International Journal of Primatology
https://doi.org/10.1007/s10764-022-00335-6

Ape Vowel-like Sounds Remain Elusive: A Comment on Grawunder et al. (2022)

Axel G. Ekström[1]

Analysis of nonhuman, great ape vocal behavior may provide insight into the evolution of human speech. Two main impositions prohibit such work. The first and most obvious is the scarcity of relevant data; even though literature on primate vocalization stretches back decades, high-quality data are rare and seldom publicly available. The second—far more subversive—is a misunderstanding of the phenomenon at hand, leading to inadequate analyses and conclusions. Here, I discuss the latter in the context of a recent publication by Grawunder et al. (2022) in *Philosophical Transactions of the Royal Society B*, purporting to show a chimpanzee (*Pan troglodytes*) "vowel-like sound space" (Fig. 1) and argue that the authors' analyses are inadequate for their intended purpose and that interpretations stemming therefrom are invalidated.

Articulate speech is a respiratory, laryngeal, and supralaryngeal phenomenon. Pulmonic airflow causes vibrations in laryngeal vocal folds and is forced through constrictions on the vocal tract. The rate of vocal fold vibration is termed the fundamental frequency (f_0). Imposition of narrow constrictions in the vocal tract results in spectral frequency peaks, resulting from resonances, termed the first and second formants (F_1 and F_2, respectively). F_1 is mainly determined by tongue body height and jaw opening; F_2 by tongue body shape and front-to-back position. A vowel space refers to a two-dimensional area plotting observed F_1 and F_2 coordinates (Fig. 1).

Grawunder et al. (2022, p. 8) claim that several analyzed chimpanzee calls "extend beyond the human vowel space." Peak F_1 values observed in human speech are for [æ], [ɑ], and [a], at ~ 1,000 Hz, and peak F_2 values are observed for [i] and [e] at ~ 3,000 Hz—both when uttered by a child. For adults, typically observed values are ~ 800 Hz (F_1) and ~ 2,400 Hz (F_2) for the same sets of vowels. By comparison, for one "scream" call in the Grawunder data—the single most extreme data point—F_1 approximates 1,800 Hz, and F_2 approximates 3,500 Hz. Other factors may well account for these discrepancies.

Handling Editor: Joanna M. Setchell

✉ Axel G. Ekström
axeleks@kth.se

[1] Speech, Music & Hearing, KTH Royal Institute of Technology, Lindstedtsvägen 24, 114 28 Stockholm, Sweden

Published online: 17 November 2022

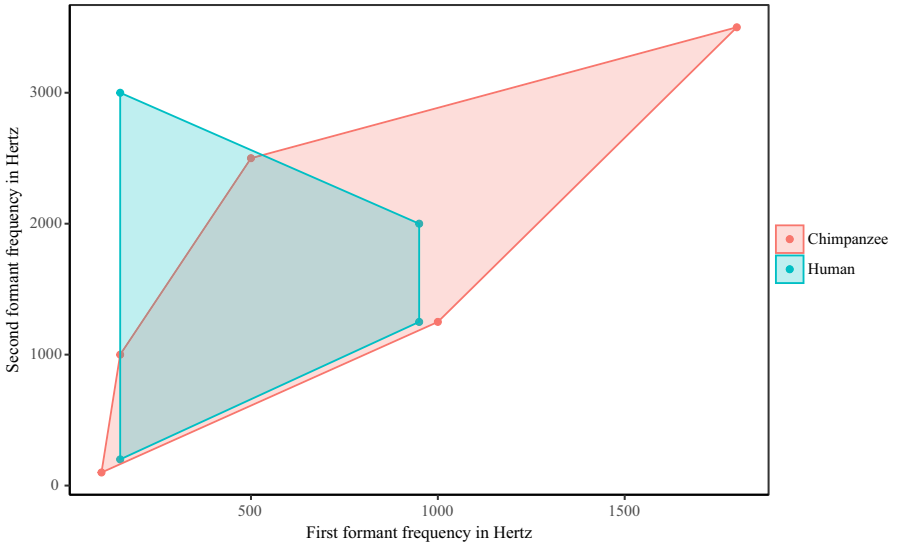

Fig. 1 Supposed chimpanzee vowel-like sound space, estimated from Grawunder et al., 2022. Human vowel space, derived from the IPA vowel chart (internationalphoneticassociation.org), is superimposed.

In Grawunder et al. (2022), an unsupervised Praat (praat.org) script is said to have estimated formants using the program's standard Formant (burg) algorithm linear predictive coding (LPC) method. Results were later cross-checked by one of the authors. However, formant estimation, by eye as well as by automatic means (e.g., LPC), becomes more sensitive to error at higher f_0. The problem is tangible already at $f_0 < 300$ Hz (where F_1, F_2, and F_3 estimation errors may be up to ± 60 Hz; Monsen & Engebretson, 1983) and becomes more significant as f_0 increases. In particular, harmonic partials are known to bias estimates of lower formants. For linear prediction methods, accuracy degrades because higher-f_0 voices have comparably large energy concentrations, which are favored by linear prediction through its least-squares error criterion, around f_0 itself and its lowest integer multiples. Further degradation may take place because of autocorrelation-domain aliasing procedures. Indeed, most F_1 and F_2 values reported for scream calls (Grawunder et al., 2022, p. 9) follow an apparent 1:2 ratio, suggesting that frequencies of two lower-spectrum partials were interpreted as formants.

While Grawunder et al. do not report f_0 values of chimpanzee screams from which formants are estimated, measurements for scream climaxes have elsewhere been reported to be as high as $f_0 > 1{,}500$ Hz (Mitani et al., 1992). The radical F_1 and F_2 values reported for chimpanzee calls (scream calls in particular, though there are suspicious data clusters for other calls types also), thus, likely reflect biased values, resulting from the application of inadequate acoustic analysis procedures incapable of accurately estimating formants from such high f_0 values. Furthermore, while a

variety of methods have been designed to counter the issue presented by formant estimation given high f_0, including modifications to traditional LP (Alku et al., 2013) and formant estimation through inverse filtering of the oral air-flow waveform (e.g., Gauffin & Sundberg, 1989), none are applied, or even discussed, by the authors.

The acoustics of nonhuman primate vocal production are poorly understood, with no agreed-upon methods of data treatment or analysis. Because of differences in both the behavior itself, as well as species' vocal anatomy, methods developed for the analysis of human speech do not constitue ready-made methodologies adequate for exploring acoustics of nonhuman primate vocalizations. Until such a method is properly explored, researched, and described, the production of vowel-like sounds by nonhuman primates, and any insights they provide hominid speech evolution, will remain elusive. The work by Grawunder et al. (2022), while presenting highly valuable and sought-after data, does little to change this. I hope that highlighting pitfalls such as those discussed above will aid the development of methods for analyzing vowel-like vocal production from in situ primate vocalization.

Acknowledgements The author gratefully acknowledges Johan Sundberg for valuable comments and discussions on the topic at hand.

Declarations

Conflict of Interest The author declares no conflicts of interest.

References

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A. M., & Story, B. H. (2013). Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America, 134*(2), 1295–1313.

Gauffin, J., & Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech, Language, and Hearing Research, 32*(3), 556–565.

Grawunder, S., Uomini, N., Samuni, L., Bortolato, T., Girard-Buttoz, C., Wittig, R. M., & Crockford, C. (2022). Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage. *Philosophical Transactions of the Royal Society B, 377*(1841), 20200455.

Mitani, J. C., Hasegawa, T., Gros-Louis, J., Marler, P., & Byrne, R. (1992). Dialects in wild chimpanzees? *American Journal of Primatology, 27*(4), 233–243.

Monsen, R. B., & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of Speech, Language, and Hearing Research, 26*(1), 89–97.