Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/158212

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
Orangutan information broadcast via consonant-like and vowel-like calls breaches mathematical models of linguistic evolution

Adriano R. Lameira¹²*, António Alexandre³, Marco Gamba⁴, Matthew G. Nowak⁵⁶, Raquel Vicente³, Serge Wich⁷⁸

¹Department of Psychology, University of Warwick, Coventry, UK
²School of Psychology and Neuroscience, University of St Andrews, Scotland, UK
³Independent researcher
⁴University of Turin, Department of Life Sciences and Systems Biology, Turin, Italy
⁵Sumatran Orangutan Research Program (PanEco-YEL), North Sumatra, Indonesia
⁶Department of Anthropology, Southern Illinois University, Illinois, USA
⁷School of Natural Sciences and Psychology, Liverpool John Moores University, UK
⁸Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, the Netherlands

*Corresponding author: adriano.lameira@warwick.ac.uk

Keywords: Language origin; language evolution; proto-consonants; proto-vowels; great apes; orangutans (Pongo ssp)
Abstract

The origin of language is one of the most significant evolutionary milestones of life on Earth, but one of the most persevering scientific unknowns. Two decades ago, game theorists and mathematicians predicted that the first words and grammar emerged as a response to transmission errors and information loss in language’s precursor system, however, empirical proof is lacking. Here, we assessed information loss in proto-consonants and proto-vowels in human pre-linguistic ancestors as proxied by orangutan consonant-like and vowel-like calls that compose syllable-like combinations. We played-back and re-recorded calls at increasing distances across a structurally complex habitat (i.e. adverse to sound transmission). Consonant-like and vowel-like calls degraded acoustically over distance, but no information loss was detected regarding three distinct classes of information (viz. individual ID, context and population ID). Our results refute prevailing mathematical predictions and herald a turning point in language evolution theory and heuristics. Namely, explaining how the vocal-verbal continuum was crossed in the hominid family will benefit from future mathematical and computational models that, in order to enjoy empirical validity and superior explanatory power, will be informed by great ape behavior and repertoire.

Introduction

Communication in natural (e.g. human language) and artificial systems (e.g. computer language) rests on three vertices: the encoder, the decoder, and the communication channel linking the two\textsuperscript{1}. With regards to language origin – the last major evolutionary transition of life on Earth\textsuperscript{2} – much attention has been dedicated to the role of the encoder (its anatomical\textsuperscript{3–6} and motoric attributes\textsuperscript{7–11}), the receiver (its anatomical\textsuperscript{12,13} and perceptual attributes\textsuperscript{14–18}) and the interactions between the two\textsuperscript{19}. Surprisingly, however, the role of the channel\textsuperscript{1} – the interval between encoder and decoder that a signal must traverse – in the emergence of language has remained virtually ignored\textsuperscript{20}.

This knowledge gap is particularly problematical in light of game theory and mathematical models of language evolution\textsuperscript{21–23}. Notably, these models have predicted that the first words and grammatical rules emerged to minimize error and information loss in language’s precursor channel. Regarding word origin, this argument asserts that the lengthier a signal combination, the lower the probability of mistaking signals for each other. Regarding syntax origin, it asserts that the more varied a sequence of signal combinations, the lower the probability of mistaking the events being referred to, with words and syntax having, thus, developed in the human lineage to decrease transmission errors. Without basic knowledge about the communication channel used by our
ancestors to broadcast information and its "error limit"\textsuperscript{21–23}, it is impossible, however, to validate these models or their proposed evolutionary scenario.

Human evolution unfolded in parallel with acute climate and ecological changes in the African continent\textsuperscript{24}, however, it is unclear when and where the first forms of language manifested among human ancestors. Regardless of whether proto-language originated in the rainforest, woodland or savannah, the hypothesis that the first linguistic structures emerged to avert error can be best tested in forested habitats, which pose the most adverse conditions to sound transmission, and thus, where signal and information limits can be assessed.

To implement an the empirical proof of the currently prevailing mathematical models of linguistic evolution, we assessed information loss in wild orangutan voiceless consonant-like and voiced vowel-like calls\textsuperscript{7}. These calls exhibit articulatorily homology with their human counterparts, and therefore, represent living proxies of spoken language’s putative pre-linguistic units\textsuperscript{25–27}. Namely, we played-back consonant-like kiss-squeaks and vowel-like grumphs\textsuperscript{28} and re-recorded these calls at increasing distances. Critically, bar humans, orangutans are the only known great ape to produce consonant-like and vowel-like calls combined into syllable-like combinations\textsuperscript{29}, therefore, presenting a privileged hominid model for this study\textsuperscript{30}.

\section*{Materials and Methods}

\textit{In brief}

Calls were originally recorded from wild orangutan individuals across contexts and populations of Sumatran (\textit{Pongo abelii}) and Bornean orangutans (\textit{P. pygmaeus}). Only consonant- and vowel-like calls that were prevenient from the same syllable-like combination were used for playback. We extracted four acoustic paraments over distance. We used individual, contextual and geographical acoustic signatures\textsuperscript{25} to assess information loss. This setup mimicked the putative proto-combinatoric conditions at the moment of language origin. Methodologically, this allowed to control for biasing factors between consonant- and vowel-like calls (e.g. individuals, context, recording settings).

\textit{Study site}

Playback experiments were conducted at the Sikundur Research Station (3°55′48.07″; 98°2′31.17″), Leuser Ecosystem, North Sumatra, Indonesia. The Sikundur forest is located on the eastern forest margin of the Alas river dividing the Leuser Ecosystem along its North-South axis and constituting a major barrier dispersal barrier for orangutans at this altitude\textsuperscript{31}. Presently, the forest is a dipterocarp tropical rainforest, comprised of disturbed primary forest and secondary/regrowth forest that was the target of previous logging operations (between 1970 and
1980, and later during 1990s. Research at the station is managed by the Sumatran Orangutan Conservation Programme (SOCP)–PanEco Foundation. The study was performed in agreement with regulations and permissions from the relevant Indonesian authorities. No animal observation, handling, contact or interaction took place during this study.

Data collection

Recordings for the playback playlist were previously collected at three research stations: Tuanan and Gunung Palung (Central and West Kalimantan, respectively, Indonesian Borneo) and Sampan Getek (North Sumatra, Indonesia). The playback playlist included 120, 118 and 249 calls to assess individual ID, context and population ID information, respectively (see more in ESM). Orangutan kiss squeaks were used as living proxies of voiceless proto-consonants, orangutan grumphs as living proxies of voiced proto-vowels.

All kiss-squeaks and grumphs were selected from call combinations composed of the two calls, specifically kiss-squeak+grumph (see Data Analyses and ESM). All recordings were set to the same peak amplitude prior to playback using Raven interactive sound analysis (version 1.2.1, Cornell Lab of Ornithology, Ithaca, New York). No further signal transformations were conducted.

Playbacks were conducted using a Marantz Digital Recorder PMD-660 (D&M Holdings, Kawasaki, Japan) connected to a Nagra DSM speaker (Audio Technology Switzerland S.A., Romanel, Switzerland). The speaker was set at 1 - 1.5m from the ground. Because Sikundur is partially a regrowth/secondary forest, with abundant undergrowth below the understory, this height offered a suitable means to explore the effects of complex habitat structure on broadcast performance. Playback volume was set at ~100dB SPL at 1-meter distance to facilitate assessment of sound degradation over distance and was not meant to emulate orangutan natural vocal loudness. Playbacks were conducted between 5h30 and 6h30 local time in absence of wind and with no rain during the previous 48 hours. This time was elected for playbacks because in this habitat, early mornings presented the time of day with least biotic noise. We made no presumptions on whether early human ancestors communicated predominantly at this time. All recordings along the same transect were conducted in the same morning.

Playbacks were conducted twice at two locations (i.e. along two transects), once at each location. Re-recordings were conducted every 25m along the two transects across the forest up until 100m away, at which point playbacks became too faint to be analysed. Transects started within 10m from each other and advanced forward in oblique direction one from other. Using different transects allowed to assess the impact of particular phonological features (e.g. larger tree trucks, leaf density) on broadcast performance. Transects were straight, flat and included no obvious canopy openings or clearings. Playbacks were re-recorded using ZOOM H4next Handy Recorder (ZOOM Corporation, Tokyo, Japan) connected to a RØDE NTG-2 directional microphone.
(RÖDE LLC, Sydney, Australia). Audio data were recorded using the WAVE PCM format at 16-bit. The microphone was set at 1 - 1.5m from the ground. Data for distance zero were extracted from the original playback recordings. In total, 7826 calls (incl. original at 0m and re-recordings up to 100m) were collected (see ESM for sample breakdown). For each transect, three playbacks sessions were conducted, one for each information type: one playlist comprised recordings varying in individual subjects, the other in context and the other in population.

**Data measurements**

We manually measured four acoustic parameters from all calls using Raven interactive sound analysis (version 1.2.1, Cornell Lab of Ornithology, Ithaca, New York) using the spectrogram window (window type: Hann; 3-dB filter bandwidth: 124 Hz; grid frequency resolution: 2.69 Hz; grid time resolution: 256 samples): duration (s), maximum frequency (Hz), maximum power (uncalibrated dB), and maximum time. Duration was the time difference between call offset and onset. Maximum frequency was the frequency with maximum energy (i.e. power, dB) in a call. Maximum power was the power of the maximum frequency. Maximum time was the moment when the maximum power occurred proportionally to the total duration of a call (e.g. max time=0.5 means it occurred half way the call’s duration). These parameters have been found to be strong descriptors of orangutan calls and their informational content\textsuperscript{25,28,33}. Critically, they were extractable from both consonant- and vowel-calls, enabling direct comparison between acoustic and information broadcast performance between the two call categories.

**Data analyses – Acoustic performance**

To assess acoustic broadcast performance during transmission, Linear Mixed Models (model type: III sum of squares; test model terms: Satterthwaite, using restricted maximum likelihood) were conducted using JASP\textsuperscript{34} (version 0.14.1). One model was generated per acoustic parameter (x4) per call type (x2), with a total of 8 models. Per model, the acoustic parameter was inserted as dependent variable (N=3560 per call type). Distance (treated as ordinal: 0, 25, 50, 75, 100m), transect (2 levels), context (3 levels: towards human observers, tiger-patterned predator-model, plain-white predator-model)\textsuperscript{29} and population (3 levels: Tuanan, Gunung Palung, Sampan Getek) were inserted as fixed effect variables. Individual (20 levels) and call number (N=249 per call type) were inserted as random effect, since some calls were re-used for different playbacks and from the same individual. Random slopes for distance and transect were allowed to vary per individual. No explicit indication of nested variables (e.g. individual within population) was provided since this is automatically identified by the model (see\textsuperscript{25} and ESM).

**Data analyses – Information performance**
To assess information broadcast performance, we conducted discriminant function analyses (DFA) per distance\textsuperscript{33}. All analyses were based on the four measured acoustic parameters simultaneously. Six analyses were conducted to test information content (x3; individual ID, context, population ID) for each call type (x2). LMM results indicated that “transect” had a significant effect acoustic performance over distance, hence, all (p)DFA analyses were conducted using one transect only. We conducted DFA with leave-one-out procedure using SPSS (IBM SPSS Statistics, version 27; ESM) to assess information content about individual identify (same context used across individuals). To assess information content about context and population, we performed permuted DFAs (pDFA) with cross-classification\textsuperscript{35}: crossed pDFA for context (to control for individual variation) and nested pDFA for population (individual variation nested within population; ESM). pDFA was conducted in R\textsuperscript{36} with MASS\textsuperscript{37} and using a function provided by R. Mundry\textsuperscript{35}. Because crossed pDFAs do not tolerate null data, only three individuals with calls in all contexts were included. Figures were prepared using ggplot2\textsuperscript{38} and gridExtra\textsuperscript{39}. A script example was:

\begin{verbatim}
pdfa.res=pDFA.crossed(test.fac="Context", contr.fac="Individual", variables=c("Duration", "Max frequency", "Max time", "Max power"), n.to.sel=NULL, n.sel=100, n.perm=1000, pdfa.data=test.data).
\end{verbatim}

**Results**

*Acoustic performance over distance*

Consonant-like and vowel-like call acoustic parameters changed significantly during transmission (Table 1, Fig. 1, ESM). This was expected since different parameters interact differentially with the environment (e.g. max power declines over distance following the general inverse square law of sound attenuation). Several significant differences were found between transects (ESM), confirming that acoustic performance was (partly) dictated by the physical structure of the transmission channel. Context had a significant effect on the acoustic performance of some parameters (ESM). Given that both call types are known to exhibit marked contextual variation\textsuperscript{25}, this shows that the acoustic features of different contextual sub-types affect how their transmission plays out. For both consonant-like and vowel-like calls, population had a significant effect on some acoustic parameters (ESM), suggesting that geographic accents\textsuperscript{25} may endow calls with better transmission properties. Given that forest structure is no longer pristine across virtually all orangutan sites, it is unclear whether these gains can be attributed to adaptive selection in some populations.

<table>
<thead>
<tr>
<th>Consonant-like calls (Kiss-squeaks)</th>
<th>Vowel-like calls (Grumps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>F</td>
</tr>
</tbody>
</table>


Despite poor acoustic performance, informational performance of consonant- and vowel-like calls was not affected during transmission (Fig. 2). Both call categories allowed correct assessment of information about individual identity, context and population well above chance levels (Fig. 2). Information loss was only observed for individual identity when transmitted by vowel-like calls, however, this effect was only observed when computing a leave-one-out DFA procedure (a more stringent model) and information performance remained overall above chance (Table 2; ESM). Information performance was equivalent between consonant- and vowel-like calls; their trend lines remained relatively parallel over distance (Fig. 2). Consonant-like calls tended to exhibit higher percentage of correct assignments, suggesting heavier information load (Fig. 2).

<table>
<thead>
<tr>
<th>Information performance over distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Despite poor acoustic performance, informational performance of consonant- and vowel-like calls was not affected during transmission (Fig. 2). Both call categories allowed correct assessment of information about individual identity, context and population well above chance levels (Fig. 2). Information loss was only observed for individual identity when transmitted by vowel-like calls, however, this effect was only observed when computing a leave-one-out DFA procedure (a more stringent model) and information performance remained overall above chance (Table 2; ESM). Information performance was equivalent between consonant- and vowel-like calls; their trend lines remained relatively parallel over distance (Fig. 2). Consonant-like calls tended to exhibit higher percentage of correct assignments, suggesting heavier information load (Fig. 2).</td>
</tr>
</tbody>
</table>

Table 2. Information performance over distance: Spearman’s Correlation Summary (n=5)

<table>
<thead>
<tr>
<th>Consonant-like calls (Kiss-squeks)</th>
<th>Vowel-like calls (Grumphs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual</td>
</tr>
<tr>
<td></td>
<td>norm</td>
</tr>
<tr>
<td>Spearman’s rho</td>
<td>-0.9</td>
</tr>
<tr>
<td>p</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Discussion
We found no evidence for information loss in the only nonhuman living hominid that combines consonant-like and vowel-like calls to produce syllable-like combinations. Information content remained uncompromised until either call type become inaudible, indicating that homologous proto-linguistic units would have remained functionally discriminable as long as they could be heard. Results refute, therefore, mathematical predictions for linguistic evolution.

Orangutan consonant-like calls exhibited extreme spectral differences compared with their vowel-like counterparts (i.e. frequency centered at ~4000 vs. 250Hz, respectively, Fig. 1A, 1D). However, both can be information-dense and their information performance was equivalent. This suggests that similar results would have been likely when other nonhuman hominid consonant- and vowel-like calls had been selected. Our analyses covered a wide frequency band wherein the actual (but now extinct) proto-linguistic units of language have probably laid.

Information loss was assessed by measuring calls’ biometric information content (i.e. about individual ID, context and population ID). There is no evidence that other types of informational content (e.g. culturally conventionalized arbitrary information, such as, a word’s meaning) transmit differently via the same acoustic signals. Some orangutan consonant-like calls exhibit arbitrary function and other great ape consonant-like and vowel-like calls are transmitted culturally. Thus, these calls are not unescapably limited to the transmission of biometric information, even though this was the information used for our empirical validation.

Findings offer three insights into language origin and linguistic evolution. First, proto-consonants and -vowels encoded ample information and were resilient against information loss up to 100m distance across channels adverse to signal transmission.

Second, the structural complexity of our first linguistic ancestors’ habitat was an unlikely source of transmission error and information loss. Paleo-climate change across African habitats brought about major habitat structural changes, and with then, new soundscapes. Open habitats offer few physical obstructions to signal transmission (e.g. savannah), thus, ecological changes happening across Africa are predicted to have diminished channel noise in language’s precursor system, not the opposite. Systematic assessment will be required for conclusive resolution.

Third, mathematical and computational approaches to language evolution have not, thus far, explicitly or implicitly modeled hominid behavior. Theoretically, current models could apply to any communication system transitioning to a combinatorial state, not necessarily within the hominid family. The fact that language transpired in the human clade, but none other, implies, thus, that “being a hominid” cannot be discounted from theoretical incursions that might stand a chance to enlighten how linguistic evolution ensued from the repertoire of an ape-like ancestor. While current models assuredly encapsulate a possible evolutionary scenario, this was not the one to have likely catalyzed language. The most beneficial future theoretical models will be those that
conform with, and factor in, the (consonant-vowel-based) combinatorics shared between great apes and humans.
Acknowledgments

We thank the Indonesian RISTEK, PHKA, TNGL, BPKE, Universitas Sumatran Utara and SOCP–PanEco Foundation. We thank Graham Usher for technical support and two anonymous reviews for insightful comments. A.R.L. was supported by the European Union’s Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie grant agreement no. 702137 and the UK Research and Innovation’s Future Leaders Fellowship grant agreement no. MR/T04229X/1.

References


13. Arora, N. et al. Effects of Pleistocene glaciations and rivers on the population structure of Bornean
(2010).


33. Lameira, A. R. & Wich, S. Orangutan Long Call Degradation and Individuality Over Distance: A

34. JASP Team (2020). JASP (v. 0.14.1) [Computer software].

35. Mundry, R. & Sommer, C. Discriminant function analysis with nonindependent data: consequences


40. Lameira, A. R. et al. Population-specific use of the same tool-assisted alarm call between two wild
orangutan populations (Pongo pygmaeus wurmbii) indicates functional arbitrariness. *Plos One* **8**, 
e69749 (2013).


42. Taglialatela, J. P., Reamer, L., Schapiro, S. J. & Hopkins, W. D. Social learning of a communicative

43. Russell, J. L., Joseph, M., Hopkins, W. D. & Taglialatela, J. P. Vocal learning of a communicative signal

64 (2009).


Figure 1. Spectrographic representation of orangutan consonant-like and vowel-like calls (A), graphic representation of the experimental set up (B) and acoustic performance during transmission (C-F; based on raw data). uncal.: uncalibrated. Box plots represent median and 25-75% interquartile range, whiskers represent lowest/highest value within 1.5 times interquartile range below/above, outliers omitted for clarity. Linear trend lines represented across distance are for visual aid only (based on raw data). *: p<0.001 (LMM ANOVA; see Table 1).

Figure 2. Graphic representation of information performance of orangutan consonant-like and vowel-like calls during transmission, as measured by percentage of correctly assigned cases over distance. Black dotted lines: chance level. A: Continuous lines: Correctly classified cases (DFA); Dashed lines: Correctly classified cross-validated cases (DFA Leave-one-out). B and C: Continuous lines: Correctly cross classified cases (pDFA); Dashed lines: Correctly classified selected cases (pDFA). *: p<0.05 (Spearman’s correlation; see Table 2).