

Reviewed Preprint

v1 • August 21, 2026

Not revised

✉ **For correspondence:**

hansbilger@utexas.edu

Competing interests: No

competing interests declared

Funding: See [page 13](#)

Reviewing editor: Daniel Y

Takahashi, Universidade Federal do Rio Grande do Norte, Brazil

© 2026, Bilger et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Measurement and comparison of acoustic space use in vocalizations of humans and close primate relatives

Hans T Bilger¹ ✉, Michael J Ryan^{1,2}, Julia A Clarke^{1,3}

¹Department of Integrative Biology; University of Texas at Austin, Austin, United States • ²Smithsonian Tropical Research Institute, Balboa, Panama • ³Department of Geological Sciences; University of Texas at Austin, Austin, United States

eLife Assessment

This study presents an **important** empirical analysis of the acoustic space of vocal repertoires across primates and humans, providing new data that challenge the widely held hypothesis that the expansion of the human vocal space, driven by modifications of the vocal tract, was a key prerequisite for the evolution of spoken language. The evidence is **convincing** in its technical implementation and acoustic measurements, offering helpful comparative insights into voiced vocalizations. However, the theoretical framing and interpretation would benefit from further discussion.

<https://doi.org/10.7554/eLife.112153.1.sa2>

Abstract

The human larynx, compared to those of closely related primates, lies deeper in the throat and lacks vocal membranes and air sacs. These shifts are usually analyzed regarding their acoustic effects on vowel-like vocalizations, since the evolution of speech was long thought to require an expansion of vocal range driven by vocal tract modifications. However, vowels are just one type of phoneme, and speech is just one class of human utterance. To understand the evolutionary underpinnings of known shifts in human vocal morphology, a broader bioacoustic comparison is needed. Specifically, the range of sounds used in human speech must be compared to that employed in other human vocalizations and in the repertoires of extant close primate relatives. Here, we measure the acoustic-feature space occupied by human speech, non-linguistic, and musical vocalizations along with the calls of chimpanzees, bonobos, and chacma baboons. We use Mel-frequency cepstral coefficients to create an acoustic space depicting the spectro-temporal features of over 750,000 brief vocal segments sourced from published databases and other verified sources. Speech and song occupied significantly less volume in this acoustic space than human non-linguistic vocalizations. In addition, the acoustic-feature volumes of speech and song were not statistically distinct from those of non-human primates. These results suggest that speech was not enabled by an expansion of human vocal acoustic space. Anatomical shifts unique to humans may have led to an elaboration of non-linguistic utterances, but learned vocalizations use a surprisingly small fraction of this space. Our understanding of human vocal evolution will be further informed by additional systematic comparisons of the function and homology of non-speech vocalizations, along with the collection and incorporation of more complete non-human primate vocal datasets, especially from Gorilla and Orangutan.

1. Introduction

Complex vocalizations play an important role in human social structures, and can extend beyond the acoustic boundaries of speech and song to encompass a broader range of expressive sounds (Freeberg et al., 2012 [↗](#); McComb & Semple, 2005 [↗](#); Peckre et al., 2019 [↗](#); Sewall, 2015 [↗](#)). Morphological changes in the vocal organ and tract that are unique to humans, relative to their

nearest relatives, affect all vocal sounds produced (Fitch, 2000; Fitch, 2018). However, complex human non-linguistic utterances produced with these modified structures have rarely been considered in comparative studies of human vocal evolution (Anikin et al., 2023).

The human vocal apparatus shows several derived morphological features compared to close primate relatives. First, the larynx lies lower in the throat and the tongue descends farther into the pharynx (D. Lieberman, 2011; P. H. Lieberman et al., 1969). Second, it lacks the supralaryngeal air sacs displayed by many monkeys and apes (De Boer, 2008; Hewitt et al., 2002). Finally, it has lost the thin, reed-like vocal membranes which lie above the vocal folds in all non-anthropoid primates (Nishimura et al., 2022). These shifts have often been linked with the evolutionary emergence of speech, which requires both the neuromotor ability to modify vocal output based on auditory experience—called vocal production learning—and a vocal apparatus mechanically capable of producing a sufficient number of perceptually contrasting sounds (Brenowitz & Beecher, 2023; Fitch, 2000; Janik & Knörnschild, 2021). Lowering of the larynx and lengthening of the tongue have been argued to increase the range of possible vowel-like articulations; air sac loss has been claimed to increase vowel discriminability; and the loss of vocal membranes (Fig. 1, A) has been proposed to stabilize the vocal source, making it simpler to produce a consistent bed of high-bandwidth tonal sound for subsequent vocal tract filtering (De Boer, 2012; Fitch, 2018; Fitch et al., 2016; Hewitt et al., 2002; P. H. Lieberman et al., 1969; Nishimura et al., 2022).

Speech-relevant shifts in the anatomy and neurology of the human vocal production system have led to distinct hypotheses of speech emergence. The “morphological” (or “peripheral”) hypothesis of speech emergence—that a morphology-driven expansion of human vocal output range enabled the evolution of speech—stands in contrast to the “neural” hypothesis of speech emergence, which ties vocal production learning (VPL) to speech origin (Darwin, 1871; Fitch, 2018). The morphological hypothesis was widely accepted for decades, but it has been challenged in recent years through studies demonstrating laryngeal descent in other taxa and the speech-capability of non-human primate vocal mechanisms (Boë et al., 2017; Fitch et al., 2016). It also remains untested in a crucial sense: Thus far, human vocal range shifts have mainly been measured using vowel-like articulations, which represent a small subset of our species’ vocal output.

Here, we assess evolutionary hypotheses of speech emergence by measuring the range of spectral envelopes, or “timbres”, present in a large portion of the human vocal output range and comparing it to those of key primate outgroups. To characterize these short-time envelopes, we used acoustic features known as mel-frequency cepstral coefficients (MFCCs). MFCCs provide a sparse, perceptually weighted numeric description of a vocalization’s short-time spectral envelope (Davis & Mermelstein, 1980). They were originally designed for human speech recognition tasks, and have since been used extensively in the study and classification of other primate vocalizations recorded in noisy and naturalistic environments (Cauzinille et al., 2024; Lakdari et al., 2024). While these features do not capture every aspect of vocal acoustic structure (such as fundamental frequency, syntactical arrangement, etc.), they have proven extremely useful for characterizing the various signatures of the upper vocal tract transfer function. The filter characteristic of the upper vocal tract is one of the chief perceptual contrast agents in human speech (it creates formant patterns and vowel identity), so the MFCC functions as a well-suited, perceptually-informed proxy of communication-relevant vocal range in humans and other primates.

Human non-linguistic utterances (cries, laughs, screams, etc.) are largely innate and produced through a broadly conserved neurological pathway older than and separate from the one used for generating speech and song (Jürgens, 2009). By comparing the volumes of the acoustic-feature spaces occupied by speech, song, non-linguistic vocalizations, and the vocal repertoires of extant primate relatives, we can empirically test for differences in the vocal output range of humans and non-human primates that lack human laryngeal novelties. We can also determine whether these differences are uniquely seen in learned vocalizations (speech, song). These analyses allow us to discriminate between alternative scenarios of speech emergence, since the morphological hypothesis predicts a speech-centered expansion of the range of human vocal sound qualities due to human-specific anatomical shifts, while the neural hypothesis does not. Specifically, if the

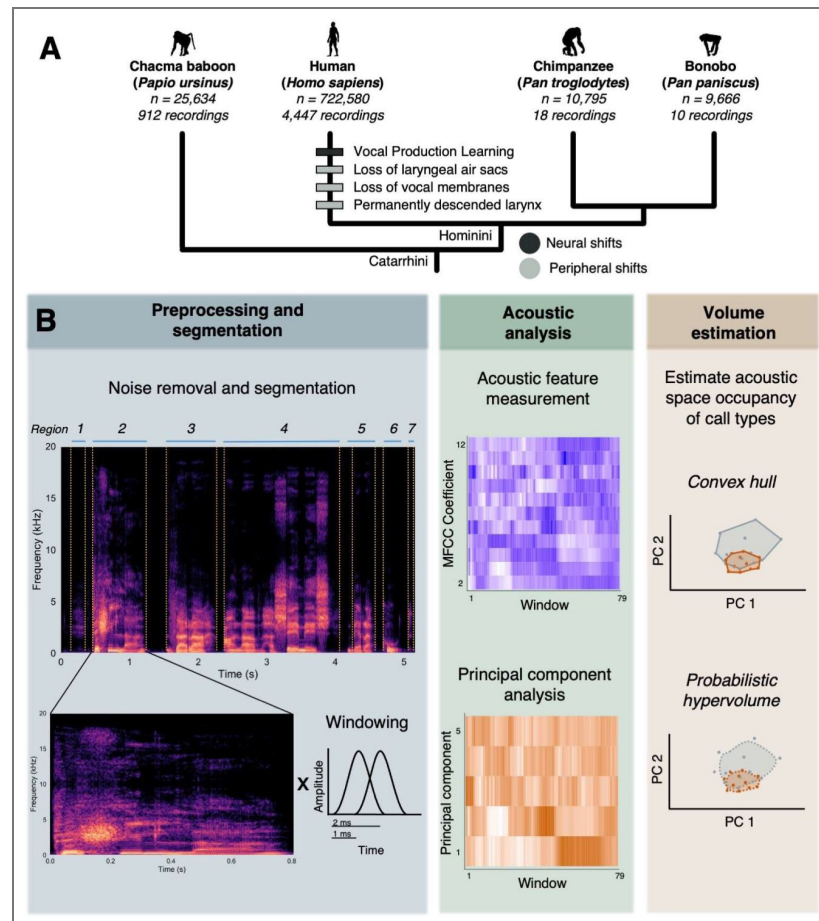


Fig. 1. Evolutionary context for speech emergence and acoustic analysis pipeline.

(A) Phylogenetic tree showing human autapomorphies relative to extant non-human primates. Shifts in the neural underpinnings of vocal control are marked in dark blue; peripheral shifts in vocal anatomy are marked in light gray. (B) Analysis pipeline for vocalization recordings. First (Preprocessing and segmentation panel), each recording was stripped of all possible extraneous noise and segmented into regions of quasi-continuous sound (top and bottom left). Then each region was split into a series of partially overlapping Hamming windows with a length of 0.02 s and an overlap of 50% (bottom right). Next (Acoustic analysis panel), each window was converted into 11 Mel-frequency cepstral coefficients. Principal component analysis (PCA) was then performed on this centered, scaled dataset of acoustic features. Finally (Volume estimation panel), the acoustic-feature space occupied by each vocalization type was estimated using convex hulls and probabilistic hypervolumes. PCs 1–5 were used to estimate these volumes. [Note to publisher: please print this figure in color.]

acoustic-feature volume occupied by human speech sounds is larger than that of nonlinguistic vocalizations or the vocalizations of nonhuman primates, this would support the morphological hypothesis. If the acoustic-feature volume of speech is less than or equal to that of the other groups, it would falsify the morphological hypothesis and be consistent with the neural hypothesis, which makes no predictions regarding evolutionary shifts in acoustic-feature volume.

2. Materials and Methods

2.1 Participant and experimental model details

Our dataset of 11,691 short segments of recorded human and non-human primate (chacma baboon, *Papio ursinus*; chimpanzee, *Pan troglodytes*; and bonobo, *Pan paniscus*) vocalizations from 5,388 individual recordings (Fig. 1A [↗](#); Table 1 [↗](#), Table 2 [↗](#)) is a novel synthesis of human and non-human Hominidae data. All recordings were sourced from published databases and other verified sources; since the study did not recruit novel participants, an ethics committee review was unnecessary. Chimpanzees and bonobos were included in the dataset because they are the closest living relatives of humans. Chacma baboons were included because they possess the most complete acoustic dataset available for the vocal output of a non-human primate. All three non-human taxa lack the human novelties described above. We were unable to locate a sufficiently comprehensive set of Gorilla (*Gorilla gorilla*) or Orangutan (*Pongo pygmaeus*) vocal recordings in the published literature, so these taxa were not included in the analysis.

2.2 Recording sources

Human speech vocalizations were compiled by the International Phonetic Association and included a mixture of recordings of single words, phrases, and full sentences from adult male and/or female speakers. Recorded single words were chosen to represent the phonemic inventory of each language (International Phonetic Association, 1999 [↗](#)). Twenty-seven languages from nine language families were represented in the corpus: American-English, Amharic, Arabic, Bulgarian, Cantonese, Catalan, Croatian, Czech, Dutch, French, Galician, German, Hausa, Hebrew, Hindi, Hungarian, Igbo, Irish, Japanese, Korean, Persian, Portuguese, Sindhi, Slovene, Swedish, Thai, and Turkish. No age or sex information was given for recordings except for those of Slovene, which contained vocalizations from male and female speakers.

Recordings of human singing were compiled from two broad cross-cultural samples of recorded human music mostly from small-scale societies (Mehr et al., 2019 [↗](#); Nettl et al., 1998 [↗](#)) and a laboratory study of biphonic throat singing (Bergevin et al., 2020 [↗](#)). In all musical corpora, only recordings of solo, unaccompanied singers were selected for further analysis. Human song can contain both linguistic and non-linguistic elements; all elements of song were included in the final dataset.

Human non-linguistic vocalizations were sourced from two databases. The first included 56 speakers (26 female, 30 male, ages 18–35, from USA, India, Kenya, and Singapore) who were asked to produce a variety of non-linguistic vocalizations relating to 30 emotional categories (Cowen et al., 2019 [↗](#)). The second was an archive of 603 spontaneously produced human non-linguistic vocalizations, representing nine emotional categories, collected from YouTube (Anikin & Persson, 2017 [↗](#)). In both cases, only vocal segments without obvious linguistic content were selected for further analysis.

Chacma baboon recordings were sourced from Wadewitz et al. (2015) [↗](#). This corpus comprised 912 recorded calls from wild chacma baboons in the Moremi Wildlife Reserve in Botswana. The dataset includes calls from 34 males, 35 females, 5 infant males, and 4 infant females. Recordings were selected to represent the species' entire described vocal repertoire. Chimpanzee vocalizations were sourced from the Macaulay Library at the Cornell Lab of Ornithology. All recordings used in the final analysis were collected in the Mahale Mountains National Park and Gombe National Park in Tanzania. The recordings included a variety of call types from male and female adults and juveniles. Only recordings with a “five-star” rating that included mostly single callers were selected for further processing. The following recordings were used in the final

Vocalization type	Data source	Link
Human nonlinguistic	A. Anikin, T. Persson, Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. Behavior research methods. 49, 758–771 (2017).	http://www.cogsci.se/publications/2017_corpus.html
Human nonlinguistic	A. S. Cowen, H. A. Elfenbein, P. Laukka, D. Keltner, Mapping 24 emotions conveyed by brief human vocalization. American Psychologist. 74, 698 (2019).	https://supp.apa.org/psycarticles/supplemental/amp0000399/amp0000399_suppl.html
Human speech	International Phonetic Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet (Cambridge University Press, 1999).	https://www.internationalphoneticassociation.org/content/ipa-handbook-downloads
Human song	B. Nettl, R. Stone, J. Porter, T. Rice, Eds., The Garland Encyclopedia of World Music (Garland, New York, 1998).	https://search.alexanderstreet.com/gln
Human song	S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, Universality and diversity in human song. Science. 366 (2019)	https://osf.io/vcybz/
Human song	C. Bergevin, C. Narayan, J. Williams, N. Mhatre, J. K. Steeves, J. G. Bernstein, B. Story, Overtone focusing in biphonic Tuvan throat singing. Elife. 9, e50476 (2020).	https://datadryad.org/stash/dataset/doi:10.5061/dryad.cvdncjt14
Chacma baboon	P. Wadewitz, K. Hammerschmidt, D. Battaglia, A. Witt, F. Wolf, J. Fischer, Characterizing vocal repertoires—Hard vs. soft classification approaches. PloS one. 10, e0125785 (2015).	Provided by original investigators upon request
Bonobo	Dr. Frans de Waal, Charles Howard Candler Professor of Primate Behavior, Department of Psychology, Emory University	Provided by original investigator upon request
Bonobo	Bonobo Conservation Initiative / Dr. Zanna Clay, Associate Professor, Department of Psychology, Durham University	https://www.bonobo.org/bonobos/what-do-bonobos-sound-like
Chimpanzee	Macaulay Library at the Cornell Lab of Ornithology	https://www.macaulaylibrary.org/

Table 1. Recording sources.

Vocalization type	Data source	Original sample rate	File type	Recordings (n)	Audio segments (n)	Time windows (n)	Learned vocalization?
Human nonlinguistic	Anikin and Persson 2017	n/a	WAV	563	1644	65691	Some
Human nonlinguistic	Cowen et al. 2019	n/a	mp3	2030	2030	199889	Some
Human speech	International Phonetic Association 1999	20 kHz / 22.05 kHz / 44.1 kHz	WAV	1660	3745	158953	Yes
Human song	Nettl et al. 1998	n/a	mp3	66	1387	221119	Yes
Human song	Mehr et al. 2019	n/a	mp3	118	354	44136	Yes
Human song	Bergevin et al. 2020	96 kHz	WAV	10	50	32792	Yes
Chacma baboon	Wadewitz et al. 2015	44.1/48 kHz; later downsampled by original researcher to 20 kHz/5 kHz	WAV	912	994	25634	No
Bonobo	Dr. Frans de Waal	44.1 kHz	WAV	4	576	8203	No
Bonobo	Bonobo Conservation Initiative	n/a	mp3	6	64	1463	No
Chimpanzee	Macaulay Library at the Cornell Lab of Ornithology	44.1 kHz	WAV	19	847	10795	No

Table 2. Acoustic data used in analyses of acoustic feature space.

The table depicts the number of original recordings downloaded from each data source (“Recordings (n)”), the number of quasi-continuous vocal segments that each group of recordings was divided into following pre-processing (“Audio segments (n)”), and the number of partially overlapping Hamming windows (length = 0.02 s; overlap = 50%) that each group of audio segments was divided into (“Time windows (n)”).

analysis: ML 53994, ML 54074, ML 162001, ML 162110, ML 162158, ML 162306, ML 163515, ML 163593, ML 163596, ML 163604, ML 163634, ML 163639, ML 163641, ML 163650, ML 189988, ML 196196, ML 196248, ML 196251, and ML 196252.

Bonobo recordings were sourced from two locations. The first corpus of six recordings was provided by Dr. Zanna Clay and published by the Bonobo Conservation Initiative (Clay, 2022 [↗](#)). This corpus included recorded calls of six types: “high hoots”, “food calls”, “travel calls”, “alarm and alert” calls, “threat barks and anger calls”, “male contest hoots,” and “laughter”. No information was provided on the age, sex, or location of callers. The second corpus was provided via personal communication by the late Dr. Frans de Waal. It included four digitized recordings of “hoots” and “screams” made by Bonobos in the San Diego Zoo. No information on caller age or sex was provided.

2.3 Bioacoustic analyses

2.3.1 Recording pre-processing and segmentation

To measure the range of sound qualities contained in each vocalization category, we constructed a morphospace derived from acoustic-features known as Mel-frequency cepstral coefficients (MFCCs). MFCCs are human hearing-informed spectral summary features used for vocal signal analysis in a variety of taxa (Clink et al., 2019 [↗](#); Lee et al., 2006 [↗](#); Mielke & Zuberbühler, 2013 [↗](#)). The MFCC was designed to characterize the short-term spectral envelope of human speech segments, so we assume it exaggerated variance in the human speech sounds in our dataset relative to other vocalizations. To prepare our audio data for MFCC extraction, each of the 5,388 recordings in the initial dataset was aurally and visually inspected. Recordings with high levels of environmental noise or obvious acoustic distortion were rejected. Pre-treatment noise attenuation has been shown to increase the performance of MFCC feature extraction (Soe Naing et al., 2020 [↗](#)), so environmental noise and other acoustic artifacts were attenuated using a suite of audio processing tools in the program iZotope RX 9 Audio Editor (v9.2.0.1278; iZotope, Inc., 2007 [↗](#)). Modules used within the program included “Voice De-noise”, “De-plosive”, and “De-clip”. Next, each recording was split into regions of quasi-continuous sound via visual assessment of spectro- and oscillograms (Fig. 1 [↗](#), B). For recordings with particularly high signal-to-noise ratios, the “Truncate Silence” feature in the audio recording and editing program Audacity (v2.4.2) was used to partially automate the segmentation process (Audacity Team, 2021 [↗](#)). This process yielded 11,691 audio segments from <0.5 s to several seconds long. Next, each audio segment was converted to WAV format if necessary and resampled to 100 kHz using the *resamp* function in the R package *seewave* (Sueur et al., 2008 [↗](#)).

2.3.2 MFCC analysis

To calculate MFCCs, we first split each segment into a series of partially overlapping Hamming windows with a length of 0.02 s and an overlap of 50% ($n=768,675$). We then estimated the power spectrum of each window by conducting a short time Fourier transform and calculating its absolute value. This spectrum was then multiplied by a bank of frequency filters based on the Mel scale, a mathematical construct meant to approximate the response of the human auditory system. Finally, a discrete cosine transform was performed on the log of each filter bank’s total energy. This step has the effect of decorrelating and removing the vocal source signal (i.e. the pitch and harmonics of the voice) from the filter signal (i.e. broad spectrum shape and formant/vowel information), making it more robust at identifying speech-relevant acoustic features. Each of these transformed filter values constituted one Mel-frequency cepstral coefficient. We collected 12 MFCCs for each window within the range 0–20 kHz. No pre-emphasis filtering or liftering was used. MFCC features were collected using the function *melfcc* in the R package *tuner* (Ligges et al., 2018 [↗](#)). Since the first MFCC for each time window chiefly describes the loudness of the segment rather than its spectral envelope, it was omitted it from subsequent analyses.

2.3.3 Acoustic-feature volume analyses

Hypervolume analysis becomes extremely computationally expensive with increasing dimensionality, so principal component analysis (PCA) was performed on the centered, scaled MFCC dataset. Then, the first five PCs from each window (which accounted for ~68% of cumulative variance) were used as input data to construct shapes for each vocalization type within acoustic feature space. The five-dimensional volumes of these shapes were then estimated using convex hulls and probabilistic hypervolumes (Blonder et al., 2018 [DOI](#); Habel et al., 2019 [DOI](#); Hutchinson, 1957 [DOI](#)).

Convex hulls were constructed using the “convhulln” function from the R package *geometry* (Habel et al., 2019 [DOI](#); R Core Team, 2022 [DOI](#)). Hypervolumes were constructed using the “hypervolume_svm” function from the R package *hypervolume* (Blonder et al., 2022 [DOI](#)). The *hypervolume* package can use three different methods to construct hypervolumes: box kernel density estimation (“box”), Gaussian kernel density estimation (“gaussian”), and one-class support vector machines (“SVM”). We chose to use the “SVM” method because it responded most efficiently to the study’s large acoustic dataset. Default parameters were used for each application of the “hypervolume_svm” function.

Estimated acoustic-feature volumes were then compared across vocal types to test and generate hypotheses of acoustic space use in hominid vocal evolution. The patterning of points in acoustic-feature space offers information about the basic range of vocal sound qualities expressed by a given taxon or vocalization category, but says nothing about their higher-level temporal ordering (i.e., the rhythm or syntax of vocal phrases).

2.4 Quantification and statistical analysis

2.4.1 Statistical validation—randomization tests

Results were statistically validated using randomization tests. For each comparison of interest (i.e. nonlinguistic vs. speech, human song vs. chacma baboon, etc.) group labels were shuffled and convex hulls/hypervolumes were recomputed. This process was repeated 10,000 times for each comparison (Figs. S1, S2 [DOI](#)). To determine whether the observed difference in hypervolume size was statistically significant, the following calculation was performed:

$$p \text{ value} = (\# \text{ times } | \text{ simulated difference } | \geq | \text{ observed difference } |) / 10,000$$

This yielded the *p*-values reported in Table S1 [DOI](#). All statistical computations were conducted in R (R Core Team, 2022 [DOI](#)).

2.5 Materials availability

Information on the sources and accessibility of the original recordings analyzed in this manuscript is provided in Table 1 [DOI](#). All recordings are available either in public repositories, for purchase from the original publisher, or through reasonable request to the respective original authors. Original recordings cannot be reproduced here because the authors of this manuscript do not have permission to mirror all original recording files used in the analysis.

3. Results and Discussion

In both convex hull and hypervolume analyses, human speech was volumetrically smaller than non-linguistic vocalizations (randomization test, Fig. 2 [DOI](#), A–C; Table 3 [DOI](#)). This result shows general agreement with the recent speech-nonverbal comparison carried out by Anikin et al. (2023) [DOI](#), in which (1) nonverbal vocalizations showed significantly more fundamental frequency modulation than speech, and (2) speech sounds occupied a significantly larger amount of vowel space (as defined by the first two formant frequencies F1 and F2) than nonverbal sounds. The second result may initially appear contradictory with the present study, but here it is important to note that while vowel/formant space and MFCCs both capture aspects of the vocal filter function,

they do not report exactly the same qualities. Formant frequencies are frequency ranges in the filter envelope that show local maxima, and they are determined solely by the configuration of the vocal tract (Fitch et al., 2025 [↗](#)). MFCCs, on the other hand, report the relative acoustic energy present across all the frequency bands of the filter envelope—and the degree which those bands are excited depends partly on the behavior of the vocal source. For example, a low frequency scream-like vocalization and a spoken vowel could both use the same vocal tract configuration, resulting identical F1 and F2 frequencies. But in the case of the scream, high-frequency MFCC features would show greater values because of the added harmonics created by nonlinearities in the vocal source. Notably, Anikin et al. (2023) [↗](#) excluded two-thirds of screams and 10–25% of nonverbal vocalization types from their formant analysis since they were either “too high pitched for formant analysis or didn’t contain sufficiently long and stable vowel-like regions.” Taken together, the results of both studies suggest that while speech made richer use of the aspect of the vocal filter most important for lexical coding—vowel/formant space—it still reduced the number of general filter envelope shapes compared to nonlinguistic vocalizations.

The volumetric difference between speech and non-linguistic vocalizations can be partly explained through three of speech’s key features: It is quiet, informationally dense, and formant-centered. First, it is *quiet*: Unlike distress signals or advertisement calls, which are designed to propagate over large distances, speech is generally reserved for moderate-intensity, affiliative social interactions (Fitch, 2000 [↗](#)). Some regions of acoustic space—roars, shrieks, hysterical laughter, etc.—can only be accessed at mechanical extremes of the vocal system. Such sounds could startle intended receivers or attract predators or eavesdroppers (Morton, 1977 [↗](#)). Speech is also *informationally dense*. The limited temporal resolving power of the human auditory system along with the general tendency for biological systems to strive for maximally reliability at minimal metabolic cost (known as Zipf’s Principle of Least Effort) has pressured speech production to become more mechanically efficient over time (Coupé et al., 2019 [↗](#); Zipf, 1949 [↗](#)). Many modern phonemes are “coarticulated”, meaning that their most extreme acoustic forms are traded for intermediate versions (Farnetani & Recasens, 2010 [↗](#)). This simplifies the transitions between vocal tract configurations at the cost of reduced acoustic contrast between phonemic elements and increased load on the auditory brain.

The relatively low-contrast, highly informational quality of speech sounds partially explains why they are so *formant-centered*. Most vowel types are perceptually identified using the first two formant frequencies, and the spectral envelopes of these lower formants are easier to discern when they are excited by a low frequency source dense with harmonics (Anikin et al., 2023 [↗](#); Nishimura et al., 2022 [↗](#); Titze & Palaparthi, 2018 [↗](#)). Furthermore, critical speech frequencies align with the region of the human auditory system with the best pitch discrimination. Average speech f_0 is approximately 131 Hz for males and 220 Hz for females, with F1 frequencies of approximately 300–1,000 Hz (J. Hillenbrand et al., 1995 [↗](#); J. M. Hillenbrand & Clark, 2009 [↗](#)). A strong sensation of “pitch” is only perceived for harmonic complex tones with a frequency range ~30–4,000 Hz. Within that range, harmonic complex tones sound most “tonelike” (a feature associated with pitch and strongly associated with the formation of clear auditory images) between 82–784 Hz, with a maximum around 300 Hz (Huron, 2016 [↗](#); Oxenham, 2012 [↗](#)). There is therefore a conspicuous alignment between frequencies used by the vocal production system to create speech-like sounds and the tuning of the auditory system that perceives them. Human screams can exceed 2 kHz in fundamental frequency (Pisanski et al., 2020 [↗](#)), rendering that class of sounds less useful for the encoding of linguistic information.

The volume of acoustic space occupied by human speech was also and not significantly different from non-human primates in the present study (randomization tests, Fig. 2, A, B, D [↗](#); Table 3 [↗](#)). This result is—withstanding the caveats above regarding MFCCs and vowel space—consistent with recent work from Fitch et al. (2016) [↗](#) and Boë et al. (2017), who found non-significant differences in vowel space size between humans, macaques, and baboons. The lack of volumetric difference between speech and non-human primates was especially striking given disparities in sample sizes. Acoustic-feature volumes can only grow with increased sampling, and speech had ~6x more observations than Chacma baboons, ~16x more than chimpanzees, and ~15x more

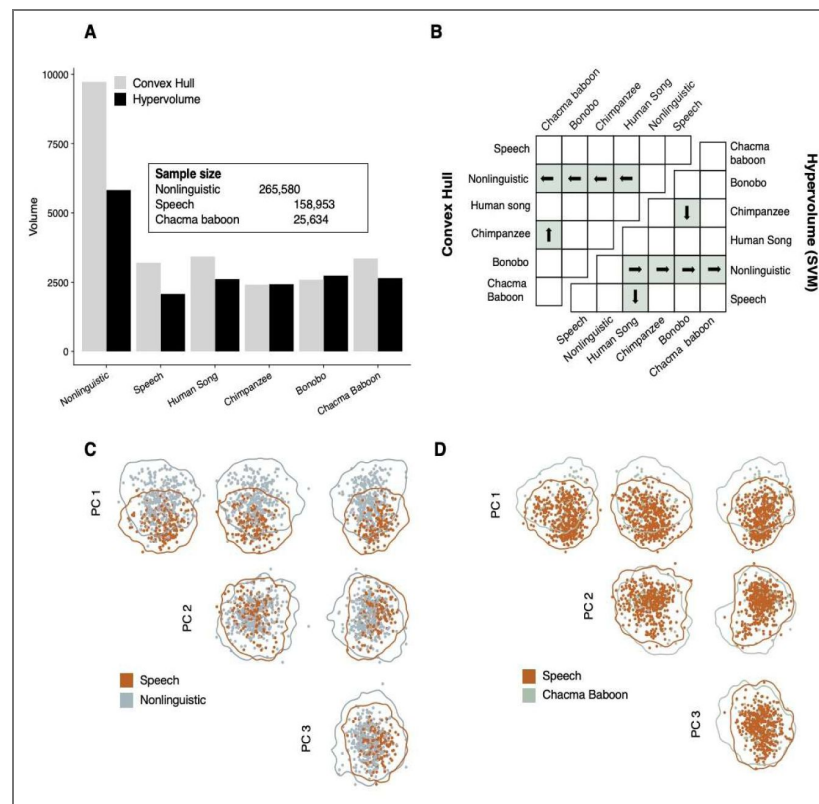


Fig. 2. Non-linguistic vocalizations occupy more acoustic-feature space than speech, while speech and non-human primate outgroup repertoires occupy similar amounts.

(A) Convex hull and probabilistic hypervolume volumes from a PCA including all data and constructed from MFCC 2–12. Non-linguistic sounds occupied more acoustic-feature space than speech, but the former group also contained more observations. Speech and chacma baboons occupied similar amounts of acoustic-feature space despite the former having more than 6x as many observations. **(B)** Differences in hypervolume size. Green color denotes significant differences; arrows point to the larger hypervolume in each dyad. Non-linguistic sounds occupied more acoustic-feature space than all other vocalization types in both analyses. No other significant group differences were recorded across both hypervolume and convex hull analyses. **(C)** Two-dimensional projections of hypervolumes for speech (orange) and non-linguistic vocalizations (blue). The larger size of the non-linguistic hypervolume is evident from the shape outlines. **(D)** Two-dimensional projections of hypervolumes for speech (orange) and chacma baboon vocalizations (green). The approximately equal size of the two hypervolumes is evident from the shape outlines. Chacma baboon was the non-human primate whose vocal repertoire was most fully captured in the dataset. [Note to publisher: please print this figure in color.]

Method	Group	Volume
Hypervolume	Nonlinguistic	5825.0
Hypervolume	Speech	2077.0
Hypervolume	Human Song	2611.5
Hypervolume	Chimpanzee	2429.0
Hypervolume	Bonobo	2734.7
Hypervolume	Chacma Baboon	2647.2
Convex Hull	Nonlinguistic	9727.7
Convex Hull	Speech	3200.9
Convex Hull	Human Song	3425.7
Convex Hull	Chimpanzee	2409.1
Convex Hull	Bonobo	2587.7
Convex Hull	Chacma Baboon	3359.8

Table 3. Observed acoustic-feature space volumes.

than bonobos. These results are notable for the inclusiveness of the sampling. While prior comparative studies, including those of [Fitch et al. \(2016\)](#), have generally relied on single or more distantly related non-human primate species for their comparisons, the present study includes samples from the two closest human relatives (chimpanzee and bonobo) along with an example from a more distant outgroup (*Chacma baboon*).

Song and speech occupied similar volumes of acoustic-feature space ([Fig. 2, A, B](#); [Table 3](#)). Song space was slightly larger in both convex hull and hypervolume analyses, but this difference was only significant in the latter. Since hypervolumes are less sensitive to outliers than convex hulls, our results indicate that the song data set contains a larger number of frequently used outliers than speech. Song is regularly transmitted over both long and short distances (consider opera singing vs. lullabies), so it might make more frequent use of acoustic extremes than speech. This quality would be in line with the finding that mammal and bird vocalizations with higher f_0 radiate acoustic power more efficiently, making them more suitable for long distance communication ([Titze & Palaparthi, 2018](#)). This result is also roughly consistent with [Anikin et al.'s \(2023\)](#) analysis, which suggested that the range of fundamental frequencies produced in singing is greater than in speech and comparable with that of nonlinguistic vocalizations, with sex specific effects on the ranges of each vocal category.

Human song also figures in an influential hypothesis of language evolution known as the “musical protolanguage” ([Darwin, 1871](#); [Fitch, 2013](#)). Since song requires vocal learning but not hierarchical syntax or complex semantics—two other key derived components of language—it has been proposed as an intermediate step between well-conserved innate calls and modern speech. This hypothesis suggests an evolutionary step where the sounds of the non-linguistic acoustic space were given a basic temporal ordering or coarse syntax. However, extant song often contains spoken elements, so it cannot be homologous with the “song” of the musical protolanguage—it is derived from speech. In our analysis, song occupied significantly less acoustic space than non-linguistic vocalizations and a similar amount to speech (randomization test, [Fig. 2, A, B](#); [Table 3](#)). These results are consistent with extant song being speech-derived. It is notable that both vocally learned vocalizations in our dataset appear to have led to a contraction of the exploited acoustic space rather than the opposite. In each case, learning seems to have led to a refinement of timbral options rather than an expansion. A comparison with [Anikin's \(2023\)](#) results is also informative here. Their analysis found that while both song and speech sustained particular fundamental frequencies for long, steady bouts, speech fundamental frequencies were lower in pitch and more restricted in range than singing fundamental frequencies.

Finally, the human non-linguistic space was significantly larger than that of all other vocalization types by at least a factor of two (randomization test, [Fig. 2, A, B](#); [Table 3](#)). This could be partly due to innovation in the vocal periphery. The human loss of supralaryngeal air sacs and vocal membranes along with the acquisition of a permanently descended larynx and tongue have often been proposed to enable speech or make it more intelligible ([De Boer, 2012](#); [Nishimura et al., 2022](#)). However, changes in the vocal organ and tract would affect all utterances. The first appearance of these shifts could equally have marked an increase in the diversity of human *non-linguistic* vocal sounds produced in new social contexts. An expanded non-linguistic repertoire, partially enabled by peripheral innovations and possibly concomitant with expansions in social-behavioral repertoire, may have preceded the evolutionary origin of spoken language. Social and communicative complexity have long been proposed to correlate, and vocal repertoire size appears to track with social bonding intensity within primates ([Freeberg et al., 2012](#); [McComb & Semple, 2005](#); [Peckre et al., 2019](#); [Sewall, 2015](#)). This hypothesis requires further testing, however; the human non-linguistic space was better sampled than the non-human primate spaces in our study. This is especially true for bonobos and chimpanzees, which had a significantly smaller number of recordings in their respective samples.

Quantitative comparisons of vocal output have greatly informed the debate on speech origins, but prior work has focused on speech-like utterances. Exploration of the diversity of non-linguistic sounds in humans is relevant to the evolution of hominin vocal behavior.

Estimating vocal output range by collecting the broadest possible sample of vocal elements (Anikin et al., 2023 [DOI](#); Odom et al., 2021 [DOI](#)) may lead to new areas of inquiry. We can better understand the evolution of vocal communication, including speech, by further embedding humans in comparative datasets that capture more inclusive samples of each taxon's vocal output.

Data availability

Acoustic-feature data and example recordings have been deposited at Open Science Framework. A public DOI will be made available as of the date of publication. Data can currently be accessed at https://osf.io/e8vsr/?view_only=1badad869e2b439da63fc8e1139fdc68 [DOI](#). All original code has been deposited at Open Science Framework. A public DOI will be made available as of the date of publication. Code can currently be accessed at https://osf.io/e8vsr/?view_only=1badad869e2b439da63fc8e1139fdc68 [DOI](#). Any additional information required to reanalyze the data reported in this paper is available from the corresponding author upon request.

Acknowledgements

The authors thank the late F. de Waal, The Macaulay Library at the Cornell Lab of Ornithology, The International Phonetic Association, the Bonobo Conservation Initiative, A. Anikin et al., A.S. Cohen et al., Bergevin et al., P. Savage, S. Mehr et al., and Z. Clay for supplying recordings; D. Spindler and T. Hess for statistical computing assistance; B. Blonder and D. Chen for guidance on probabilistic hypervolume analysis; A. Vickers for statistical guidance; and M. Knörnschild, B. Sealey, A.L. Baier, L. Legendre, and M. Hauser for support and comments. This work was completed in partial fulfillment of the requirements for a PhD at The University of Texas.

Additional information

Author contributions

Conceptualization: HTB

Methodology: HTB, JAC, MJR

Data curation: HTB

Formal analysis: HTB

Investigation: HTB, JAC, MJR

Visualization: HTB

Funding acquisition: HTB

Project administration: HTB

Writing – original draft: HTB

Writing – review & editing: JAC, MJR

Supervision: JAC, MJR

Funding

Funder	Grant reference number	Author
NSF National Science Foundation Graduate Research Fellowship Program (GRFP)	Fellow ID 2018262669	Hans Bilger

Author ORCID iDs

Hans T Bilger:  <https://orcid.org/0000-0001-8545-3065>

Additional files

[Supplementary file 1](#) 

References

- Anikin A**, Canessa-Pollard V, Pisanski K, Massenet M, Reby D (2023) Beyond speech: Exploring diversity in the human voice. *Isience* **26** <https://doi.org/10.1016/j.isci.2023.108204> | [PubMed](#)
- Anikin A**, Persson T (2017) Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior Research Methods* **49**:758-771 <https://doi.org/10.3758/s13428-016-0736-y> | [PubMed](#)
- Muse Group**, contributors (2021) Audacity(R): Free Audio Editor and Recorder. version: 2.4.2 <https://audacityteam.org/>
- Bergevin C**, Narayan C, Williams J, Mhatre N, Steeves JK, Bernstein JG, Story B (2020) Overtone focusing in biphonic Tuvan throat singing. *eLife* **9**:e50476 <https://doi.org/10.7554/eLife.50476> | [PubMed](#)
- Blonder B**, Morrow CB, Harris DJ, Brown S, Butruille G, Laini A, Chen D (2022) hypervolume: High Dimensional Geometry, Set Operations, Projection, and Inference Using Kernel Density Estimation, Support Vector Machines, and Convex Hulls. CRAN. version: 3.0.1 <https://doi.org/10.32614/CRAN.package.hypervolume>
- Blonder B**, Morrow CB, Maitner B, Harris DJ, Lamanna C, Violle C, Enquist BJ, Kerkhoff AJ (2018) New approaches for delineating n-dimensional hypervolumes. *Methods in Ecology and Evolution* **9**:305-319 <https://doi.org/10.1111/2041-210X.12865>
- Boë L.-J.**, 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 <https://doi.org/10.1371/journal.pone.0169321> | [PubMed](#)
- Brenowitz EA**, Beecher MD (2023) An ecological and neurobiological perspective on the evolution of vocal learning. *Frontiers in Ecology and Evolution* **11**:1193903 <https://doi.org/10.3389/fevo.2023.1193903>
- Cauzinille J**, Favre B, Marxer R, Rey A (2024) Applying machine learning to primate bioacoustics: Review and perspectives. *American Journal of Primatology* **86**:e23666 <https://doi.org/10.1002/ajp.23666> | [PubMed](#)
- Clay Z** (2022) What do Bonobos Sound Like? Bonobo Conservation Initiative. The Bonobo Conservation Initiative. <https://www.bonobo.org/bonobos/what-do-bonobos-sound-like>
- Clink DJ**, Crofoot MC, Marshall AJ (2019) Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia. *Bioacoustics* **28**:193-209 <https://doi.org/10.1080/09524622.2018.1426042>
- Coupé C**, Oh YM, Dediu D, Pellegrino F (2019) Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* **5**:eaaw2594 <https://doi.org/10.1126/sciadv.aaw2594> | [PubMed](#)
- Cowen AS**, Elfenbein HA, Laukka P, Keltner D (2019) Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist* **74**:698 <https://doi.org/10.1037/amp0000399> | [PubMed](#)
- Darwin C** (1871) *The Descent of Man and Selection in Relation to Sex* Murray.
- Davis S**, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**:357-366 <https://doi.org/10.1109/tassp.1980.1163420>
- De Boer B** (2008) The acoustic role of supralaryngeal air sacs. *Journal of the Acoustical Society of America* **123**:3779-3779 <https://doi.org/10.1121/1.2935420>
- De Boer B** (2012) Loss of air sacs improved hominin speech abilities. *Journal of Human Evolution* **62**:1-6 <https://doi.org/10.1016/j.jhevol.2011.07.007> | [PubMed](#)

- Farnetani E, Recasens D (2010) Coarticulation and Connected Speech Processes. In: Hardcastle WJ, Gibbon FE, Laver J (Eds). *The Handbook of Phonetic Sciences* Blackwell Publishing Ltd.. pp. 316-352 <https://doi.org/10.1002/9781444317251.ch9>
- Fitch WT (2000) The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4:258-267 [https://doi.org/10.1016/s1364-6613\(00\)01494-7](https://doi.org/10.1016/s1364-6613(00)01494-7) | PubMed
- Fitch WT (2013) Musical protolanguage: Darwin's theory of language evolution revisited. In: Bolhuis JJ, Everaert M (Eds). *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain* MIT Press. pp. 489-503 <https://doi.org/10.7551/mitpress/9322.003.0032>
- Fitch WT (2018) The biology and evolution of speech: A comparative analysis. *Annual Review of Linguistics* 4:255-279 <https://doi.org/10.1146/annurev-linguistics-011817-045748>
- Fitch WT, Anikin A, Pisanski K, Valente D, Reby D (2025) Formant analysis of vertebrate vocalizations: Achievements, pitfalls, and promises. *BMC Biology* 23:92 <https://doi.org/10.1186/s12915-025-02188-w> | PubMed
- Fitch WT, De Boer B, Mathur N, Ghazanfar AA (2016) Monkey vocal tracts are speech-ready. *Science Advances* 2:e1600723 <https://doi.org/10.1126/sciadv.1600723> | PubMed
- Freeberg TM, Dunbar RI, Ord TJ (2012) Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367:1785-1801 <https://doi.org/10.1098/rstb.2011.0213> | PubMed
- Habel K, Grasman R, Gramacy RB, Mozharovskiy P, Sterratt DC (2019) geometry: Mesh Generation and Surface Tessellation. CRAN. version: 0.4.5 <https://doi.org/10.32614/CRAN.package.geometry>
- Hewitt G, MacLarnon A, Jones KE (2002) The functions of laryngeal air sacs in primates: A new hypothesis. *Folia Primatologica* 73:70-94 <https://doi.org/10.1159/000064786> | PubMed
- Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97:3099-3111 <https://doi.org/10.1121/1.411872> | PubMed
- Hillenbrand JM, Clark MJ (2009) The role of f_0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics* 71:1150-1166 <https://doi.org/10.3758/app.71.5.1150> | PubMed
- Huron D (2016) *Voice leading: The science behind a musical art* MIT press.
- Hutchinson GE (1957) Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22:415-427 <https://doi.org/10.1101/SQB.1957.022.01.039>
- International Phonetic Association (1999) *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* Cambridge University Press. <https://www.internationalphoneticassociation.org/content/ipa-handbook-downloads>
- iZotope, Inc (2007) iZotope RX 9. version: 9.2.0.1278
- Janik VM, Knörnschild M (2021) Vocal production learning in mammals revisited. *Philosophical Transactions of the Royal Society B* 376:20200244 <https://doi.org/10.1098/rstb.2020.0244> | PubMed
- Jürgens U (2009) The neural control of vocalization in mammals: A review. *Journal of Voice* 23:1-10 <https://doi.org/10.1016/j.jvoice.2007.07.005> | PubMed
- Lakdari MW, Ahmad AH, Sethi S, Bohn GA, Clink DJ (2024) Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. *Ecological Informatics* 80:102457 <https://doi.org/10.1016/j.ecoinf.2023.102457>
- Lee C.-H, Chou C.-H, Han C.-C, Huang R.-Z (2006) Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. *Pattern Recognition Letters* 27:93-101 <https://doi.org/10.1016/j.patrec.2005.07.004>
- Lieberman D (2011) *The Evolution of the Human Head* Harvard University Press.

- Lieberman PH, Klatt DH, Wilson WH (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science* **164**:1185-1187
<https://doi.org/10.1126/science.164.3884.1185> | PubMed
- Ligges U, Krey S, Mersmann O, Schnackenberg S (2018) tuneR: Analysis of Music and Speech. CRAN. version: 1.4.7 <https://doi.org/10.32614/CRAN.package.tuneR>
- McComb K, Semple S (2005) Coevolution of vocal communication and sociality in primates. *Biology Letters* **1**:381-385 <https://doi.org/10.1098/rsbl.2005.0366> | PubMed
- Mehr SA, Singh M, Knox D, Ketter DM, Pickens-Jones D, Atwood S, Lucas C, Jacoby N, Egner AA, Hopkins EJ (2019) Universality and diversity in human song. *Science* **366**:eaax0868
<https://doi.org/10.1126/science.aax0868> | PubMed
- Mielke A, Zuberbühler K (2013) A method for automated individual, species and call type recognition in free-ranging animals. *Animal Behaviour* **86**:475-482 <https://doi.org/10.1016/j.anbehav.2013.04.017>
- Morton ES (1977) On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist* **111**:855-869 <https://doi.org/10.1086/283219>
- Nettl B, Stone R, Porter J, Rice T (1998) *The Garland Encyclopedia of World Music* Garland.
- Nishimura T, Tokuda IT, Miyachi S, Dunn JC, Herbst CT, Ishimura K, Kaneko A, Kinoshita Y, Koda H, Saers JPP, et al. (2022) Evolutionary loss of complexity in human vocal anatomy as an adaptation for speech. *Science* **377**:760-763 <https://doi.org/10.1126/science.abm1574> | PubMed
- Odom KJ, Araya-Salas M, Morano JL, Ligon RA, Leighton GM, Taff CC, Dalziel AH, Billings AC, Germain RR, Pardo M (2021) Comparative bioacoustics: A roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews* **96**:1135-1159 <https://doi.org/10.1111/brv.12695> | PubMed
- Oxenham AJ (2012) Pitch perception. *The Journal of Neuroscience* **32**:13335-13338
<https://doi.org/10.1523/JNEUROSCI.3815-12.2012> | PubMed
- Peckre L, Kappeler PM, Fichtel C (2019) Clarifying and expanding the social complexity hypothesis for communicative complexity. *Behavioral Ecology and Sociobiology* **73**:1-19
<https://doi.org/10.1007/s00265-018-2605-4>
- Pisanski K, Raine J, Reby D (2020) Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries. *Royal Society Open Science* **7**:191642
<https://doi.org/10.1098/rsos.191642> | PubMed
- R Core Team (2022) R: A language and environment for statistical computing. version: 4.5.1
<http://www.R-project.org/>
- Sewall KB (2015) Social complexity as a driver of communication and cognition. *Integrative and Comparative Biology* **55**:384-395 <https://doi.org/10.1093/icb/icv064> | PubMed
- Soe Naing HM, Hidayat R, Hartanto R (2020) Discrete wavelet denoising into MFCC for noise suppressive in automatic speech recognition system. *International Journal of Intelligent Engineering & Systems* **13** <https://doi.org/10.22266/ijies2020.0430.08>
- Sueur J, Aubin T, Simonis C (2008) Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics* **18**:213-226 <https://doi.org/10.1080/09524622.2008.9753600>
- Titze IR, Palaparthi A (2018) Radiation efficiency for long-range vocal communication in mammals and birds. *The Journal of the Acoustical Society of America* **143**:2813-2824
<https://doi.org/10.1121/1.5034768> | PubMed
- Wadewitz P, Hammerschmidt K, Battaglia D, Witt A, Wolf F, Fischer J (2015) Characterizing vocal repertoires—Hard vs. Soft classification approaches. *PloS One* **10**:e0125785
<https://doi.org/10.1371/journal.pone.0125785> | PubMed
- Zipf GK (1949) *Human behavior and the principle of least effort: An introduction to human ecology* Addison-Wesley Press.
- Bilger HT, Ryan MJ, Clarke JA (2026) Acoustic-feature data and example recordings. OSF. ID e8vsr
https://osf.io/e8vsr/?view_only=1badad869e2b439da63fc8e1139fdc68

Bergevin C, Narayan C, Williams J, Mhatre N, Steeves JK, Bernstein JG, Story B (2020) Overtone focusing in biphonic Tuvan throat singing. Dryad. <https://doi.org/10.5061/dryad.cvdncjt14>

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors conducted a comparative acoustic analysis of primate vocal repertoires, focusing on the assumption that speech and language evolution required and involved an expansion in the acoustic space of voiced vocalizations from non-human primates to humans. Results challenge this idea. The study compiles and analyzes a large dataset of calls to quantify differences in vocal production space.

Strengths:

The study is technically sound, with a solid implementation of acoustic measurements and a valuable new dataset that brings empirical rigor to test a dominant, yet hitherto strictly theoretical, notion about what speech and language evolution entailed. It provides concrete comparative acoustic data across species to disprove that speech and language required an increase in the range of voiced calls, and thus, by extension, of vowels. The approach is methodologically rigorous and directly engages with the relevant data, rather than relying on untested presumptions of what great apes "ought" to be able to do or not.

Weaknesses:

The theoretical contextualization should be strengthened and updated, as several aspects contain inaccuracies, most notably by equating voiced calls or vocalizations with speech (overlooking the critical role of consonants, as human languages typically show vowel:consonant ratios of 1:4 or greater) and misrepresenting the premises and current status of the neural (Kuypers-Jürgens) hypothesis.

The discussion drifts into speculative territory on features like syntax and co-articulation that fall outside the paper's scope and data, and it does not sufficiently engage recent evidence on vocal learning and consonant-like capacities in great apes.

Minor issues include incomplete sampling justifications, imprecise terminology, and reliance on references that have been critiqued in more recent work.

<https://doi.org/10.7554/eLife.112153.1.sa1>

Reviewer #2 (Public review):

This study examines the evolutionary context of the emergence of human speech. The authors address the widely held hypothesis that the expansion of the human vocal space, resulting from modifications of the vocal tract, was a key prerequisite for the evolution of spoken language.

To test this hypothesis, the authors quantified the acoustic space of human speech, non-linguistic vocalizations, and musical vocalizations and compared it with that of nonhuman primates, chimpanzees, bonobos, and chacma baboons.

The authors found that speech and song occupied significantly less volume in the acoustic space than human non-linguistic vocalizations. In addition, the acoustic-feature volume of speech and song was not statistically distinct from that of non-human primates. Accordingly,

the authors conclude that the evolution of human speech did not depend on an expansion of the human vocal acoustic space.

I find the analysis presented in this manuscript highly convincing. It is conducted at a contemporary scientific standard, and the results provide strong support for the authors' conclusions. I particularly appreciate that the authors explicitly discuss the limitations of their approach. For example, they acknowledge that MFCCs cannot capture all aspects of acoustic structure.

I have only three minor comments:

First, the authors may wish to briefly summarize the main findings of the study by Anikin et al., as it represents the central reference for the present work. A concise summary in two or three sentences would help readers who are not familiar with that study.

Second, I would appreciate a brief explanation of why the authors chose this particular statistical approach.

Third, the authors could briefly mention that the Chacma baboon dataset provides a very comprehensive representation of the vocal repertoire of this species, although a small number of rare vocalizations are not included. I am not sure whether a similar limitation also applies to the chimpanzee and bonobo datasets, but if so, it would be useful to mention this as well.

<https://doi.org/10.7554/eLife.112153.1.sa0>