PREPRINT

Cross-modal constraints in multimodal vocalizations in Siamang (*Syndactylus symphalangus*)

Wim Pouw* [1] **Mounia Kehy** [2] **Marco Gamba** [3] **Andrea Ravignani** [4, 5, 6]
[1] Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen, The Netherlands
[2] Equipe de Neuro-Ethologie Sensorielle, Université Jean Monnet, France
[3] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Torino, Italy
[4] Max Planck for Psycholinguistics Nijmegen, The Netherlands
[5] Center for Music in the Brain, Department of Clinical Medicine, Aarhus University The Royal Academy of Music, Aarhus, Denmark
[5] Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
*Correspondence: wim.pouw@donders.ru.nl

ABSTRACT

Gibbons (*Hylobatidae*) sing loudly, reaching over a hundred Decibels - about the sound level of a rock concert. Qualitative observations report that, during song climaxes, individuals move in a coordinated way with their singing. We hypothesize that vigorous thorax-loading movements such as brachiation induce physical constraints on the respiratory-vocal system. This coordination and possibly underlying biomechanics have never been studied, leaving a blind spot for a possible shared variance between vocal and locomotor repertoires that would indicate a co-evolution of two systems. Here, we recorded over a hundred stereotypical multimodal vocal calls from over 7 hours of singing in two captive siamangs (*Syndactylus symphalangus*), a type of gibbon that are the largest in size. These stereotypical calls coincided with a bodily display during solo singing, and were exclusively performed by juvenile individuals. We used computer vision methods (opencv and deeplabcut) to quantify upper body acceleration of these multimodal displays of two individuals; we found that body acceleration statistically predicted the nearest peak in the amplitude envelope of the call. The results indicate that Siamang singing likely co-evolved with movement due to physical constraints of pectoral limb-respiratory-vocal interactions, similar to birds, bats, and rodents, as well as humans. This has important implications for the locomotor hypotheses of rhythm origins as our results suggest a biomechanical and basic coupling of vocal systems with the (loco)-motor system: in singing Siamangs, and perhaps humans too, as there are homologies to be drawn between how humans and Siamang move in gestural ways with their vocalizations.

Keywords: Siamang, Locomotor-vocal coupling, Locomotion, Respiration, Vocalization, Multimodal displays

INTRODUCTION

To master the biophysics of vocal production, the jaw, tongue, velum, and larynx must move. Vocal folds will need to vibrate. And the surrounding structures of the lungs must move to produce air flow. In complex vocal sound production, it takes a village of body parts to sound the right way. While there is classically a lot of scientific interest in the potential top-down sound-modification abilities of animals [17,27,37,41,54] - often so as to compare to articulatory dexterity in humans [16,32] - some bottom-up constraints on vocalization might have impacted vocal evolution in a wide range of animal taxa.

Examples of bottom-up constraints on vocalization exist in a few species. Multiple bat species (e.g., *Phyllostomus hastatus*) synchronize - in 1:1 or reliably polyrhythmic fashion - echo-locating pulses with their wingbeats while flying [29,55]. Twelve species of North- and South-American birds show a similar coupling, where vocalizations are timed with the downbeats, and the wingbeat duration shows an allometric scaling with their vocal unit durations [4]. Gerbils (*Meriones unguiculatus*) locomote in a saltating way, when they hop and hit the ground with their forelimbs they synchronously emit a vocalization [6]. The male brown-headed cowbird uses vigorous wingbeats in their court-ship displays, and said activity affects respiratory-vocal activity [11]. Finally, humans too, show synchronization of impulses of their upper-limb movements with their vocalization [46,48]. What is the key (hypothesized) mechanism for all these findings? In these species, pectoral limb activity and the wider muscle chains involved lead to compressive effects on the surrounding structures of the lungs, thereby biomechanically modulating egressive drive that powers vocalizations.

Surprisingly, little is known about these physical constraints in non-human primates, despite an intensive interest in their vocal abilities. One group of primates seems the perfect model to understand motion-phonation interactions: the gibbons and siamang.

They are highly vocal species and they load their entire body weight on their pectoral system during their primary mode of locomotion - brachiation. Here we focus on the heavier weight siamang (*Syndactylus symphalangus*).

Siamang and Gibbons diverge in several ways from great apes. They are a highly vocal species, performing extremely loud (sometimes > 100 Decibels)[27,36], and rhythmically coordinated[51], duetting songs (e.g., see here), supporting family-bonding and territory-marking, and on occasion alarming. These apes also move at extremely fast speeds (sometimes >45 kms/h) through the canopy using hand-over-hand grasps, also known as brachiation. Small Asian Apes also move *while* they sing (e.g., see here). Interestingly, Haimoff (1981, p. 135)[20] observes a temporal coordination of locomotion and vocalization in the wild siamang they studied: Whereupon, the male vigorously moved away from the stationary female, and presented a rapid series of short barks, in association with the duration of the movement. These and other widespread *qualitative* observations[38,53] suggest that singing in siamang and gibbons may at times be a multimodal display, where movement and vocal sequences are interdependent, which may be comparable to other species that move and vocalize in coordinated ways[4,6,11,29]. The nature of multimodal performances in siamang and gibbons are difficult to make sense of, when we consider that the loud booming calls have been argued to be an adaptation for communication with distant conspecifics in thickly foliaged canopies where visual signaling reliability is reduced[38].

The production of loud calls requires a strong respiratory drive. A range of animals are constrained in their respiratory drive by their locomotion. Animals that include their pectoral limbs for locomotion, as in quadrupeds (e.g., dogs, horses, rhinoceros), synchronize their locomotor cycles at increasing gait speeds with respiratory cycles[8]. While it is generally acknowledged that brachiation must have effects on respiratory drive and thus vocalization[21,22,39,40,46] no quantitative evidence exists to support this mechanistic link. Green-rumped parrotlets (Forpus passerinus) overcome respiratory challenges to produce adult-like calls only once these birds start to use their wings in explorative ways, with their vocalizations becoming indistinguishable from adults after taking their first virgin flights[3]. Berg and colleagues (2013) suggest that by utilizing the high expiratory pressure generated during wing-powered flight, individuals could serendipitously and efficiently amplify their high frequency calls. Hence, not only could brachiation similarly interact with singing in the Siamang, it can be aligned in a way so that it is minimally not counteradaptive, and possibly adaptive to do so.

In this study we opportunistically observed captive siamang and audio-visually recorded singing for a period of 21 days. We noticed characteristic multimodal vocalizations, with stereotypical vocalizations that synchronized with pulse-like movements (e.g., for examples see here). These calls with ululating screams as main units[19] were produced by juveniles during solo singing, that is, after the duetting singing was completed or was winding down. Such non-overlapping and loud solo songs may serve as an honest signaling function towards prospective mates (see e.g., [50]). We would suspect that if these high-amplitude solo calls are designed to inform others about ones adaptive fitness it is likely that all available bodily resources are recruited to stabilize and amplify said calls. The pulse-like movements synchronized with solo calls are a possible candidate acting as an embodied resource for supporting vocalization, and we will assess whether said movements associate with the loudness of these solo-songs. Given that we know from biomechanics in humans and other animals that the physical impact of a movement on the musculo-skeletal system is during acceleration or deceleration (as forces are a function by mass (c) and acceleration)[2,6,12,23,28,45–47]; here, we test whether thorax accelerations during multimodal vocalizations statistically predict the amplitude of concurrent vocalization in Siamang.

METHODS

Data recording

Audiovisual recordings of a family of siamang (6 members; female adult, male adult, two male juveniles, one infant, and a newborn) were collected in the June and August of 2022 over two visits at the Jaderpark Zoo in Lower Saxony, Germany. This yielded over 7 hours of recorded singing, collected by the first and second authors. The siamang sang primarily in the morning around 9-10am, or after their fruit and vegetable lunchtime, around 1pm, and occasionally around 5pm. Only two juvenile/young adult individuals performed the stereotypical multimodal performances we consider here, namely Baju (7 years and 8m) and Fajar (4 years and 11m). Baju and Fajar were both born in the Jaderpak zoo. During our stay, Baju was separated from the family due to risk of injury after a

fight where all family members attacked Baju after a transgression. This also means we could not collect more data from Baju during our second visit.

Audio-visual recording

Four GOPRO Hero9 were installed, set at 1080 quality, sampling at 59.74fps, with linear lens settings. We then cropped frames, re-compressed to sample at a regular 50fps. The camera positions were positioned as orthogonally from each other as the site allowed (for a sketched map with (geometrically) estimated distances based on a laser-pointer measurement device, see here).

We further use in this study the four audio sources that were available from the GO-PRO compliant MKE400 Sennheiser microphones with windjammers, sampling at 48Khz. We combined these four acoustic waveforms to yield a single-channel audio source. Since the cameras were placed at near-orthogonal angles from each other, we can be confident that amplitude peaks are much less affected by differences in direction of sound radiation, or location of the singing individual.

Identification of multimodal vocalizations and related features

The first and the second author identified opportunistically as many multimodal vocalizations by going through all the recorded songs and annotating these events in ELAN [58]. These vocalizations were easy to identify because they all consisted of a ululating scream as a main unit, and any variability in the call structure was stereotypical within the two individuals. Interestingly, these stereotypical vocalizations always co-occurred in close synchrony with a pulse-like movement, though with variable intensity and different types of locomotion. The annotator (second and first author) drew a liberal boundary around the movement + vocalization event, such that at least the movement and the vocalization sequence was contained. We will refer to these annotated events as multimodal vocalizations throughout.

The types of physical constraints expected are likely dependent on the type of action performed during vocalizations; therefore we attempted to apply a standardized description to locomotor actions. We used Hunts typology of locomotion types [24] to characterize the action that occurred during the vocalization (Figure 1). In some instances, we slightly deviated from Hunts 35 categories to accommodate for a particular locomotion action (e.g., drop fore limb swing: the individual sits on top of horizontal structure, and then scoots backwards or forward to drop and swing forward with two extended arms). Note, a common mode of locomotion for the Siamang, ricochetal brachiation (e.g., see here), is absent in our dataset, possibly due to the facility having more ropes than rigid and connected supports, and thus being more tailored towards swinging movements rather than ricochetal brachiation . Some multimodal vocalizations remained undefined as they did not fall into a clear category (i.e., mixed locomotion modes). We will make a crude binary distinction between locomotion types which load the entire weight on the thorax via the shoulder girdle(s) (forelimb only[0]), or those that involve distribution of weight or support via the lower limbs (other). In the case of forelimb only loads one would particularly expect accelerations to constrain respiratory-vocal interactions.

Video Preprocessing and post-processing

The footage of the events each has four potential camera angles. We first checked all these camera angles to see whether the individual was not or badly visible. If the individual was not visible, it was excluded as a potential camera angle submitted for analysis. Video processing was performed in Python, the specific steps discussed below. Further processing to prepare the dataset for statistical analyses was performed in R and R-studio.

Cutting scenes and performing initial motion detection with OpenCv2

Firstly, a custom Python script automatically cut the videos based on the ELAN annotations (see here) using moviepy, ffmpeg and pydub. Secondly, we determined regions of interest for each scene. The installed cameras had a field of view to opportunistically capture behavior at different locations. This makes tracking with supervised computer vision more computationally costly as there may be many individuals moving in a complex structured site. As a pre-processing step we therefore created a python pipeline (see see here) using OpenCv2 that determined pixel deviations from the median to ascertain a key area of movement per frame, which

[0] Strictly speaking, since siamang are primarily bipedal, we could have also referred to the pectoral limbs as upper limbs (rather than forelimbs). But we decided to follow Hunts categories here.

was after some processing steps (smoothing, obtaining maxima) used to determine a static bounding box that would further serve to crop the video to primarily contain the movement of the siamang and exclude the rest of the complex scene (for an example see here).

Tracking DeepLabCut

A convolutional neural network (Resnet-50) was trained using DeepLabCut (version 2)[35] with 250 hand-labeled frames (download weights and other info here). Two key points were used for the training set, one for the thorax (i.e., which was used for acceleration) and another for the underside of the individual (which was used for the body normalization of the acceleration magnitudes). We trained the model with half a million iterations, reaching an average error rate of 1 pixel (for keypoints with 60% confidence rates) in the training set, and 25 pixels in the test set. If we normalize these errors by the original frame-sizes (1500x1080), then we get an error of 0,0015%.

When using the trained model to extract position traces for the video recordings we applied the model to all events with DLCs native filtering option to remove noise-related jitter, yielding x,y position traces for the two key points (see e.g., here) and likelihoods. Since derivatives increase power of noise-related jitter relative to slower frequencies[57], we also applied extra smoothing of the resultant position traces with a 9th order Kolmogorov Zurbenko filter with a span of 110ms (**R**-package **kza**).

Likelihoods were further used for data quality curation. Specifically, if a camera angle had tracking that dropped for more than 5% below a threshold of .80, than we did not submit that particular camera for further analyses; the remaining 5% of the data that had low confidences, were linearly interpolated (na.approx function using **R**-package **zoo**) using the surrounding high-confidence tracking samples. Note that the DLC team has recommended a likelihood of at least .60 for good tracking, so we are being slightly more conservative.

Normalization

The thorax accelerations are calculated by differentiating speed measured in pixels over time. However, different camera positions, and different locations of the Siamang, make pixel-units problematic. Therefore, we normalized the kinematics to a dimensionless quantity by scaling the position traces by the mean body size of the Siamang detected (for all frames that had a DeepLabCut confidence estimate of 100%). This means that all kinematics in this report are normalized by body size units.

Approximating kinematics from multiple camera angles

Depending on the location of the individual, we may have more than one camera angle that recorded the multimodal vocalization. The many objects on the site and the distances of the cameras did not allow us to perform stereoscopic reconstruction of the camera angles to estimate 3D postures using a Charuco board[25,52,56]. Obviously we cannot take the Euclidean norm of all the accelerations recorded for different camera angles, as this would yield an overestimation given that camera angles have correlated information (for example, because they all capture vertical displacement of the individual).

We combined position traces of n cameras (c) available $M = [x_{c=1}, y_{c=1}, x_{c=2}, y_{c=2}, ...x_{c=n}, y_{c=n}]$ to a principal component analysis, to calculate the eigenvectors (v), where we extract the three highest loading principal components, $PC_{1-3} = [Mv(:,1), Mv(:2), Mv(:3)]$, which will approximate positional information about orthogonal planes. Subsequently, we take the Euclidean norm of the derivatives of the first three principal components to get an approximation of 3D speed vector $\vec{s}$, such that $\vec{s} = \sqrt{\Delta PC_1^2 + \Delta PC_2^2 + \Delta PC_3^2}$, which is then differentiated once more, and absolutized, to yield an approximated 3D acceleration magnitude vector, $\vec{a} = (|\Delta\vec{s}|)$. Note after each differentiation we smooth the data with a 5th order Kolmogorov Zurbenko filter with a span of 50ms.

Acoustics: Smoothed amplitude envelope

We extracted a smoothed amplitude envelope of the waveform from the combined audio sources using a common approach (see here for example code), by first applying a Hilbert transformation[1], and then taking the complex modulus. This resulted in a one-dimensional time series, which was downsampled to 100Hz and further smoothed with a 12 Hz Hanning window.

Peak analyses

For each multimodal event we determined the global maximum in absolutized acceleration (max acc), i.e., the largest magnitude of either acceleration or deceleration. Then we obtained the local maxima in the smoothed amplitude envelope, with a `findpeaks` function using `R`-package `pracma`. This allowed us to select the magnitude of the local peak in the envelope nearest to max acc. Since the acoustic and acceleration peak data were long-tailed distributed we log transformed the variables (which also improved model fits), and z-normalized.

Aggregation

We synchronized the motion tracking data and the acoustic data by first aligning the data samples in time, and upsampling the motion tracking to 100Hz to preserve the high-sampling rate of the acoustics. We upsample the motion tracking data by linearly interpolating the data along a time vector using `R`-package zoo (function `na.approx`[59]) so as to have a regular sampling rate at 100Hz that exactly matches the sampling times of the amplitude envelope. This yields a combined time series with acoustics and motion tracking that could be further processed for analysis. `R`-code performing the above-mentioned processing steps is available on Github.

Final processed dataset

After having to exclude 29 events that had low-confidence tracking the final dataset consisted of 83 events (of which; 26 Baju and 57 Fajar).

RESULTS

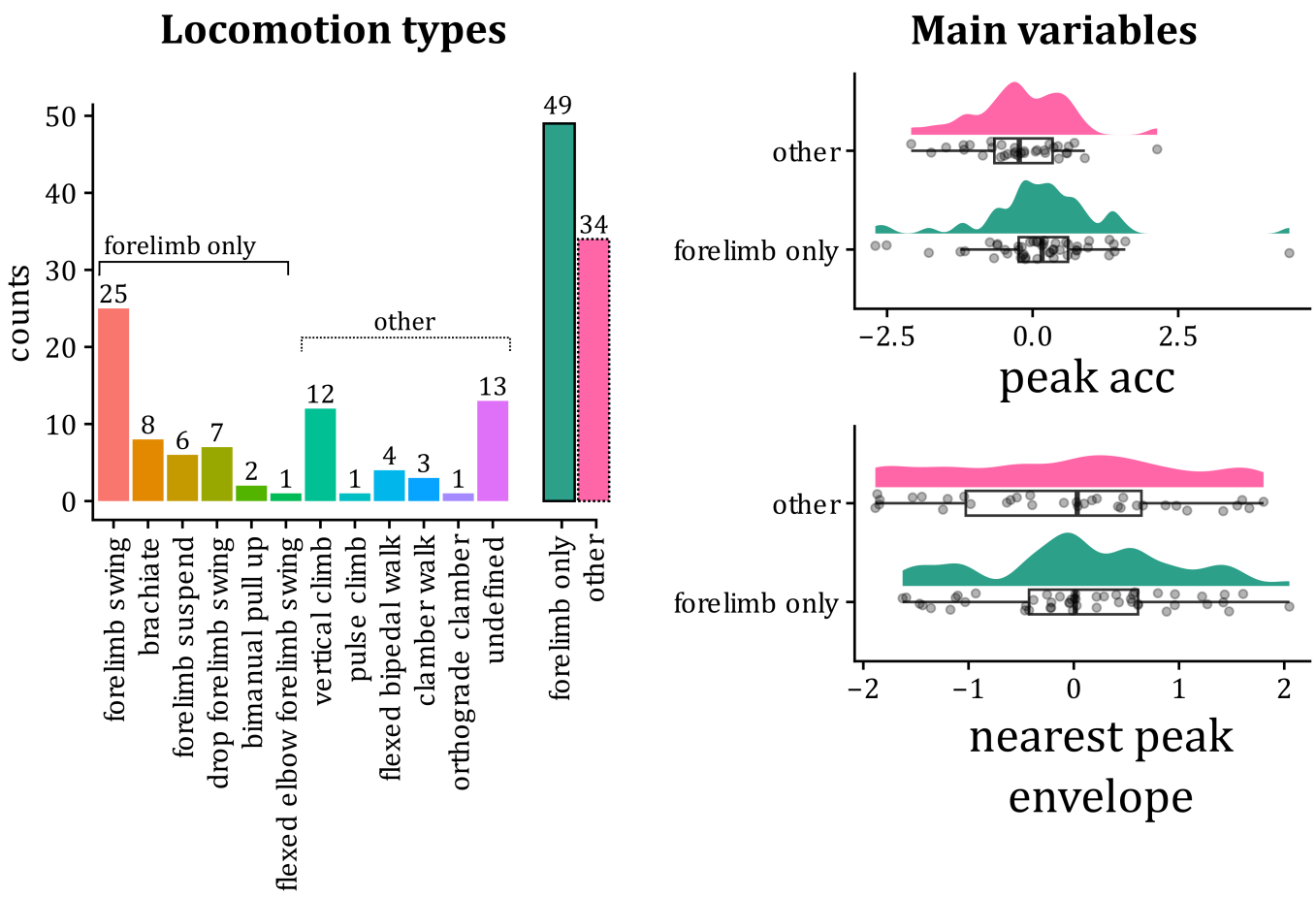
An `R`-markdown script supporting the statistical analyses can be found here. Table 1 provides information about the main variables, as well as the time windows of the annotations (approximate length of the multimodal events). The temporal inter-peak distance between the global maximum in acceleration and the nearest local maximum in amplitude was on average 120 milliseconds (*SD* = 160). This low temporal distance provides confidence that the two point-estimates for kinematics and acoustics occurred sufficiently close in time to be possibly coupled. Table 1 further shows that, based on 95% confidence intervals, the older juvenile Baju seemed to generate higher peaks in the amplitude envelope (nearest to peak acceleration) as compared to his younger brother Fajar, while being comparable in their body accelerations peaks. Further, forelimb only locomotion tends to have higher accelerations than other locomotion types. The main variables did not dramatically differ either by individual or locomotion type.

Table 1: Descriptives stats main variables.

Variable	Overall	Baju	Fajar	Forelimb only	Other
Time annotation *M* (*SD*)	2851 (491)	2495 (268)	3014 (485)	2815 (485)	2905 (504)
Time annotation 95%CI[lower, upper]	[2744, 2959]	[2386, 2603]	[2886, 3143]	[2675, 2954]	[2729,3080]
Nearest peak envelope (z) *M* (*SD*)		0.52 (0.82)	-0.24 (.99)	0.07 (0.91)	-0.10 (1.12)
Nearest peak envelope (z) 95%CI[lower, upper]		[0.19, 0.85]	[-0.50, 0.03]	[-0.19, 0.33]	[-0.49, 0.29]
Peak acceleration (z) *M* (*SD*)		0.29 (0.69)	-0.13 (1.09)	0.15 (1.09)	-0.21 (0.83)
Peak acceleration (z) 95%CI[lower, upper]		[0.01, 0.57]	[-0.42, 0.16]	[-0.21, 0.46]	[-0.50, 0.08]
Inter-peak distance *M* (*SD*)	121 ms (160)	157ms (165)	104 (156)	107 (154)	141 (169)
Inter-peak distance 95%CI[lower, upper]	[86, 156]	[90, 224]	[62, 146]	[62, 151]	[83, 200]

Note. Max nearest envelope is the z-scaled magnitude of log peak smoothed amplitude envelope. Note that there are no overall descriptives for variables that have been z-scaled (amounting to *M* = 0, *sd* = 1)

Figure 1: Frequency distributions for the locomotion types, peak acceleration, and limbs .



Note. The left panel shows the number of different locomotion types we observed. Since there are too many categories, with few instances, we created super categories that indicate locomotion actions that only included the fore/upper limbs (forelimb only), versus those that included another limb (other). On the right panel, the distributions are shown for the magnitude in the global peak of acceleration, per locomotion category.

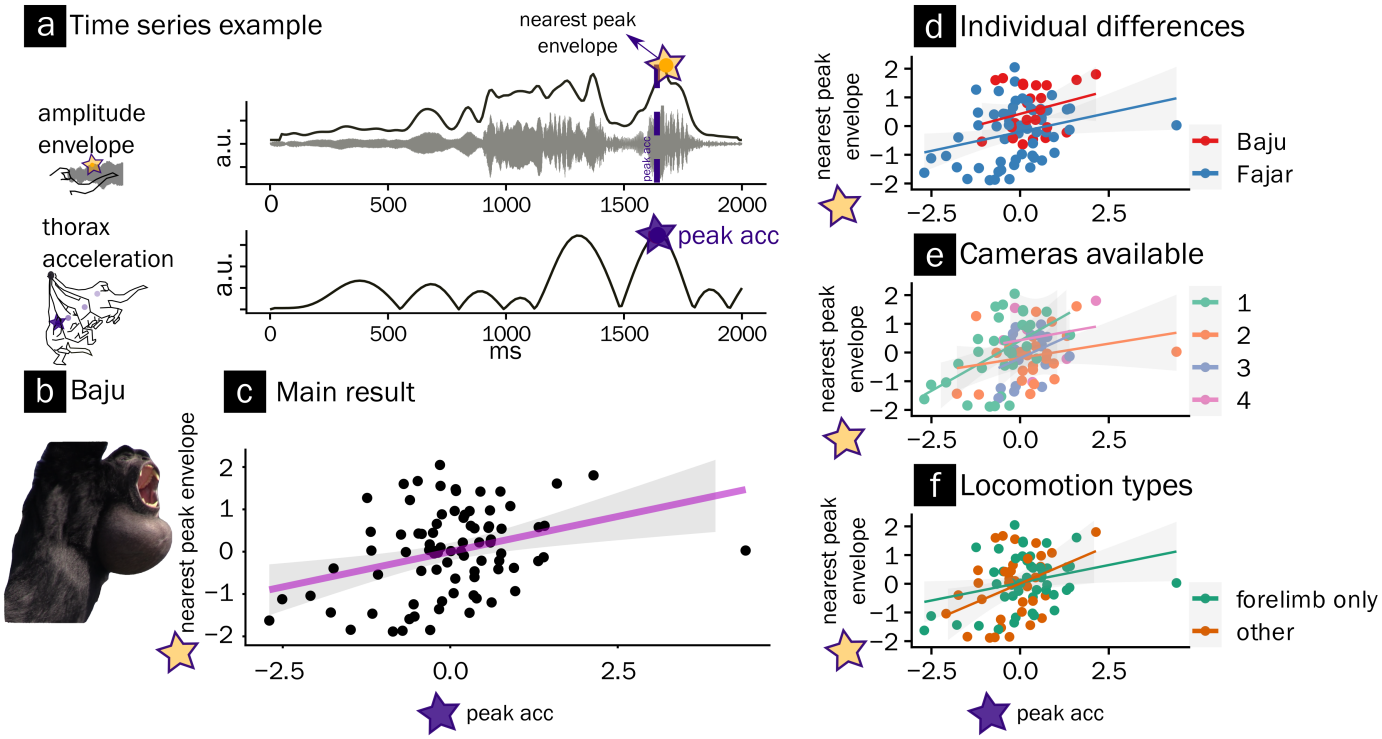
Main results

To assess whether body accelerations constrain vocalizations amplitude, we assess whether both parameters reliably scale in magnitude. Figure 2 provides a summary of the procedure and the main results, also partitioned for possible moderating/confounding variables.

A mixed linear regression model was performed associating peak acceleration with nearest peak envelope (using maximum likelihood; `R`-package `nlme`[43]). Individual (Fajar, Baju) was set as random intercept, but a model with random slopes for individual did not converge. The model statistically predicting peak envelope from peak acceleration reliably outperformed a base model predicting the overall mean in amplitude; change in ^{2} (1) = 7.68, p = 006. The model coefficients indicated that a higher magnitude peak in acceleration reliably associated with higher magnitude envelope peaks, b = 0.29, b 95%CI = [0.08, 0.49], t(80) 2.82, p = .006, intercept b = 0.09, t (80) = 0.41, emphp = 0.682. Since the model is performed on a log-log scaled, we can interpret the coefficients as indicating that for a unit of increase in acceleration there is a 30% increase in the magnitude of the nearest peak envelope (i.e., 1/3 power law). A simple regression analysis yielded a similar conclusion, r = .33, t(81)= 3.18, p = .002. Also, note that if we remove the possible outlier shown in Figure 1B (see supplemental plot without outlier) the conclusions remain unchanged, r = .38, t(80)= 3.1829, p < .0001.

We explored possible confounds (e.g., number camera angles available) or moderators (e.g., locomotion type) of this kinematic-acoustic coupling via interactions, but such interaction models were not reliably outperforming our main model, ^{2} (1) < 8.44, ps

Figure 2: Nearest peak analyses and results for acceleration and amplitude envelope.



Note. The time series example (a) shows data for a single event (of 83 in total). The upper panel of (a) shows the smoothed amplitude envelope of the call; the lower panel of (a) shows the acceleration of the thorax, both given in arbitrary units (a.u.). The purple star shows the global maximum for acceleration. This maximum is then used to determine the nearest local maximum in the amplitude envelope (yellow star, with purple outline), which in this case also happens to be the global maximum in amplitude. b) shows the individual Baju while vocalizing. The point-estimates are then correlated in (c), and plotted separately for each individual, Baju and Fajar (d), different cameras available (e), and locomotion type (f). To explore these data with the relevant events, please follow this (link).

> .21. Figure 2 further provides confidence that effects of acceleration and vocalization are stable among different measurement conditions, individuals, and locomotion types.

Thus we can conclude that the magnitude of vocalization amplitude peaks that occur around moments when the thorax undergoes its maximum acceleration or deceleration associates positively with the magnitude of that acceleration. For an inspection of the multimodal vocalizations for one camera angle see our dynamic data dashboard (see code here for reproducibility).

DISCUSSION

The evolution of vocal production can be framed as a continuous stabilization of multiple interacting subsystems originally evolved for different purposes[33,34,49]. Vocalizing is thus an emergent product of aligning lower-level dynamical processes to sound just right[13,14]. Many bodily systems must fall in line to produce such skilled and energetic vocal duet singing in Gibbons and Siamang. This duet singing is certainly shaped, as hypothesized till now, by several adaptations, such as complex vocal tract shaping[26] and the presence of laryngeal air sacs[7,15]. However, here we show that physical constraints of co-vocal movement partly shape singing. We find stereotypical pulse-like movements that co-occur with vocalizations in two juvenile Siamang engaged in solo singing. Using unsupervised and supervised computer vision and data science methods we approximated the magnitude of peaks in 3D acceleration/deceleration of the thorax, which we then related to the magnitude of the peak in the smoothed amplitude envelope nearest to the moment of peak acceleration. We obtained that the more the Siamangs thorax undergoes acceleration, the higher the amplitude of the vocalization. This suggests that the strong respiratory drive needed to vocalize, might be modulated by (more) accelerative locomotor actions, much like a range of other animals including humans[11,29,42,46,55]. The mechanism is hypothesized to reside in biomechanics, such that thorax-loading activities can increase egressive flow (for an overview[46]).

Our results have several implications. Firstly, vigorous movements during singing in Siamang have generally been described as multimodal displays. However, our results suggest that these behaviors should also be understood as coordination, where the biophysics of movement participates in the production of sounds [46]. This further has important implications for the large field on primate gesture studies, specifically as it tailors to the need for a better understanding of how different modalities contribute to communicative signaling [31] - after all, it seems that Siamang signal by using their body and their vocalization in an aligned way. Siamang, then, vocalize in a synchronized way with movement; this synchrony reminds how professional human singers couple their upper limb movements during vocalizations [42], and may be mechanistically analogous to how other non-human animals couple their pectoral limb activity to vocalization [4,6,11,30]. Our findings further tie in with recent research suggesting that species with more arboreal locomotion repertoires also have increased vocal singing abilities citepschruthEvolutionPrimateProtomusicality2021. More generally, by showing that body accelerations couple to vocalizations in singing Siamang we provide a concrete coupling mechanism that may figure into more general accounts that maintain that vocal capabilities such as rhythm may emerge from interactions with locomotion

However, contrary to such sentiments, coupling of the respiratory-vocal system with peripheral bodily movements is not necessarily something that might drive vocal flexibility over evolutionary time [4,9]. Several avian vocal learners tend to have a looser allometric scaling of their wingbeat duration with their vocalization duration, as compared to birds who are not vocal learners [4]. Mammalian vocal learners also tend to escape allometric scaling laws that relate vocalization and vocal tract size [18]. In humans, the increased flexibility in respiratory control has been in part attributed to a weakening of biomechanical constraint of locomotion with respiration (as the thorax was no longer impacted by locomotion) [9]. These considerations raise the possibility that the evolutionary pressure that is generated by a (too rigid) biomechanical coupling of vocalization with other physical constraints may lead to limiting effects on a species vocal evolutionary trajectory.

There are several limitations to the current study. It is based on data obtained from captive rather than wild Siamangs, from a single zoo, and it only takes into account certain (multimodal) vocalizations. These particular issues can only be resolved by collecting and analyzing more data and behaviors. Further, our measurements do not allow at present for more fine-grained 3D tracking of the animals; they also do not allow for perfect vocal amplitude measurements since that would require much more controlled parameters (e.g., constant distance between source, constant direction of radiation). We should further highlight that our hypothesized mechanism of vocal-motor interaction is on the level of biomechanics (kinetics) suggesting that mechanical loading on the thorax increases eggressive flow for vocalization (raising amplitude), but our analysis is on the level of movement (kinematics). Accordingly more work is needed on the level of biomechanics too [5,10,30,44], evaluating how locomotion might affect the surrounding myofascial-skeletal structure around the respiratory system and thereby far-from-equilibrium subglottal pressures needed for vocalizing.

ACKNOWLEDGMENTS

Correspondences can be addressed to Wim Pouw (wim.pouw@donders.ru.nl). We would like to thank the Jaderpark Tier- und Freizeitpark an der Nordsee for allowing us to record audiovisual data of the Siamang family at their facility. This work has been supported by the Max Planck Institute (MPI) for Psycholinguistics Nijmegen, and the Donders Institute for Cognition, Brain, and Behavior. We would like to thank Jeroen Geerts of the MPI, for support of organizing the audiovisual equipment, and Maarten Snellen with his help setting up a dedicated server to run our dashboard application. We would like to thank Diandra Duengen for her support of making the recordings possible. WP is funded by a VENI grant (VI.Veni 0.201G.047: PI Wim Pouw) and was further supported by a Donders Postdoctoral Development fund. The Comparative Bioacoustics Group is supported by Max Planck Independent Research Group Leader funding to A.R. Center for Music in the Brain is funded by the Danish National Research Foundation (DNRF117).

DATA AVAILABILITY

All data and code supporting this manuscript can be found on github.

REFERENCES

[1] (2017). Amplitude envelope kinematics of speech signal: Parameter extraction and applications. In *Konferenz Elektronische Sprachsignalverarbeitung*, pages 107–113, Saarbrücken.

[2] Aruin, A. S. and Latash, M. L. (1995). Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements. *Experimental Brain Research*, 103(2):323–332.

[3] Berg, K. S., Beissinger, S., and Bradbury, J. (2013). Factors shaping the ontogeny of vocal signals in a wild parrot. *Journal of Experimental Biology*, 216(2):338–345.

[4] Berg, K. S., Delgado, S., and Mata-Betancourt, A. (2019). Phylogenetic and kinematic constraints on avian flight signals. *Proceedings of the Royal Society B: Biological Sciences*, 286(1911):20191083.

[5] Bertram, J. E. A. (2004). New perspectives on brachiation mechanics. *American Journal of Physical Anthropology*, Suppl 39:100–117.

[6] Blumberg, M. (1992). Rodent ultrasonic short calls: Locomotion, biomechanics, and communication. *Journal of Comparative Psychology*, 106(4):360–365.

[7] Boer, B. D. (2012). Air sacs and vocal fold vibration: Implications for evolution of speech. *Theoria et Historia Scientiarum*, 9(0):13–28.

[8] Boggs, D. F. (2002). Interactions between locomotion and ventilation in tetrapods. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 133(2):269–288.

[9] Bramble, D. and Carrier, D. R. (1983). Running and breathing in mammals. *Science*, 219(4582):251–256.

[10] Cheyne, S. M. (2011). Gibbon Locomotion Research in the Field: Problems, Possibilities, and Benefits for Conservation. In D'Août, K. and Vereecke, E. E., editors, *Primate Locomotion: Linking Field and Laboratory Research*, pages 201–213. Springer, New York, NY.

[11] Cooper, B. G. and Goller, F. (2004). Multimodal signals: Enhancement and constraint of song motor patterns by visual display. *Science*, 303(5657):544–546.

[12] Daley, M. A., Bramble, D. M., and Carrier, D. R. (2013). Impact loading and locomotor-respiratory coordination significantly influence breathing dynamics in running humans. *PLOS ONE*, 8(8):e70752.

[13] Deacon, T. W. (1998). *The Symbolic Species: The Co-Evolution of Language and the Brain*. W.W. Norton.

[14] Deacon, T. W. (2013). *Incomplete Nature: How Mind Emerged from Matter*. W.W. Norton.

[15] Dunn, J. C. (2018). Sexual selection and the loss of laryngeal air sacs during the evolution of speech. *Anthropological Science*, 126(1):29–34.

[16] Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press.

[17] Fitch, W. T., de Boer, B., Mathur, N., and Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *Science Advances*, 2(12):e1600723.

[18] Garcia, M. and Ravignani, A. (2020). Acoustic allometry and vocal learning in mammals. *Biology Letters*, 16(7):20200081.

[19] Geissmann, T. (1999). Duet Songs of the Siamang, Hylobates Syndactylus: II. Testing the Pair-Bonding Hypothesis during a Partner Exchange. *Behaviour*, 136(8):1005–1039.

[20] Haimoff, E. H. (1981). Video Analysis of Siamang (Hylobates syndactylus) Songs. *Behaviour*, 76(1/2):128–151.

[21] Harrison, D. F. N. (1995). *The Anatomy and Physiology of the Mammalian Larynx*. Cambridge University Press, Cambridge.

[22] Hayama, S. (1996). The origin of the completely closed glottis. Why does not the monkey fall from a tree? *Primate Research*, 12(2):179–206.

[23] Hodges, P. W. and Richardson, C. A. (1997). Feedforward contraction of transversus abdominis is not influenced by the direction of arm movement. *Experimental Brain Research*, 114(2):362–370.

[24] Hunt, K. D., Cant, J. G. H., Gebo, D. L., Rose, M. D., Walker, S. E., and Youlatos, D. (1996). Standardized descriptions of primate locomotor and postural modes. *Primates*, 37(4):363–387.

[25] Karashchuk, P., Rupp, K. L., Dickinson, E. S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B. W., and Tuthill, J. C. (2021). Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports*, 36(13):109730.

[26] Koda, H. (2016). Gibbon Songs: Understanding the Evolution and Development of This Unique Form of Vocal Communication. In Reichard, U. H., Hirai, H., and Barelli, C., editors, *Evolution of Gibbons and Siamang: Phylogeny, Morphology, and Cognition*, Developments in Primatology: Progress and Prospects, pages 349–359. Springer, New York, NY.

[27] Koda, H., Nishimura, T., Tokuda, I. T., Oyakawa, C., Nihonmatsu, T., and Masataka, N. (2012). Soprano singing in gibbons. *American Journal of Physical Anthropology*, 149(3):347–355.

[28] Lafortuna, C. L., Reinach, E., and Saibene, F. (1996). The effects of locomotor-respiratory coupling on the pattern of breathing in horses. *The Journal of Physiology*, 492 (Pt 2):587–596.

[29] Lancaster, W. C., Henson, O. W., and Keating, A. W. (1995a). Respiratory muscle activity in relation to vocalization in flying bats. *The Journal of Experimental Biology*, 198(Pt 1):175–191.

[30] Lancaster, W. C., Henson, O. W., and Keating, A. W. (1995b). Respiratory muscle activity in relation to vocalization in flying bats. *Journal of Experimental Biology*, 198(1):175–191.

[31] Liebal, K., Slocombe, K. E., and Waller, B. M. (2022). The language void 10 years on: Multimodal primate communication research is still uncommon. *Ethology Ecology & Evolution*, 0(0):1–14.

[32] Lieberman, P. (2003). Motor Control, Speech, and the Evolution of Human Language. In *Language Evolution*. Oxford University Press, Oxford.

[33] MacLarnon, A. M. and Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology*, 109(3):341–363.

[34] MacNeilage, P. F. (2010). *The Origin of Speech*. Oxford University Press.

[35] Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289.

[36] McAngus Todd, N. P. and Merker, B. (2004). Siamang gibbons exceed the saccular threshold: Intensity of the song of *Hylobates syndactylus*. *The Journal of the Acoustical Society of America*, 115(6):3077–3080.

[37] Micheletta, J., Engelhardt, A., Matthews, L., Agil, M., and Waller, B. M. (2013). Multicomponent and multimodal lipsmacking in crested macaques (*Macaca nigra*): Lipsmacking behavior in crested macaques. *American Journal of Primatology*, 75(7):763–773.

[38] Mitani, J. C. (1985). Gibbon Song Duets and Intergroup Spacing. *Behaviour*, 92(1/2):59–96.

[39] Mott, F. (1924). A Study by Serial Sections of the Structure of the Larynx of Hylobates syndactylus (Siamang Gibbon). *Proceedings of the Zoological Society of London*, 94(4):1161–1170.

[40] Negus, V. E. (1949). *The Comparative Anatomy and Physiology of the Larynx*. William Heinemann Medical Books, London.

[41] Nishimura, T., Tokuda, I. T., Miyachi, S., Dunn, J. C., Herbst, C. T., Ishimura, K., Kaneko, A., Kinoshita, Y., Koda, H., Saers, J. P. P., Imai, H., Matsuda, T., Larsen, O. N., Jürgens, U., Hirabayashi, H., Kojima, S., and Fitch, W. T. (2022). Evolutionary loss of complexity in human vocal anatomy as an adaptation for speech. *Science*, 377(6607):760–763.

[42] Pearson, L. and Pouw, W. (2022). Gesture–vocal coupling in Karnatak music performance: A neuro–bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences*, n/a(n/a).

[43] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Team, R. C. (2019). Nlme: Linear and nonlinear mixed effects models.

[44] Pouw, W., Burchardt, L. S., and Selen, L. (2023). Postural and muscular effects of upper-limb movements on voicing.

[45] Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., and Dixon, J. A. (2020a). Gesture-speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, 1491(1):89–105.

[46] Pouw, W. and Fuchs, S. (2022). Origins of vocal-entangled gesture. *Neuroscience & Biobehavioral Reviews*, 141:104836.

[47] Pouw, W., Harrison, S. J., Esteve-Gibert, N., and Dixon, J. A. (2020b). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, 148(3):1231–1247.

[48] Pouw, W., Paxton, A., Harrison, S. J., and Dixon, J. A. (2020c). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences*, 117(12):11364–11367.

[49] Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., Schaefer, R., and Wiggins, G. (2021). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

[50] Raemaekers, J. J., Raemaekers, P. M., and Haimoff, E. H. (1984). Loud Calls of the Gibbon (Hylobates lar): Repertoire, Organisation and Context. *Behaviour*, 91(1/3):146–189.

[51] Raimondi, T., Di Panfilo, G., Pasquali, M., Zarantonello, M., Favaro, L., Savini, T., Gamba, M., and Ravignani, A. (2023). Isochrony and rhythmic interaction in ape duetting. *Proceedings of the Royal Society B: Biological Sciences*, 290(1990):20222244.

[52] Rameau, F., Park, J., Bailo, O., and Kweon, I. S. (2022). MC-Calib: A generic and robust calibration toolbox for multi-camera systems. *Computer Vision and Image Understanding*, 217:103353.

[53] Redmond, J. and Lamperez, A. (2004). Leading limb preference during brachiation in the gibbon family member, Hylobates syndactylus (siamangs): A study of the effects of singing on lateralisation. *Laterality*, 9(4):381–396.

[54] Risueno-Segovia, C. and Hage, S. R. (2020). Theta synchronization of phonatory and articulatory systems in marmoset monkey vocal production. *Current Biology*.

[55] Suthers, R. A., Thomas, S. P., and Suthers, B. J. (1972). Respiration, wing-beat and ultrasonic pulse emission in an echo-locating bat. *Journal of Experimental Biology*, 56(1):37–48.

[56] Theriault, D. H., Fuller, N. W., Jackson, B. E., Bluhm, E., Evangelista, D., Wu, Z., Betke, M., and Hedrick, T. L. (2014). A protocol and calibration method for accurate multi-camera field videography. *Journal of Experimental Biology*, page jeb.100529.

[57] Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons.

[58] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. page 4.

[59] Zeileis, A. and Grothendieck, G. (2005). Zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14:1–27.