PREQUEL: SUPERVISED PHONETIC APPROACHES TO ANALYSES OF GREAT APE QUASI-VOWELS

Axel G. Ekström[1], Steven Moran[2,3], Johan Sundberg[1,4], Adriano R. Lameira[5]

[1]KTH Royal Institute of Technology, [2]University of Neuchâtel, [3]University of Miami,
[4]University College of Music Education Stockholm, [5]University of Warwick
axeleks@kth.se, steven.moran@unine.ch, jsu@kth.se, adriano.lameira@warwick.ac.uk

ABSTRACT

There is renewed interest in potential vowel production by nonhuman primates, but no agreed-upon methodologies for its estimation from real-life vocalizations. Here, we present a set of supervised approaches for estimating primate vowel-like articulation, with reference to orangutan long call pulses (N=36). We summarize our approach as a cohesive framework, the Primate Quasi-Vowel (PREQUEL) protocol. We *(1)* estimated f_0 from correlograms, *(2)* and vocal tract resonances (formants) from spectrograms, *(3)* the results of which were then compared against synthesized vowels for those frequency values; and *(4)* presented to uninformed listeners (N=16), who largely agreed on the categorization of vowel-like qualities for vocalizations (Cronbach's alpha=.701). We also provide descriptions of methods that are seemingly inadequate for formant estimation in great ape calls. We argue that a combination of phonetic methods is required to develop a science of nonhuman primate articulation.

Keywords: Primatology, Speech acoustics, Vowel space, Evolution of speech, Bioacoustics.

1. INTRODUCTION

Vowels are central to all of the world's spoken languages. Because aspects of human vocal anatomy is evolutionarily old and shared with primates and other mammals [1, 2, 3, 4], it is conceivable that comparative analyses of voiced calls by nonhuman great apes (hereafter great apes) – our closest living relatives – could inform knowledge of evolutionary origins of human vowel production capacities [5, 6, 7, 8]. Empirically, this approach has, however, grappled with major pitfalls, delaying insight towards a clearer picture of speech evolution. These include ethical and practical difficulties (e.g., apes cannot report their experience, or undergo imaging technology methods without some level of anesthesia), which may be circumvented non-invasively by approaching and analyzing great ape voiced calls as "acoustic casts" of vocal tract shape. However, because of the highly variable fundamental frequencies (f_0) characterizing various great ape calls, harmonic partials of f_0 may easily be mistaken for resonances [9] (formants) when estimations are calculated via unsupervised methods. Here, we present in detail a new approach and showcase its potential by confidently identifying [u]-like vocalizations by wild orangutans. We dub our methodological approach the primate quasi vowel (PREQUEL) protocol.

2. DATA

For the present study, we analyzed six long call pulses produced by each of six flanged male orangutans in the wild (N=36) [10] (duration M=.81s). Loud (long-distance) calls, including orangutan long calls and chimpanzee pant hoots, are a shared trait between primates that may date to the last common ancestor of primate species (around 65 Mya). In addition, due to their conspicuousness, they tend to be the best studied components of primate call repertoires. Our approach can, thus, potentially be extended to a variety of primate species. Among great apes, orangutan long calls are particularly slow paced (compared to e.g., chimpanzee pant hoots) and produced solitarily (i.e., without interference from other individuals' calls), and thus constitute a strong candidate for researchers seeking to develop methods of analysis of ape vowel-like production. We adopt the terminology of phonetics, referring to apparent vocal tract resonances as formants.

3. FUNDAMENTAL FREQUENCY

Orangutan long calls typically consist of three acoustically distinguishable phases, *(1)* the build-up grumbling phase, *(2)* climax pulses and *(3)* the bubbling tail-off. Because of apparent aperiodicity in *(1)* and *(3)*, only *(2)* was selected for acoustic

analysis.

Fundamental frequency (f_0) reflects the rate of vocal fold oscillation and corresponds to pitch in perception. In our orangutan data, selected pulses were subject to manual pitch estimation using waveform-matching correlograms [10] and verified with syntheses from the *Madde* additive voice synthesizer software [11]. This procedure was employed because high-f_0 signals are widespread in primate vocalization [12, 13] and therefore present a challenge to phonetic analyses, as the risk of biased estimates increases with f_0. Average f_0 of analyzed pulses (N=36) was 296.38Hz (SD=59.41, Min=172.76, Max=378.3), roughly corresponding to a high-pitched human voice, and within the range for a human male tenor singer. Thus, while significantly lower than those those observed in e.g., chimpanzee screams, values observed still raised the possibility of biased formant estimations, if using hand annotation or linear prediction.

4. FORMANT ESTIMATION

Problems with formant estimations are known from acoustic analyses of human speech, where $f_0 >$ 300 may result in formant estimation errors of $\pm$ 60 Hz [14]. To cancel out any such biasing effect of f_0, we initially attempted to estimate formant values by applying inverse filtering (in the *Sopran* software [11]) to all segments, resampled at 16 kHz (to smooth the spectral curve) – the main effect of which is that it allows analysis of components up to 8 kHz, well above the highest relevant spectrum components.

In speech audio analysis, inverse filtering procedures allow for the cancelling out of effects of formants on the radiated sound (the sum of voice source and vocal tract resonances). Two criteria are applied for tuning the inverse filters, each of which corresponds to a formant, a ripple-free closed phase in the waveform, and a source spectrum envelope void of local peaks and troughs near the formants (e.g., [15]). For our orangutan data, however, because inverse filtering assumes signal periodicity, satisfactory accuracy in tuning of inverse filters was not possible. Additionally, we observed that orangutan calls are breathy to the ear, possibly reflecting incomplete glottal closure in the vibration. The present work is exploratory; these initial observations indicate that a different approach may be necessary to analyze high-frequency ape calls (although higher-quality recordings may also help resolve this issue).

Instead, first and second formant frequencies (F_1, F_2) were annotated by hand; higher resonances were typically unavailable due to low signal-to-noise ratios for frequencies $>$ 2 kHz. In analyses of human voices, narrow-band analyses (45 Hz) are typically applied to reveal harmonic structure, while wide-band analyses (300 Hz) are generally used to reveal formant structures. This is so because narrow-bandwidth analysis admits one voice harmonic at a time, while wider passbands admit multiple voice harmonics at a time, rendering formants as remaining energy peaks (Fig. 1) (although higher-pitched voices may require spectrogram bandwidths at ~600 Hz for intelligible rendering of formants [16, 17]). Because our data was collected in the wild, and includes background noise at higher frequencies, spectrogram bandwidths were set to 300 Hz. F_1-F_2 dispersions and were found to largely overlap with approximate human vowels [u] and [ʊ], with M=338 (SD=55.59) for F_1, and M=969 (SD=163.56) for F_2 (Fig. 2) (Tab. 1). Next, to validate estimated formant values, we sought to investigate whether the calls were indeed perceived as appropriate vowel qualities by human listeners.

Vowel quality	F_1(Hz)	F_2(Hz)
Orangutan long call	338	969
[u]	310	870
[ʊ]	450	1030
[ɔ]	590	880
[o]	360	640
[ɤ]	460	1310
[ɯ]	300	1390

Table 1: Average formant values observed for long calls, and approximate formant values for a set of adjacent (in apparent vowel quality) human vowels (male speaker) [18, 19].

To diminish the risk of mistaking harmonic partials of the f_0 for formant frequencies [9, 14], each candidate formant identified in the previous step was compared with frequencies of f_0 harmonic partials. Values were compared with output from the *Madde* software: each set of F_1-F_2 dispersions were input into *Madde* (set to exclude additional formants, and matched for pitch), to ensure that annotated quasi vowel qualities were adequate. Note that this method assumes that segment f_0 is sufficiently low that it is does not interfere with vowel perception. Next, we sought to investigate whether the identified [u]-like vocalizations were perceived as such by uninformed human listeners.

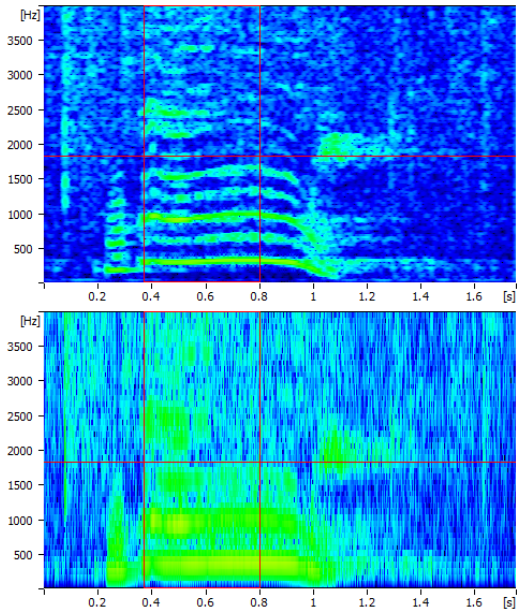

Figure 1: An orangutan long call [10], rendered as a narrow-band spectrogram (above, 45 Hz) with visible harmonic partials of the f_0; and broad-band spectrogram (below, 300 Hz), visibly dissolving partials. Spectrograms were rendered in the *Sopran* software. Length of highlighted segment is 42ms.

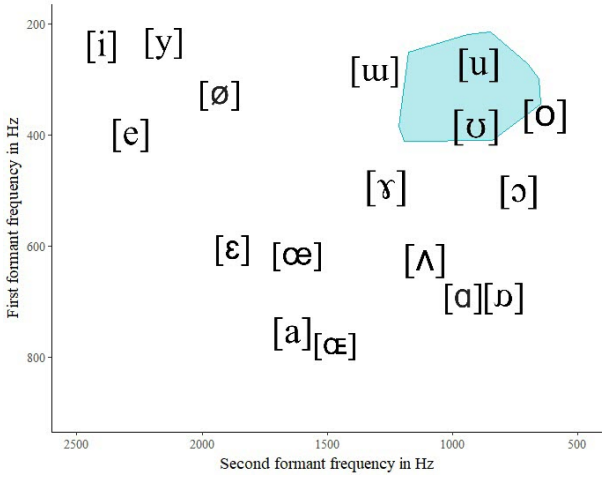

Figure 2: Orangutan [u]-like space. For comparison, average human vowel ranges (adult male speaker) are superimposed in brackets.

5. LISTENING AND SORTING TASKS

Participants (N=16, 9 women and 7 men) were presented with a visual listening sorting task using the visor software [20]. In total, participants listened to two subsets of 8 calls (randomly selected from the larger set) across two presentation sessions (Fig. 3). In the first session, participants were asked

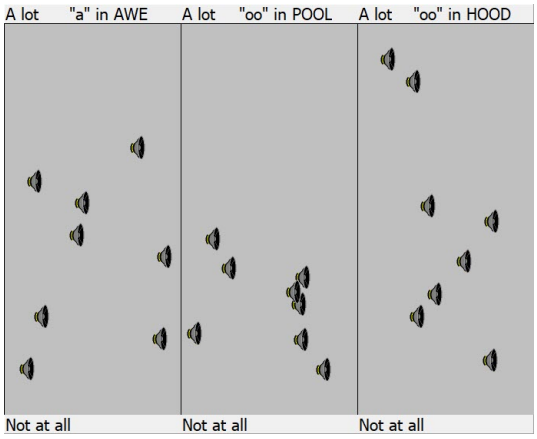

Figure 3: Listening task paradigm, set up in *Visor* software [11]. Files are repeated across each category. Participants freely listen to each sound by clicking and sort them by dragging. Starting positions are random.

to rate each of the eight calls on two dimensions, "sounds like AAH" and "sounds like OO". In the second, they were asked to rate each of 8 (different) calls on three dimensions, where the dimensions corresponded to the degree to which each sound, "sounds like 'a' in AWE [ɔ] / 'oo' in POOL [u] / 'oo' in HOOD [ʊ]". The origin of the recordings and true purpose of the ratings were withheld from transcribers until after the task was concluded. Listeners chose freely how many times to listen to each sound; on average, segments were played 2.2 times across both sessions.

Results of the ratings indicate substantive inter-rater reliability for both the first (Cronbach's alpha = .84) and second sessions (Cronbach's alpha = .701). For both sessions, a paired-samples t-test was computed to investigate whether participants perceived the segments as belonging to one category more than another. For the first session, participants indicated that the segments were more readily perceived as "OO" than "AAH" (p=.002). For the second session, three paired-sample t-tests were computed to test each of the three categories against one another. Results suggest that segments were readily perceived as /u/ compared with /ɔ/ (p=.03). No other statistically significant effects were observed; a larger-scale study should seek to investigate various factors possibly affecting the perception of great ape quasi-vowels by human listeners, as well as mapping acoustic properties corresponding to the perception of such qualities.

Finally, resynthesis of observed values may constitute an alternative method to that described here. Such syntheses can be accomplished using a variety of available software [21, 22].

6. TOWARD VOCAL TRACT MODELS

The human vocal tract was reconfigured in evolution, including changes in the marked flexure of the skull base, expansion of the pharyngeal cavity, descent of the larynx and descent of tongue root into the throat [23, 24], and concomitant "particulation" (rounding) of the tongue into independently movable sections [25]. Through its reconfiguration, the human vocal tract attains a roughly 1:1 relationship between vertical and horizontal sections, affording extreme articulations necessary for quantal vowels [24, 26]. Great apes, like other nonhuman mammals have a "flat" tongue residing almost entirely in the mouth [1, 24, 27], and narrow pharynx (i.e., back cavity), while that of (adult) humans is partially descended into the throat.

While much remains unknown about articulatory capacities by nonhuman great apes, there are currently no indications that the vocal tracts of, e.g., orangutans, are capable of achieving the extreme 1:10 midpoint discontinuities required for quantal vowels [28, 24]. In this context, it is worth noting that work by Fitch et al. is commonly claimed to have refuted this idea [8]. However, the extent of the macaque data presented by those authors were driven by outlier data observed when the animal was yawning, and also did not extend to quantal vowels [i] or [u] [29]. Thus it does not constitute evidence against purported anatomical limitations on nonhuman primates' articulatory capacities [24, 29, 30, 31].

Thus, having confidently observed [u]-like formant dispersions, we are faced with an intriguing question: given an orangutan vocal tract, how may such acoustic properties otherwise be produced? Two clues come from *(1)* visual observations of chimpanzee "hoots" – which are seemingly acoustically and perceptually comparable to orangutan long call segments analyzed above, which shows that these calls are produced with protruded rounded lips [32]. And *(2)*, previous research on the acoustic consequence of laryngeal air sacs (found in most primates, including orangutans) [33], which indicates that air sacs shift down formants, and shifts them closer together.

To verify estimated formants and putative vowel-quality identified in orangutan long calls, we therefore seek to reverse engineer a plausible vocal tract shape for the animal producing those sounds. We performed this modeling effort using the *Wormfrek* software (J. Liljencrants, KTH), which allows the systematic variation of the vocal tract area transfer function as a prediction based on a sequence of uniform tubes [34] (Fig. 4). These efforts, while tentative, are suggestive of vocal tract shapes allowing animals with "unconfigured" vocal tracts to achieve [u]-like formant dispersions and vowel-like qualities. A larger-scale study should be conducted to determine the validity of these impressions. The acquisition of real-life vocal tract area dimensions and measurements promises to improve the performance of future such efforts.

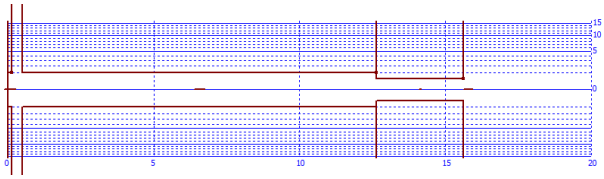

Figure 4: Tube model rendered in the *Wormfrek* software. The model assumes a vocal tract length of 12 cm, elongated with protruding (3 cm) rounded (.4 cm^2) lips, narrowing the lip passage. A .5cm cavity of 30cm^2 is assumed after a brief constriction (.125cm), simulating air sacs. The model predicts F_1=273 Hz, and F_2=1070 Hz.

7. DISCUSSION

Using a novel protocol of supervised acoustics paradigms, PREQUEL, we have confidently identified for the first time [u]-like productions in wild orangutans. Importantly, our approach allows for bypassing common hazards in related work, such as mistaking harmonic partials for formants. Though applied here to orangutans, PREQUEL can theoretically be applied to any primate call. This poses the exciting possibility for charting vowel-like articulations that, among hominids, may be older than speech itself. Moving forward, we intend to apply the PREQUEL protocol to a range of vocalizations by great apes as well as non-great ape primates, with the ultimate goal of mapping ancestral human articulatory capacities.

8. ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the UK Research & Innovation, Future Leaders Fellowship (grant agreement number MR/T04229X/1; ARL), and the Swiss National Science Foundation (PCEFP1_186841; SM). The results of this work and the tools used will be made more widely accessible through the national infrastructure Sprakbanken Tal under funding from the Swedish Research Council (2017-00626).

9. REFERENCES

[1] V. E. Negus, *The comparative anatomy and physiology of the larynx*. Heinemann, 1949.

[2] D. F. N. Harrison, *The anatomy and physiology of the mammalian larynx*. Cambridge University Press, 1995.

[3] G. Kelemen, "Comparative anatomy and performance of the vocal organ in vertebrates," in *Acoustic behaviour of animals*, R. Busnel, Ed., 1963, pp. 489–521.

[4] M. J. Owren, R. M. Seyfarth, and D. L. Cheney, "The acoustic features of vowel-like grunt calls in chacma baboons (papio cyncephalus ursinus): Implications for production processes and functions," *Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2951–2963, 1997.

[5] P. H. Lieberman, D. H. Klatt, and W. H. Wilson, "Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates," *Science*, vol. 164, pp. 1185–1187, 1969.

[6] P. Lieberman, E. S. Crelin, and D. H. Klatt, "Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee," *American Anthropologist*, vol. 74, pp. 287–307, 1972.

[7] L. J. Boë, F. Berthommier, T. Legou, G. Captier, C. Kemp, . T. R. Sawallis, and J. Fagot, "Evidence of a vocalic proto-system in the baboon (papio papio) suggests pre-hominin speech precursors," *PloS one*, vol. 1, no. 12, pp. 3329–3343, 2017.

[8] W. T. Fitch, B. D. Boer, N. Mathur, and A. A. Ghazanfar, "Monkey vocal tracts are speech-ready," *Science Advances*, vol. 12, no. 2, p. e1600723, 2016.

[9] A. G. Ekström, "Ape vowel spaces remain elusive: A comment on grawunder et al. (2022)," *International Journal of Primatology*, pp. 1–3, 2022.

[10] A. R. Lameira and S. A. Wich, "Orangutan long call degradation and individuality over distance: a playback approach," *International Journal of Primatology*, vol. 29, pp. 615–625, 2008.

[11] S. Granqvist, "tolvan.com," Downloaded 2023.

[12] D. L. Bowling, M. Garcia, J. C. Dunn, R. Ruprecht, A. Stewart, K. H. Frommolt, and W. T. Fitch, "Body size and vocalization in primates and carnivores," *Scientific reports*, vol. 7, no. 2, pp. 1–11, 2017.

[13] J. C. Mitani, T. Hasegawa, J. Gros-Louis, P. Marler, and R. Byrne, "Dialects in wild chimpanzees?" *American Journal of Primatology*, vol. 27, pp. 233–243, 1992.

[14] R. B. Monsen and A. M. Engebretson, "The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction," *Journal of Speech, Language and Hearing Research*, vol. 26, pp. 89–97, 1983.

[15] J. Sundberg, "Articulatory configuration and pitch in a classically trained soprano singer," *Journal of Voice*, vol. 23, pp. 546–551, 2009.

[16] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1960.

[17] ——, "Descriptive analysis of the acoustic aspects of speech," *Logos*, vol. 5, pp. 3–17, 1962.

[18] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage Learning, 2011.

[19] J. C. Catford, *A practical Introduction to Phonetics*. Oxford University Press, 1988.

[20] S. Granqvist, "The visual sort and rate method for perceptual evaluation in listening tests," *Logopedics Phoniatrics Vocology*, vol. 28, pp. 109–116, 2003.

[21] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 3, no. 67, p. 971â995, 1980.

[22] S. Barreda, *phonTools: Functions for phonetics in R.*, 2015, r package version 0.2-2.1.

[23] J. T. Laitman and R. C. Heimbuch, "The basicranium of plio-pleistocene hominids as an indicator of their upper respiratory systems," *American Journal of Physical Anthropology*, vol. 3, no. 59, pp. 323–343, 1982.

[24] P. Lieberman, "Vocal tract anatomy and the neural bases of talking," *Journal of Phonetics*, vol. 4, no. 40, pp. 608–622, 2012.

[25] M. Studdert-Kennedy, "The particulate origins of language generativity: From syllable to gesture," in *Approaches to the evolution of language*, 1998, pp. 202–221.

[26] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 1, no. 17, pp. 3–45, 1989.

[27] E. S. Crelin, *The human vocal tract*. Vantage Press, 1987.

[28] B. De Boer and T. W. Fitch, "Computer models of vocal tract evolution: An overview and critique," *Adaptive Behavior*, vol. 1, no. 18, pp. 36–47, 2010.

[29] P. Lieberman, "Comment on "monkey vocal tracts are speech-ready"," *Science Advances*, vol. 7, no. 3, p. e1700442, 2017.

[30] H. Takemoto, "Morphological analyses and 3d modeling of the tongue musculature of the chimpanzee (pan troglodytes)," *American Journal of Primatology*, vol. 70, pp. 966–975, 2008.

[31] A. G. Ekström, "Viki's first words: A comparative phonetics case study," *International Journal of Primatology*, pp. 1–5, 2023.

[32] S. Grawunder, N. Uomini, L. Samuni, T. Bortolato, C. Girard-Buttoz, R. M. Wittig, and C. Crockford, "Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage," *Philosophical Transactions of the Royal Society B*, vol. 1841, no. 377, 2022.

[33] B. de Boer, "Acoustic analysis of primate air sacs and their effect on vocalization," *The Journal of the Acoustical Society of America*, vol. 6, no. 126, pp. 3329–3343, Dec. 2009.

[34] J. Liljencrants and G. Fant, "Computer program for vt-resonance frequency calculations," *STL-QPSR*, vol. 16, pp. 15–21, 1975.