Vocal state change through laryngeal development

Yisi S. Zhang¹, Daniel Y. Takahashi¹, Diana A. Liao¹, Asif A. Ghazanfar¹,²,³,⁵* & Coen P.H. Elemans⁴,⁵*

Across vertebrates, progressive changes in vocal behavior during postnatal development are typically attributed solely to developing neural circuits. How the changing body influences vocal development remains unknown. Here we show that state changes in the contact vocalizations of infant marmoset monkeys, which transition from noisy, low frequency cries to tonal, higher pitched vocalizations in adults, are caused partially by laryngeal development. Combining analyses of natural vocalizations, motorized excised larynx experiments, tensile material tests and high-speed imaging, we show that vocal state transition occurs via a sound source switch from vocal folds to apical vocal membranes, producing louder vocalizations with higher efficiency. We show with an empirically based model of descending motor control how neural circuits could interact with changing laryngeal dynamics, leading to adaptive vocal development. Our results emphasize the importance of embodied approaches to vocal development, where exploiting biomechanical consequences of changing material properties can simplify motor control, reducing the computational load on the developing brain.

¹Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA. ²Department of Psychology, Princeton University, Princeton, NJ 08544, USA. ³Department of Ecology & Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. ⁴Department of Biology, University of Southern Denmark, 5230 Odense M, Denmark. ⁵These authors jointly supervised this work: Asif A. Ghazanfar, Coen P.H. Elemans. *email: asifg@princeton.edu; coen@biology.sdu.dk
Vocal development is marked by progressive changes in behaviors. The development of human speech, for example, begins with crying and protophones, continues through babbling and the production of speech-like words to, eventually, complex sentences. The brainstem central pattern generators (CPGs) encoding vocal motor output are well-documented brain areas in vertebrates (birds, bats, fish, and frogs), including monkeys and humans, and their modulation by higher-order brain areas is also known. However, can these circuits fully account for the dramatic changes in infant vocal output?

From 3 to 9 months of age, human infant vocalizations, transitioning from cries to cooing sounds, change in their fundamental frequencies and exhibit signatures of nonlinear dynamics such as noisiness and subharmonics. Later, infants can produce sounds without these chaotic elements, and the ability to do so has long been hypothesized to be due solely to increased neural control of the vocal organ. However, the larynx itself is also changing during postnatal development. Thus, an alternative hypothesis (though, not mutually exclusive) is that changes in infant vocalizations are related to changing laryngeal dynamics. Such hypotheses are difficult to test in humans. We thus lack experimental evidence that causally relates changes in vocal output with laryngeal change as a function of growth and development.

We used infant marmoset monkeys as a model to test the role of the changing larynx in the development of vocalizations. Like human infants, infant marmosets at the same life history stage spontaneously produce sequences of immature and mature vocalizations. Over the course of ~2 months, infant vocalizations exhibit changes in their acoustic properties that reflect a transition from producing mostly immature-sounding contact calls (i.e., cries) to mature contact calls (i.e., phees). This transformation of cries into adult-like contact calls includes (as they do in human infants) changes in their fundamental frequencies as well as signatures of nonlinear dynamics such as noisiness and subharmonics; these latter elements of instability disappear as the infants get older. We tested the hypothesis that the transition from cry-like states to mature-like states in the marmoset monkey vocal repertoire is due to the changing biomechanical properties of the developing larynx.

We show that vocal state changes in infant marmoset monkeys are in part the adaptive consequences of the gradually changing laryngeal properties. A sound source switch from the vocal fold (VF) proper to the apical vocal membrane (VM) causes the transition from noisy, low-frequency cries to tonal, higher pitched contact vocalizations. Because of its increased VM stiffness, the adult larynx can only produce loud, higher pitched vocalizations produced with higher efficiency. Furthermore, using an empirically based computational model of descending motor control, we show how neural circuits can interact with changing laryngeal dynamics, leading to adaptive vocal development. Our results emphasize the need for an embodied framework of vocal motor control, especially during postnatal development where the brain must coordinate its activity with a rapidly changing body.

The infant larynx has a greater acoustic capacity. To test if these attractors result from constraints in laryngeal dynamics that change over development, we investigated the vibratory behavior of excised larynges from infant (N = 3) and adult (N = 4) marmosets. Our setup allowed us to carefully and independently control both VF tension and subglottal pressure (see Methods). Our first specific hypothesis was that the difference between infant and adult vocalizations was that the amount of subglottal pressure needed to induce VF vibration is lower in infants than in adults. This hypothesis is derived from biomechanical models that simulate vocal production using two time-varying, lumped parameters—the subglottal pressure and laryngeal tension. These models all posit that the production of immature calls can be explained by a smaller subglottal pressure and VF tension. Thus, we wanted to know if infants have a lower VF tension because the minimum pressure needed for the infant larynx to produce sounds (aka the phonation threshold pressure, PTP) is lower than for adult larynges. We induced phonation in excised larynges of infants and adults, precisely controlling subglottal pressure and VF length. We quantified the PTP of vocalizations by applying subglottal pressure ramps from 0 to 3 kPa at 1 kPa s⁻¹. Our results falsify the model hypothesis by revealing that the PTP was significantly higher in the larynges of infants (2.05 ± 0.37 kPa) versus adults (0.62 ± 0.13 kPa) (p < 0.001, linear mixed-effects model) (Fig. 2c). These results show that the production of immature calls—the cry attractor state—by infant larynges cannot be explained by a lower PTP.

An alternative hypothesis is one of capacity, that the adult larynx simply cannot produce the cry attractor state like the infant larynx. That is, cry production is outside the available range of the adult vocal control space (Fig. 2a, right panel). If the capacity hypothesis is true, then the variation of acoustic features driven by the same change of control parameters should be greater for the infant larynx. To test this hypothesis, we applied sweeps through the vocal control parameter space by independently varying VF strain and subglottal pressure (Fig. 2d–g). The VF strain was calculated as the fractional change of VF length with respect to the resting length. The mean resting lengths of infant and adult VFs were 1.4 ± 0.2 mm and 2.8 ± 0.1 mm, respectively. In the infant larynx, sounds produced at low VF strain were very similar to the infant marmoset cry (Fig. 2d, left panel; Supplementary Fig. 1). At a specific VF length change...
corresponding to 17–45% strain, we observed a sharp transition in the $f_0$ in each individual from ~2 to ~9 kHz, the latter corresponding well to the mature-sounding contact call (Fig. 2d, left panel). The sound pressure phase portrait suggests a transition between two distinct vibratory modes (Supplementary Fig. 2). In contrast, in the adult, $f_0$ transitions were never observed (Fig. 2d, right panel). The different behaviors of the infant and adult larynges yielded distinct $f_0$ maps in the pressure–strain space, in which a steep $f_0$ transition exists in the infant map that is absent in the adult map (Fig. 2f, g). At very low strain and high pressure (>5 kPa), both infant and adult larynges entered a chaotic acoustic region featured by high WE value that did not correspond to any natural marmoset monkey calls. We also assessed the contribution of control parameters to the variation of acoustic features and found that $f_0$ control is redundant with respect to pressure: $f_0$ is primarily driven by VF length. Infant larynges yielded greater variation than adult larynges in the control space for both $f_0$ (Fig. 2h; $p < 0.001$, test for slope equal to 1) and WE (Fig. 2i; $p < 0.001$, test for slope equal to 1). These findings support the hypothesis that the cry attractor state is outside the vocal capacity of the adult larynx.

**Changing VF material properties.** VF material properties are a likely candidate to induce laryngeal changes at developmental timescales, but we lack the knowledge of how VF material properties could change over time. While VF vibratory kinematics, and laryngeal acoustic output, results from the complex interplay of viscoelastic properties, fluid flow, and acoustics, modeling the VF as a 1-dimensional vibrating string can, at a first approximation, partially explain $f_0$ ranges observed across a variety of species, even after accounting for body size. Such string models predict that $f_0$ is proportional to tissue stress over density and inversely proportional to VF length. VF stress typically increases nonlinearly with strain, which allows an extension of $f_0$ by two different mechanisms. First, as the maximum VF length is constrained by laryngeal geometry, nonlinearly increasing VF stress with strain allows an upward extension of $f_0$ range. Second, any nonlinearity in the dynamical

---

Fig. 1 Infant and adult marmoset vocalizations fall into different attractors in the acoustic space. a Spectrograms of infant (top) and adult (bottom) marmoset undirected vocalizations. b Exemplars of three different contact calls: Phee, Cry, and Phee-cry. Vertical lines in the spectrogram indicate the segment used for calculating the power spectral density (PSD) below. c, d The acoustic features of infant and adult contact calls in the Wiener entropy (WE)–$f_0$ space. e, f The max-scaled probability density of infant and adult vocal acoustic features in the WE–$f_0$ space. Source data are provided as a Source Data file.
Fig. 2 The infant larynx has greater vocal capacity than the adult larynx. a Two alternative hypotheses for infant cry production. Threshold hypothesis: both the infant and the adult larynxes have the same vocal range, but the adult larynx starts phonation at a higher threshold. Capacity hypothesis: the infant larynx has greater capacity than the adult larynx. Color indicates f0. b Excised larynx setup: an infant marmoset larynx mounted on an artificial trachea with arytenoid cartilage adducted and the position of thyroid cartilage manipulated by the lever arm of an ergometer. c Comparison of phonation threshold pressure between infant and adult excised larynges. d Exemplars of sound generated from an infant and an adult larynx while the vocal folds were stretched at fixed pressures. From top to bottom: the real-time vocal fold strain and subglottal pressure, sound pressure, and sound spectrogram. e f0 maps of an infant and an adult larynx in the control space of pressure vs. vocal fold strain. f, g Mean f0 maps (f) and mean WE maps (g) of infant (n = 3) and adult (n = 4) larynges. Iso-f0 contours (in Hz) and iso-WE contours (in dB) are overlaid, showing that the pressure control is redundant. h, i Comparisons between the infant and adult f0s (h) and WEs (i) generated at the same pressure and vocal fold strain settings. Slopes significantly greater than 1 indicate greater capacity of the infant larynx than the adult larynx. Source data are provided as a Source Data file.
system that sets the VF vibration kinematics can lead to multiple discrete and stable attractor states of VF vibratory patterns22,27.

We measured the stress–strain properties of VF tissues in infant and adult larynges (see Methods). A salient part of the VF geometry in some species are thin apical extensions called VMs28,29. Their function in vocal production remains untested experimentally but it has been suggested that they act as low mass oscillators that can vibrate almost independently of the VF proper28,29, and thus support, for instance, the production of high-frequency vocalizations29,30. We therefore quantified the stiffness of VF proper and VM separately. The stress–strain curves were highly nonlinear and could be fitted with combined linear and exponential models at low and high strain regions (Fig. 3b). The linear elastic limit was around 12–20% strain (fitted parameters summarized in Table 1). Overall, for both infant and adult larynges, as the strain increased, the VM stress increased more rapidly than the VF stress (p < 0.001, linear model). Following the string model, at a strain of 50%, both the infant and adult VMs can yield a 3–4 times higher f0 than the VF. Consistent with predictions for the role of the VM, this would enable high-frequency vocalizations at a greater efficiency30. However, it does not explain the f0 magnitude difference between the observed immature and mature call attractors. The stress–strain responses between infant and adult larynges were significantly different between their VMs (p < 0.001, linear model), but not between their VFs (p = 0.42, linear model). Thus, the maturation of larynx is driven at least in part by a viscoelastic change in the VM. The VM trajectories (Fig. 3b) show that in low strain regions, the stress of the infant VM only starts deviating from the VF at ~17% up to ~45% strain, much later in the trajectory than that of the adult VM (3.9% strain; p < 0.001, bootstrap). The infant VM exhibited a similar elasticity as its VF at a low strain region. The divergent point around 17% also coincides with the strain where the frequency jump occurred in the phonation tests of the infant larynx. In contrast, because the stress–strain responses of the adult VM and VF diverge almost immediately; their vibratory modes become distinct very quickly.

**Shift to VM vibrations in adult contact calls.** We speculated that it is this divergence in the elastic properties between the VF proper and the VM that accounts for a switch between distinct vibratory modes. To test this, we quantified glottal dynamics during sound production using a high-speed camera (Fig. 3c–f). During the sub-1-kHz sound produced by the excised infant larynx at low strain (<15%, i.e., relaxed state), we observed vibration with the caudal edges (VF proper) always closing prior to the cranial edges (the VM) (Fig. 3c, top two panels; Supplementary Movie S1). This phase difference along the caudo–cranial axis reflects energy transfer through the tissue33,36. Sound excitation occurred during glottal opening and closing (Fig. 3c, bottom panel). However, when the VF and VM were stretched (strain >15%, i.e., stretched state), glottal area dynamics were defined solely by the vibration of the VM (Fig. 3d; Supplementary Movie S2); no underlying VF vibration could be observed. Thus, the rapid transition from low to high f0 was caused by a distinct shift of sound production from the VF to the VM. This transition was accompanied by a shift from relatively complex to regular vibratory patterns (Fig. 3d). In the adult larynx, glottal area dynamics was set exclusively by regular vibration of the VM (Fig. 3e, f; Supplementary Movie S3), which is consistent with our finding that VM stress deviated from VF stress at a low strain in the adult larynges (Fig. 3b).

The adult larynx is more vocally efficient. Our data show that the function of the VMs—a morphological innovation observed in New World monkeys, cats, and bats—is to (i) increase the f0 range and (ii) decrease PTP, corroborating previous theoretical predictions22,30,37. Because decreased PTP is indicative of increased vocal efficiency (leading to higher amplitude vibrations for a given lung pressure)30, vocalizations could theoretically be produced louder with less energy, thereby extending their communicative range. To test this, we calculated the sound source level (SL) as a function of pressure and VF (VF proper and VM) strain. We found that the SL of the emitted sound was positively correlated with pressure as well as strain in both infant and adult larynges (p < 0.001 for slopes of multiple linear regression; Fig. 4a). This is consistent with the prediction that the VM—the source oscillator at the higher strain levels—is more energy efficient than the VF proper. Overall, the adult larynx produced sound about 15 dB louder than the infant larynx across all sets of parameters (Fig. 4a). These findings from the excised larynx preparation are consistent with what is observed in naturally produced vocalizations (Fig. 4b). The infant’s mature-sounding contact calls are generally louder than its cries, and the adult contact calls are louder than infant’s contact calls (p < 0.001, ANOVA). Assuming sound transmission loss by spherical spreading38, the 15 dB loudness increase of the contact call from infant to adult results in a roughly six times larger communicative distance. To produce louder sounds, the adult larynx may simply draw more mechanical energy or it might be more efficient in converting mechanical energy to sound. To test this, we estimated its mechanical efficiency (ME): the fraction of the aerodynamic power that was converted to acoustic power39. We found that the adult larynx was more efficient at all parameter settings (Fig. 4c), demonstrating that ME increases over development. In addition, the ME was positively correlated with strain for both infant and adult larynges (p < 0.001 for slopes of multiple linear regression; Fig. 4d). Thus, it is more energy efficient to produce high f0 sounds when energetic constraints allow for them.

**Synchronization of laryngeal and respiratory output.** Our findings do not address the role of the concurrently changing nervous system—the vocal CPGs3,4 and descending control5. In the case of marmoset monkeys, this is important because infants raised without parental care continue to produce noisy, cry-like vocalizations well beyond the age when such calls should have disappeared from the vocal repertoire40,41. These animals are not stunted in their bodily growth and it is unlikely that their laryngeal development is dependent upon parental care. It is therefore logical to infer a role for experience-dependent neural control.

To define a neuromechanical framework for embodied vocal motor control, we reverse-engineered the potential real-time control parameters—pressure and strain—that change over development (Fig. 5a). This was possible because the control space mapped from excised larynges matched the range of in vivo marmoset vocalizations (Fig. 2f, g). The probability density maps of the control parameters show two and one attractors present in the infant and adult control space, respectively (Fig. 5b). For infants, the two attractors occur at 10 and 45% strain at subglottal pressure ~3 kPa (Fig. 5b, left panel). These attractors correspond to the production of immature and mature calls, respectively. For adult vocalizations, the attractor at 45% strain and ~5 kPa pressure represents the production of mature contact calls (Fig. 5b, right panel). Thus, in addition to the inability to produce vocalizations with a f0 < 1 kHz, adults also do not produce contact vocalizations at low strain.

When these two control parameters are in phase, mature sounding contact calls are produced in both infant and adults (Fig. 5c). Infant marmosets, however, produced immature cry
Table 1 Stress–strain response fitting parameters

<table>
<thead>
<tr>
<th></th>
<th>( a ) (kPa)</th>
<th>( b ) (kPa)</th>
<th>( A ) (kPa)</th>
<th>( B ) (kPa(^{-1}))</th>
<th>( \epsilon_0 ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infant VF</td>
<td>(-0.19 \pm 0.21)</td>
<td>(3.13 \pm 2.15)</td>
<td>(0.10 \pm 0.08)</td>
<td>(8.73 \pm 3.75)</td>
<td>(18.4 \pm 6.4)</td>
</tr>
<tr>
<td>Adult VF</td>
<td>(-0.18 \pm 0.13)</td>
<td>(4.19 \pm 2.97)</td>
<td>(0.27 \pm 0.18)</td>
<td>(5.94 \pm 0.19)</td>
<td>(21.1 \pm 0.7)</td>
</tr>
<tr>
<td>Infant VM</td>
<td>(-0.31 \pm 0.30)</td>
<td>(5.77 \pm 5.33)</td>
<td>(0.15 \pm 0.21)</td>
<td>(13.65 \pm 3.34)</td>
<td>(13.1 \pm 2.6)</td>
</tr>
<tr>
<td>Adult VM</td>
<td>(-0.83 \pm 0.52)</td>
<td>(22.71 \pm 10.77)</td>
<td>(1.26 \pm 0.66)</td>
<td>(7.23 \pm 0.57)</td>
<td>(17.4 \pm 1.7)</td>
</tr>
</tbody>
</table>

The stress–strain response was modeled with a linear model at low strain and an exponential model at high strain: 
\[
\sigma = \begin{cases} 
  a + b \epsilon, & \epsilon \leq \epsilon_0 \\
  A e^{B \epsilon}, & \epsilon > \epsilon_0 
\end{cases}
\]
where \( \epsilon_0 \) is the linear limit.
calls when the VF strain lagged pressure at >90° (Fig. 5c). The in-phase control of pressure and VF strain that produces mature calls indicates synchronous activity of the laryngeal and respiratory muscles. To test this, we recorded electromyography signals from intramuscularly implanted electrodes in the cricothyroid (CT, i.e. the laryngeal muscle controlling VF length) and surface electrodes over the abdominal muscles (controlling subglottal pressure) in one adult marmoset. We observed that during mature calls, the CT and expiratory muscle activity was highly synchronized (Fig. 5d). The phase difference between expiration and CT muscle activity was near zero (0.00 ± 0.58 rads, mean ± std, n = 87 calls; Fig. 5e), supporting the idea that the production of mature contact calls is due to strongly synchronized laryngeal (representing strain) and respiratory (representing pressure) muscle activities.

Such phase-locked oscillations can emerge in the dynamics of coupled oscillators. To show how the synchronized laryngeal–respiratory control may arise over the course of development (perhaps, facilitated by social reinforcement), we generated a coupled oscillator model of laryngeal–respiratory CPGs, where their synchronization is positively correlated with their degree of coupling. To account for the higher order control that may coordinate vocal CPG output, the model CPGs received a common input through which the CPGs are indirectly coupled. Within a range of parameter settings, the common input itself oscillates and forces the CPGs to oscillate at the same frequency with a fixed phase difference. The degree of coupling, negatively related to the phase difference between the CPGs, is controlled by the strength of this input (Fig. 5f). Based on this model, shifts in phase difference towards zero between pressure and strain over the first 2 months of life require increases in the input strength from this putative higher order area to the CPGs (Fig. 5h, i).

**Discussion**

Understanding the mechanisms underlying vocal behaviors remains a major challenge, and investigations of various species have revealed key neurobiological principles underlying vocal development but without considering the role of the changing body. Across vertebrate taxa, changes occur during vocal development in a variety of parameters such as decreases in fundamental frequencies, and increasing syllable repetition rates, but the contribution of the body—the vocal organ—remains unknown. Our work shows how gradual developmental changes in the vocal organ itself can induce discrete state changes in vocal behavior. In infant marmoset monkeys, the transition between the two attractor states—from producing noisy, low fundamental frequency cries to tonal, higher pitched contact calls—occurs via a sound source switch from the VF proper to the VM. The adult larynx does not have this capacity. Because of its increased VM stiffness, the adult larynx only produces one attractor state characterized by stable, higher pitched sounds. Such changes in the material properties of the vocal organ, and its biomechanical consequences, can be used to locally ‘compute’ solutions, thereby significantly simplifying motor control and reducing the computational load, and thus energy consumption, of the developing brain. Our work thus demonstrates the need for an embodied framework of vocal motor control, especially during postnatal development where the brain is navigating a rapidly changing body.

The mechanisms that account for the change in material properties of the developing marmoset larynx are unknown. In
Vocal maturation requires the development of motor control. An exemplar showing the reverse-engineering of vocal fold strain and subglottal pressure from a sequence of infant vocalizations. The probability density of the input strength during the production of infant and adult vocalizations. The distributions of the phase difference between pressure and vocal fold strain during the production of infant and adult vocalizations. EMG recordings at the right cricothyroid (CT) muscle and the abdominal muscle (for expiration) during adult phee call production. The amplitude modulation of the EMG signals is shown in bold (a.u., arbitrary units). The distribution of the phase difference between the amplitude modulation of the expiratory muscle activity (corresponding to positive subglottal pressure) and the CT muscle activity. A model for oscillators representing the respiratory and the laryngeal CPGs coupled with a common vocal control input. The probability density of the phase difference between pressure and vocal fold strain during the production of infant and adult vocalizations.

Our study focused on the contact call development—from cries to phees—of marmoset monkeys. This is because it is the only call in their repertoire that undergoes a significant transformation in acoustic properties during the course of development. It’s long duration, loudness and clear tonality (in mature versions) makes them uniquely difficult, more energetically costly, to produce when compared to other calls in the species repertoire. Thus, while marmoset monkeys have a rich repertoire of vocalizations, all but the contact call sound adult-like at the beginning of postnatal life. A number of call types, such as ‘trills’, ‘trill-threes’, and ‘twitters’, have a similar fundamental frequencies as the contact call. However, these other call types are all very short in duration and without a sustained, tonally flat contour (that is initially noisy). Other call types in the species repertoire can sometimes have fundamental frequencies above 10 kHz (‘tsik’ calls) as well as below 1 kHz (like the ‘egg’ calls); these calls are
very short in duration\textsuperscript{59}. Our study did not attempt to account for the production of these call types; there is no reason to think that the changing material properties of the larynx prevent the production of these other call types via a different set of laryngeal control dynamics. For example, the sub-1-kHz calls may be produced when the vocalis muscles within the VFVs are shortened, a possibility that cannot be simulated in our experimental setup. Along similar lines, human VFVs can exhibit several stable modes of vibration with different \( f_c \) ranges, such as chest and falsetto regime\textsuperscript{30}. We cannot exclude that multiple vibratory modes in VF or VM are possible as observed in the human larynx. For contact call development in marmosets, however, our data and neural modeling suggest that the combination of changing laryngeal material properties (switching to VMs as the sound source) and the development of neural control of respiratory and laryngeal synchronization that transforms them from long-duration, noisy cries into long duration, tonal phee calls.

We provided the first experimental evidence for the role that VMs play in vocal production. Our finding that that they act as low-mass oscillators to produce high frequency vocalizations is consistent with previous hypotheses\textsuperscript{29,31,47}; that these calls are lower and more efficiently produced in our study is consistent with model predictions for the role of VMs in mammals\textsuperscript{30}. Although some coupling between VF and VM cannot be ruled out, our observations support earlier suggestions that apical VMs vibrate relatively independently from the VF to aid extremely high-frequency calls, allowing an even wider fundamental frequency range\textsuperscript{29,32,47}. The VM likely also produces the high frequency vibration during echolocation and social calls in bats\textsuperscript{61}

Methods

Along similar lines, human VFVs can exhibit several stable modes of vibration. For example, the sub-1-kHz calls may be produced when the vocalis muscles within the VFVs are shortened, a possibility that cannot be simulated in our experimental setup. Along similar lines, human VFVs can exhibit several stable modes of vibration with different \( f_c \) ranges, such as chest and falsetto regime\textsuperscript{30}. We cannot exclude that multiple vibratory modes in VF or VM are possible as observed in the human larynx. For contact call development in marmosets, however, our data and neural modeling suggest that the combination of changing laryngeal material properties (switching to VMs as the sound source) and the development of neural control of respiratory and laryngeal synchronization that transforms them from long-duration, noisy cries into long duration, tonal phee calls.

We provided the first experimental evidence for the role that VMs play in vocal production. Our finding that that they act as low-mass oscillators to produce high frequency vocalizations is consistent with previous hypotheses\textsuperscript{29,31,47}; that these calls are lower and more efficiently produced in our study is consistent with model predictions for the role of VMs in mammals\textsuperscript{30}. Although some coupling between VF and VM cannot be ruled out, our observations support earlier suggestions that apical VMs vibrate relatively independently from the VF to aid extremely high-frequency calls, allowing an even wider fundamental frequency range\textsuperscript{29,32,47}. The VM likely also produces the high frequency vibration during echolocation and social calls in bats\textsuperscript{61}.

In conclusion, our study established that vocal state changes in infant marmoset monkeys are in part adaptive consequences of the changing laryngeal properties. The transition from low frequency cries to high-pitched phee calls is through a sound source switch from the VF to the apical VM, allowing the animal to produce louder vocalizations with higher efficiency. These results suggest that the computational load on motor control can be reduced by the optimal solutions provided by the peripheral biomechanics. For example, in addition to the material change to the larynx over the development, our neural model posits that the strength of an input from a higher order structure increasingly synchronizes laryngeal and respiratory CPGs resulting in the generation of mature contact calls. The strength of this input could be modulated by contingent social reinforcement by parents: infant marmosets who receive greater contingent responses from their parents produce fewer cries and more mature-sounding contact calls quicker\textsuperscript{17,43}, and infants who do not receive parental care continue to produce cries\textsuperscript{40,41}.

Data acquisition. The audio signal was recorded with a ½-inch pressure microphone (46AD with pre-amplifier type 26AH, G.R.A.S., Denmark), amplified and high-pass filtered (above 10 Hz, 3-pole Butterworth filter, model 12AQ, G.R.A.S. Denmark). The microphone was calibrated before every experiment. The microphone was positioned close to the larynx, the distance measurements corresponding to maximal stretches were also marked. As the stretching process was smooth, we resampled the sparsely measured VF length at 357 Hz using the cubic spline method with no smoothing.

To quantify the glottal areas enclosed by different depths of the VF, we used a supervised learning workflow implemented in ilastik (http://ilastik.org)\textsuperscript{9}. In short, we trained a random forest model to classify the cranial edges (the VM), the caudal edges (the VF) and the rest of the larynx. To calculate the enclosed areas bounded by VM and VF, we first skeletonized the detected pixels of the edges using MATLAB routine bwmorph, and then identified the boundaries of the edges using MATLAB routine boundary. For frames where the caudal edges were occluded by the cranial edges, we set the two areas equal.

Phonation threshold pressure. To measure the PTP, we adducted the arytenoid cartilages and provided the larynx with ramping pressure while leaving the VF unstretched. The larynx were subject to ramping pressure up to 3.0 kPa at a rate of 1.0 kPa s\textsuperscript{-1}. PTP was identified as the pressure where the sound was first produced on the ramping up phase. A mild hysteresis was observed during the down phase. In this study, we only took the up phase into account.

Control space construction for acoustic maps. To systematically explore acoustic output in the control space of subglottal pressure and VF strain, we subjected the larynx to a set of subglottal pressure between 0 and 5 kPa with 0.5-kPa intervals approved by the Princeton University Institutional Animal Care and Use Committee (Protocol #1908–18).

Experiment setup. The larynx together with the trachea and the tongue were carefully extracted from the deceased animal. The day before the larynx experiment, the tissue stored in —80 °C was gradually warmed up in refrigerator over-night and was submerged in Ringers solution (5 °C)\textsuperscript{62}. It was then transferred to a Sylgard-coated petri dish. The larynx was separated from the tongue, fat and tracheal cartilages. It was mounted with the rostral side up onto a polyethylene tube via the trachea and was secured using 5/0 silk suture. To expose the VF, the hyoid bone, the epiglottis, and the vestibular fold were removed. The larynx was surrounded with Ringers-solution-soaked Gelfoam to be kept humidified. To mark the position of the VF, two knots were made on the two ends of the VF membrane using 10/0 silk suture.

The tube coupled with the trachea was oriented vertically and was connected to the air supply system with bronchial pressure controlled by dual valve differential pressure PID controllers, with respect to atmospheric pressure. Airflow was measured with mass flow sensors (PMF series, Posifa Microsystems, San Jose, USA). The pressurized air supplied to the controllers was warmed up to 37 °C and humidified with deionized water in a pressure cooker. The tip of the ventral side of the thyroid cartilage was attached to the lever arm of the ergometer (Model 300C, Aurora Scientific, Ontario, Canada) via 5/0 suture threaded through a small hole made on the cartilage. The displacement of the lever was controlled to move in a negative sawtooth waveform, with the first 80% cycle stretching the VF, and the last 20% cycle back to the zero-point at a rate of 2 Hz. To visualize the VF lengthening, a 5 × 5 × 5 mm 45°-angled aluminum surface-coated prism (Thorlabs, Newton, New Jersey, USA) was placed on the right side of the larynx. The right side of the thyroid cartilage was removed to reveal the stereoscopic view of the VF in the process to cumulate enough subglottal pressure. Power VF vibration, the arytenoid cartilages were adducted by two bent 27-gauge needles mounted on the micro-manipulators (model MM3, Narishige, Japan).

Image acquisition and analysis. To measure the dimension of the VF, a camera was mounted on a Leica M165-FC stereoscope (Leica Microsystems, Wetzlar, Germany). To record the varying VF length during phonation, a high-speed camera (MotionPro-X4, 12-bit CMOS sensor, Photron, Tokyo, Japan) was mounted on the microscope and the images were acquired at a rate of 300 Hz. In addition, the laryngeal dynamics was acquired with an ultra-high-speed camera (FASTCAM SA1, 16-bit, Photron, San Diego, CA, USA) at a frame rate of 30 kHz. Light for illumination was provided by a plasma light source (HPLS200, Thorlabs, Newton, New Jersey, USA). A thin liquid light path was reflected off a 45° prism to reduce heat. Video acquisition was automatically initiated by a synchronization signal upon the start of audio acquisition with a pre-fixed delay.

The momentary VF length was measured from videos taken from the 12-bit camera. The pixel distance for each video was calibrated. The suture marks were hand labeled in the images from the view of the prism at every 18 frames. The distance measurements corresponding to maximal stretches were also marked. As the stretching process was smooth, we resampled the sparsely measured VF length at 357 Hz using the cubic spline method with no smoothing.

To quantify the glottal areas enclosed by different depths of the VF, we used a supervised learning workflow implemented in ilastik (http://ilastik.org)\textsuperscript{9}. In short, we trained a random forest model to classify the cranial edges (the VM), the caudal edges (the VF) and the rest of the larynx. To calculate the enclosed areas bounded by VM and VF, we first skeletonized the detected pixels of the edges using MATLAB routine bwmorph, and then identified the boundaries of the edges using MATLAB routine boundary. For frames where the caudal edges were occluded by the cranial edges, we set the two areas equal.

Data acquisition. The audio signal was recorded with a ½-inch pressure microphone (46AD with pre-amplifier type 26AH, G.R.A.S., Denmark), amplified and high-pass filtered (above 10 Hz, 3-pole Butterworth filter, model 12AQ, G.R.A.S. Denmark). The microphone was calibrated before every experiment. The microphone was positioned close to the larynx, the distance measurements corresponding to maximal stretches were also marked. As the stretching process was smooth, we resampled the sparsely measured VF length at 357 Hz using the cubic spline method with no smoothing.
while manipulating the VF length from 0 to 50% strain. By slowly changing the VF length (sawtooth at 2 Hz), we predict that VF vibratory behavior was in a quasi-steady state. The strain of the VF was calculated as \( \epsilon(t) = (L(t) - L_0)/L_0 \), where \( L_0 \) is the VF length at rest. The maximal strain was set to \(-50\%\), \( f_0 \), and WE were calculated on a sliding window of 12.8 ms with 10 ms overlap when sound was produced. The pressure and the VF strain corresponding to the midpoint of the sliding window were identified. To compare the acoustic range between infant and adult larynxes, we calculated \( f_0 \) and WE within grids of VF strain (bin size = 0.05) and pressure (bin size = 0.5 kPa). If their vocal ranges were the same, one would expect that the acoustic features within the same grid were the same and they should fall onto the 45° line of the infant vs. adult plots. We thus compared the slopes of those plots to 1. A slope >1 means a greater range in the infant larynx than the adult’s, and vice versa.

Elastic property measurement. The larynges were cut in half along the midline, and the VFs were dissected with a small portion of the arytenoid and thyroid cartilages attached on the two ends. The membrane was isolated from the VF. The specimen was mounted onto the two ergometer arms of a uniaxial force transducer (Model 400A, Aurora Scientific) using silk suture along the dorsal-ventral direction. The specimen was submerged in Ringer’s solution at a physiological pH of 7.4 and a temperature of 38 °C, resembling the marmoset body temperature. The elastic properties of the specimens were quantified by real-time measurement of force and displacement. The samples were preconditioned with 0.1 mm sinusoidal strain waves at 1 Hz of three cycles each starting position and then were subjected to 20 cycles of a sinusoidal displacement signal of 50% strain at 1 Hz. The force and displacement signals were filtered with third-order low-pass Butterworth filter with a cutoff frequency at 100 Hz.

The strain was defined similarly as above:

\[
e = \frac{\Delta L}{L_0}
\]  

where \( L_0 \) is the resting length and \( \Delta L \) the displacement. The stress was defined as:

\[
\sigma = \frac{F}{A}
\]  

where \( F \) is the force measured by the stretch apparatus and \( A \) the cross-sectional area of the specimen. Usually the volume of the VF is considered incompressible, the cross-sectional area must shrink as the tissue is stretched. The area as a function of strain can be estimated as:

\[
A = \frac{V}{L} = \frac{m}{\rho k_0 (1 + \epsilon)}
\]  

where \( m \) is the VF (or membrane) mass and \( \rho = 1.04 \times 10^3 \text{ kg m}^{-3} \) the commonly used density in VF studies. The stress–strain response was modeled with a linear model at low strain and an exponential model at high strain:

\[
\sigma = \frac{a + b \epsilon, \epsilon \leq \epsilon_0}{Ae^{b\epsilon}, \epsilon > \epsilon_0}
\]

where \( \epsilon_0 \) is the linear limit. To satisfy the continuous and differentiable requirements, we set \( a - b \epsilon_0 = Ae^{b\epsilon_0} \) and \( b = Ae^{b\epsilon_0} \), and thus the linear limit must have \( \epsilon_0 = \frac{b}{a - b} \). The parameters were determined by fitting the model using MATLAB routine lsqcurvefit.

To plot acoustic features in the pressure–stress space, we converted VF strain into stress using the fitted model for the VM stress–strain response.

Efficiency analysis. To estimate the SL of the emitted sound from the excised larynx, we calculated the root mean square (RMS) of the calibrated sound pressure within 12.8 ms sliding windows. The SL at 1-m distance was estimated using

\[
\text{SL} = 20 \log_{10} \frac{P_{0}}{P_{r}} = 20 \log_{10} \frac{P_{0}}{P_{t}}
\]

where \( P_{r} \) is the RMS of sound pressure, \( P_{t} = 20 \mu Pa \) is the reference pressure, and \( r \) the distance from the excised larynx to the microphone.

To calibrate the SL of marmoset vocalizations, a speaker broadcasting a constant 45-db SPL pink noise was placed in the room to mask occasional noises. The sound pressure of marmoset vocalization was calibrated to the background noise. A microphone was positioned at a distance of 0.76 m to the testing box where the marmoset was housed. The RMS of the calibrated sound pressure was used to calculate the SL using the above formula.

The ME was calculated using

\[
\text{ME} = 10 \log_{10} \frac{P_{\text{Acoustic}}}{P_{\text{Energetic}}}
\]

where the acoustic power in Watts \( P_{\text{Acoustic}} = IA \) and the aerodynamic power in Watts \( P_{\text{Energetic}} = \frac{1}{2} \rho V^2 \), \( V \) Here we used air density \( \rho = 1.2 \text{ kg m}^{-3} \), speed of sound \( v = 344 \text{ m s}^{-1} \), \( \rho \) the RMS sound pressure in Pa, \( P_t \) the subglottal pressure in Pa, and \( V \) the glottal flow rate in m s\(^{-1}\) (ref. 39,86).

Statistical analysis. To compare the stress–strain response between the VF and the VM, as well as the difference in the response due to age, we used a multiple linear regression model:

\[
in = a + b \epsilon + c \epsilon + d \text{ age} \epsilon + e \epsilon + f \text{ age} \epsilon + g \epsilon + h \text{ age} + i \epsilon + j \text{ age} + k \epsilon + l \text{ age} + m \epsilon + n \text{ age} + o \epsilon + p \text{ age} + q \epsilon + r \text{ age} + s \epsilon + t \text{ age} + u \epsilon + v \text{ age} + w \text{ age} + x \text{ age} + y \text{ age} + z \text{ age} + \text{ error}
\]

where (type (VF or VM) and age (adult or infant) are dummy variables, which were coded as 0 or 1 for each category.

To estimate the strain where the VM exceeded the VF in stress, we randomly resampled the stress–strain curves for each age and specimen type group with replacement for 1000 times and calculated the difference between the mean of the VM and VF curves for each repetition. The divergent point was determined as the point where the lower bound of the 95% confidence interval of the difference crossed zero. To estimate the mean and the uncertainty of this point, we repeated the procedure for 1000 times. The p-value corresponding to the difference in the mean divergent point between infant and adult was estimated by assuming normality of the distributions.

To compare the SL among infant cry, infant phee call, and adult phee call, ANOVA was performed in combination with Tukey’s honest significant difference post hoc test.

To test the SL (or ME) change in response to pressure and VF strain, we carried out multiple linear regression:

\[
\text{SL} = a + b \epsilon + c \epsilon + d \text{ age} \epsilon + e \epsilon + f \text{ age} \epsilon + g \epsilon + h \text{ age} + i \epsilon + j \text{ age} + k \epsilon + l \text{ age} + m \epsilon + n \text{ age} + o \epsilon + p \text{ age} + q \epsilon + r \text{ age} + s \epsilon + t \text{ age} + u \epsilon + v \text{ age} + w \text{ age} + x \text{ age} + y \text{ age} + z \text{ age} + \text{ error}
\]

where \( \epsilon \) is the VF strain and \( P_t \) the subglottal pressure, for infant and adult larynxes, respectively.

Control parameter reconstruction. The control parameters were estimated on a sliding window of 12.8 ms with 10 ms overlap. Within each window, \( f_0 \) and WE were calculated and then z-score-normalized. In the normalized \( f_0 \)–WE space, a tolerance circle of radius 1 (1-sigma away from the center) was created. Within this circle, the control parameters closest to the previous action in Euclidean distance were identified as the control parameters generating the concurrent sound. The control parameters were only evaluated during vocalization, and the initial conditions for VF strain and pressure were set to 0 and the PTP, respectively. The 2D probability density of the control parameters was constructed from all the monophasic control parameters.

To construct the polar density, the VF lengthening and pressure were converted to values between \(-n/2 \) and \( n/2 \) corresponding to no lengthening (or non-positive pressure) and maximum lengthening and pressure. The polar density was calculated using MATLAB exchange file circ_kdensity.

Intramuscular electromyography (EMG) surgical procedure. All procedures were approved by the Princeton University Institutional Animal Care and Use Committee. The implant surgery was performed under sterile conditions with the animal deeply anesthetized with isoflurane (1.0–2.0%) in oxygen gas. Vital signs including heart rate, respiratory rate, SPO2, and body temperature were monitored throughout the surgery. A small incision was made to expose the skull for connector attachment. A connector (Omnetics A75859) soldered with stainless steel wire electrodes (A-M Systems 79900) was secured on the surface of the skull with adhesive cement (C&B Metabond). Each pair of electrodes was attached to a 27-gauge needle tip with polyamide tubing protected. The animal was then rotated to expose the ventral side of the neck. A 2-cm-long incision between the pognion and the cranial end of the sternum was made on the midline ventral neck. To expose the cricothyroid muscles, subcutaneous fat, glandula mandibularis and sternohyoideus muscles were separated using blunt scissors. All electrode wires were passed from the skull subcutaneously. Around 3 cm from the tip on each pair of electrodes, a 1-mm opening was made on the insulation (as electrode contacts) with 2-mm separation between the two wires. A bead made from colored Kwiksil was placed next to the contact for visualization. The needle was passed twice through the fascia on top of the CT muscle using a needle driver till the beads touched the fascia and a suture knot was made to secure the wires in place. The needle tip and excessive wire were removed. After the electrodes implanted, the incisions were closed using 5/0 suture with a drop of Vetbond applied.

EMG signal acquisition and processing. The EMG signal of CT muscle activity was recorded from the intramuscular implant and that of the abdominal muscle activity was recorded from a pair of Ag–AgCl surface electrodes (Grass technology) attached to the torso of the subject with an elastic band. The EMG signal from each pair of electrodes was differentially amplified (250x) and the resulting signal was acquired by Plexon Omniplex System. The signal was notch-filtered at 60 Hz and digitized at a sample rate of 40 kHz.

The EMG signal was bandpass filtered between 100 Hz and 3 kHz using a zero-phase elliptic filter. The signal was rectified. The amplitude modulation of the EMG signal was calculated as the envelope of the rectified signal. Vocalizations during EMG recording were identified using threshold crossing. The EMG signal 0.5 s before and 0.5 s after vocal production was extracted.
To calculate the phase difference between abdominal and CT muscle contraction, we calculated the cross-correlation between the amplitude modulation of the two EMG signals around phee calls. The time delay \( \tau \) was determined as the location of the first peak of the cross-correlation. We then calculated the mean duration \( d \) of the utterances. The phase difference was then estimated as \( \delta \phi = 2\pi \omega t/d \).

**Coupled oscillator model.** We simulated the oscillatory dynamics for the CPGs of the laryngeal and respiratory activities using circular oscillators coupled with a common oscillatory input. The equations are given as follows:

\[
\begin{align*}
\dot{\theta}_1 &= \omega_1 + A_1 \sin(\phi - \theta_1) \\
\dot{\theta}_2 &= \omega_2 + A_2 \sin(\phi - \theta_2), \\
\phi &= \Omega,
\end{align*}
\]

where \( \theta_1 \) and \( \theta_2 \) are the respiratory and the laryngeal phases, \( \omega_{1,2} \) \( (\omega_1 > \omega_2) \) the intrinsic angular frequencies, \( A_{1,2} \) the input strength, and \( \phi \) the input phase and frequency, respectively. For illustration purpose, we let \( A_1 = A_2 = A \) and \( \Omega = \omega_{2,1} \) (these conditions are not necessary). When \( \Omega \) fixed, synchronization between the two oscillators occurs when \( \omega_2 = \omega_1 \). The phase difference between the two oscillators is thus \( \mu \left(\sqrt{1 - \mu^2} - \sqrt{1 - \mu^2}\right) \), where \( \mu_1 = \omega_1/\omega_2, i = 1, 2 \). Obviously, \( \theta_1 \) and \( \theta_2 \) tend to oscillate in phase as the input strength \( A \) increases. To estimate the control input strength for marmoset development data, we first calculated the phase difference between pressure and VF length for the vocalizations produced during the first 2 months. The values were then used to estimate \( A \). Here we let \( \omega_1 = 1, \omega_2 = 0.5 \), and set the boundaries for \( A \) between 0.25 and 3.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Data are provided for Figs. 1c, d, 2c, h, i and 4b in Source Data file. The remaining datasets and codes that support the findings of the current study are available in the DRYAD Digital repository (https://doi.org/10.5061/dryad.h7en66).

Received: 20 February 2019; Accepted: 13 September 2019; Published online: 09 October 2019

**References**


**Acknowledgements**

We like to thank Lasse Jakobsen for equipment loan. This research was supported by the NIH NINDS (R01NS054898) to A.A.G. and the Danish Research Foundation and NovoNordisk Foundation (NNF17OC0028928) to C.P.H.E.

**Author contributions**


**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41467-019-12588-6.

**Correspondence** and requests for materials should be addressed to A.A.G. or C.P.H.E.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2019