 42023, St-Étienne, France
[3] CNRS, French National Centre for Scientific Research, Laboratoire de Dynamique du Langage, University of Lyon 2, 69007 Lyon, France

ORCIDS: A. Anikin 0000-0002-1250-8261; K. Pisanski 0000-0003-0992-2477; M. Massenet 0000-0002-0085-1871; D. Reby 0000-0001-9261-1711

*Author for correspondence (andrey.anikin@lucs.lu.se).
†Address: Cognitive Science, Lund University, 22100, Lund, Sweden

Stimuli

Vocalizations were synthesized in *soundgen* (1). Voice synthesis was controlled manually for each of 82 prototype vocalizations, but the general principle was the same. First, we created a simple harmonic sound with some aspiration noise and with the f_0 contour of the original vocalization, including NLPs as per condition. The smoothed spectral and amplitude envelopes of the original were then applied to the synthetic sound using the *transplantEnvelope()* function in *soundgen*. The manipulation of NLPs, especially of chaos, has a large effect on the spectral envelope comparable to the addition of broadband noise – for example, it "dilutes" the harmonic structure and shifts spectral energy to formants. We were therefore careful to control the amount of spectral smoothing, which regulates how closely the synthetic sound conforms to the original spectral envelope. This was done on a case-by-case basis to achieve a precision sufficient to create an authentic-sounding vocalization, yet generalizable enough to accommodate the changes in glottal source associated with the presence or absence of NLPs (2). Screams with little broadband noise and with f_0 above the first formant ($F1$) are particularly challenging because their original formant structure is essentially invisible. These sounds were therefore resynthesized with simple schwa-like formant structure estimated for a randomly chosen and sex-appropriate vocal tract length (~13-17 cm), otherwise the absence of an audible $F1$ in the synthetic version with chaos sounded unnatural. In sum, although parametric (re)synthesis unavoidably involves making certain assumptions and compromises, we endeavored to create an array of vocalizations that sounded as natural as possible, yet differed systematically only in the feature of interest, namely the type of NLPs that they contained.

The synthesized stimuli were analyzed acoustically with the *soundgen* function *analyze()* to confirm the effect of NLPs on harmonics-to-noise ratio (HNR) and roughness (proportion of amplitude modulation within the roughness zone, 30 to 150 Hz in the spectrotemporal modulation spectrum . As shown in Table S1, we confirmed that vocalizations without NLPs were, as expected, the most tonal, whereas those with chaos were the roughest, and those with amplitude modulation and subharmonics were intermediate. The code for creating the stimuli, the original recordings, and their manipulated synthetic versions can be downloaded from http://cogsci.se/publications.html.

Table S1. Acoustic descriptives of experimental stimuli: Mean ± SD [range].

NLP Condition	N stimuli (male/female)	Duration, ms	Median pitch, Hz	HNR, dB	Roughness, %
No NLPs	82 (39/43)	1000 ± 531 [250, 2500]	859 ± 594 [242, 2420]	17.1 ± 2.1	11.9 ± 2.5
Amplitude modulation (AM)	82 (39/43)			13.6 ± 1.9	13.2 ± 3.0
Subharmonics	82 (39/43)			14.2 ± 1.9	15.7 ± 2.4
Chaos	82 (39/43)			6.1 ± 2.4	22.9 ± 3.7

Procedure

Rating test

Response scales used in the rating experiment are shown in Figure S1.

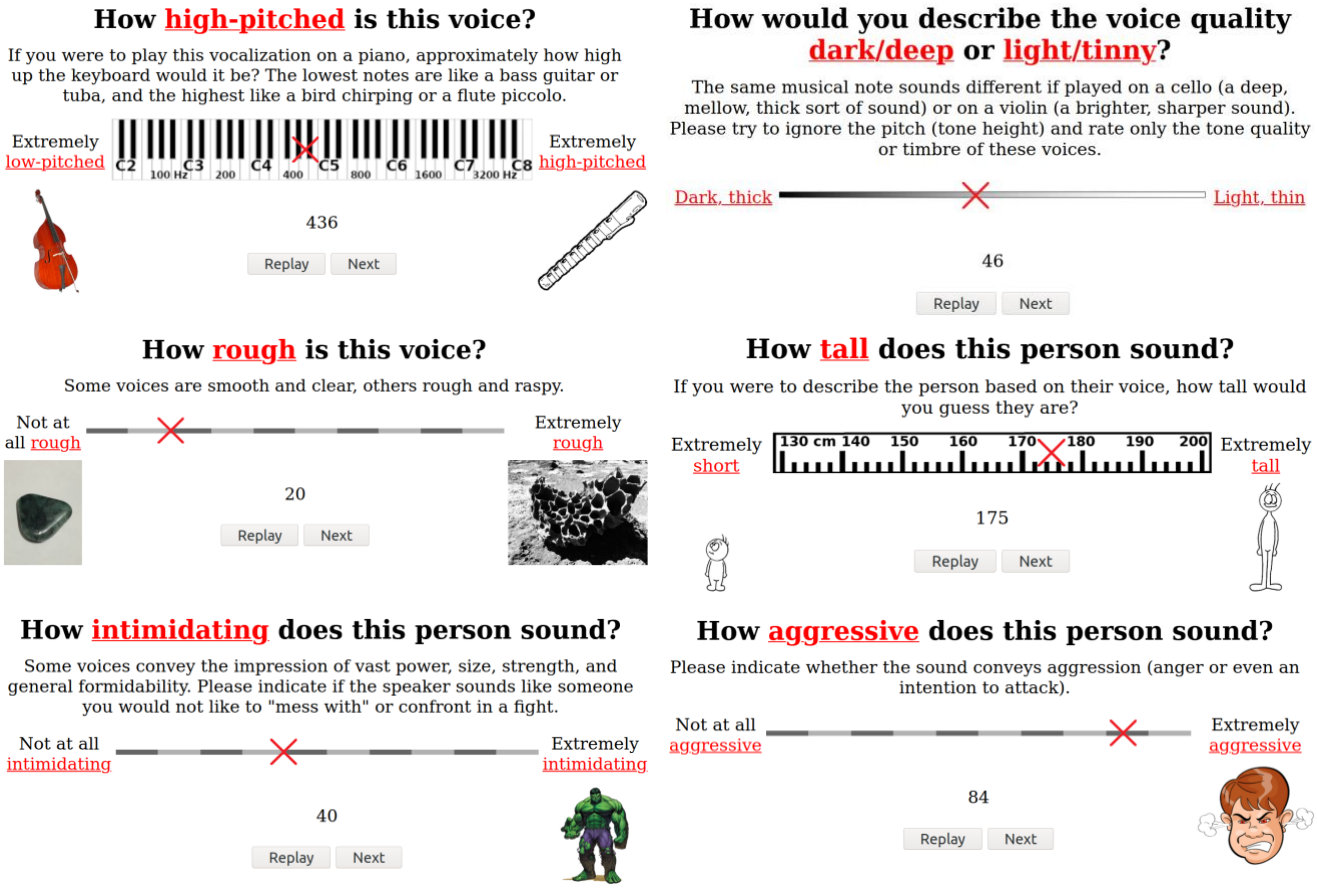

Figure S1. Response scales.

Implicit Associations Test

We implemented a web-based version of the Implicit Associations Test as described by Parise and Spence (3). 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. The pair of

visual stimuli was always the same, namely the images of a short and tall person used to illustrate the height scale in Experiment 1. The pairs of acoustic stimuli varied across experiments, but one was always a synthetic vocalization without NLPs and the other the same vocalization with some NLP (chaos, amplitude modulation, or subharmonics). The pairing rule changed in every block of 16 trials. For example, in one block of trials the image of a tall person and the sound with chaos might be assigned to the left arrow key, and the short person and the sound without NLPs to the right arrow key (a congruent combination). 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 blocks 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 pairs of 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 about 500 to 600 ms in duration, but participants could respond before the end of playback to ensure a fast pace. 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 (3).

Data analysis

Rating test

A single mixed model was fit to all 14911 trials using the R package *brms* (4), which predicted the rating in an individual trial as a function of Condition (4 levels: no NLPs, amplitude modulation, subharmonics, and chaos) and Scale (6 levels: pitch, timbre, roughness, height, formidability, and aggression), with an interaction. The effect of both predictors was assumed to vary across subjects (random slopes per subject, 301 levels) and across prototype vocalizations (random slopes per prototype, 82 levels). We also fitted a random intercept for each unique stimulus (328 levels); thus, the model structure in *brms* syntax was as follows:

*response ~ condition * scale + (condition * scale|subject) + (scale|sound) + (condition * scale|prototype)*

The outcome variable was the rating of a vocalization on a continuous scale, which was re-encoded to range from 0 to 1 for modeling purposes, in the case of pitch with an additional logarithmic transformation that converted Hz to semitones above the low end of the scale (62 Hz). These normalized ratings were then modeled with zero-one-inflated beta distribution (5) with four separately modeled parameters: (1) *mu*: the mean of beta distribution capturing non-extreme responses between 0 and 1; (2) *phi*: the precision of beta distribution; (3) *zoi*: zero-one inflation, the probability of answering 0 or 1 rather than a number in the interval (0, 1); (4) *coi*: conditional one-inflation, the probability of answering 1 rather than 0.

Two-alternative forced choice task

A multilevel ordinal logistic regression model was fit to estimate the most credible differences in perceived size between the six possible pairs of NLP conditions. The model was of the form:

size ~ pair + (pair|subject) + (pair|prototype),

where size was encoded as “1” (first person judged as taller), “2” (no difference), or “3” (second person judged as taller), and size preferences were allowed to vary both across subjects and across the 82 prototypes. The measure of interest was the posterior distribution of the difference in the probability of

obtaining a size rating of “3” rather than “1” contrasted for each pair of NLP conditions both globally (population effect) and separately for each prototype (group-specific slopes).

Implicit Associations Test

All training trials were discarded, and only test trials were analyzed. A single model was fit to this unaggregated dataset to analyze the accuracy from all six experimental blocks (N = 46,717 trials), and another to analyze response times (RT) in trials with correct responses and RT under 5 s (N = 43,754 trials). The models were as follows, in *brms* / *lme4* syntax:

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

*RT (lognormal): responseTime ~ experiment * 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 vs. touchscreen). Finally, congruence effects were allowed to vary across participants.

References

1. Anikin A. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. Behavior research methods. 2019;51(2):778–92.

2. Anikin A. The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. Bioacoustics. 2020;29(2):226-247.

3. Parise CV, Spence C. Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. Experimental Brain Research. 2012;220(3–4):319–33.

4. Bürkner P-C. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software. 2017;80(1):1–28.

5. Ospina R, Ferrari SL. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis. 2012;56(6):1609–23.