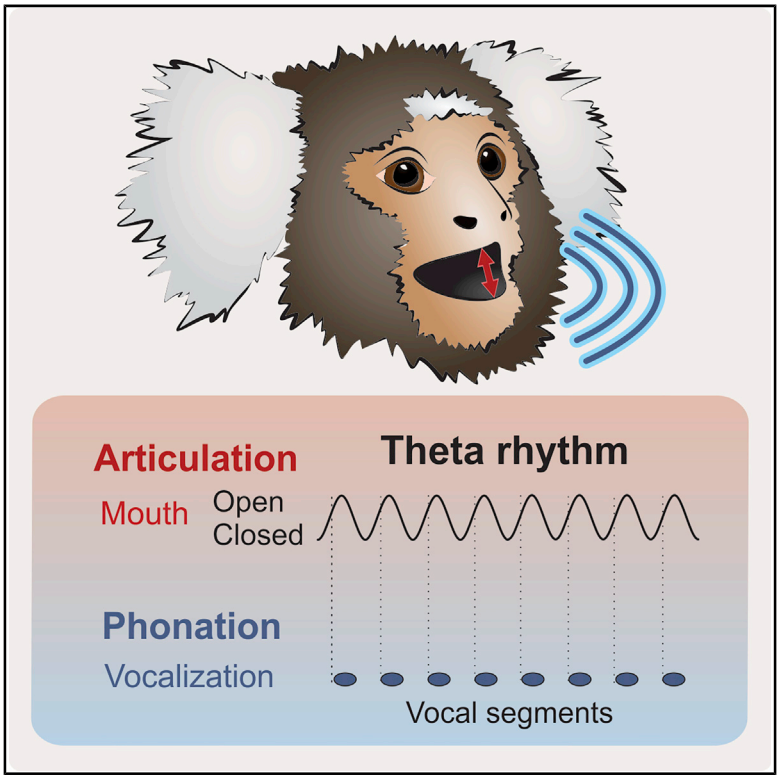
Report

Current Biology

Theta Synchronization of Phonatory and Articulatory Systems in Marmoset Monkey Vocal Production

Graphical Abstract



Highlights

- Marmosets exhibit vocalizations with coupled theta phono-articulatory oscillations
- The amplitude of segmented phee calls is phase locked to the inter-lip distance
- Human speech rhythms might be shared with the bi-motor rhythmicity of marmosets
- Rhythms underlying human speech might have evolved early in the primate lineage

Authors

Cristina Risueno-Segovia,
Steffen R. Hage

Correspondence

steffen.hage@uni-tuebingen.de

In Brief

Risueno-Segovia and Hage show that marmoset monkeys exhibit vocalizations with coupled phono-articulatory oscillations that are synchronized and phase-locked at theta rhythms, similar to those found in human speech. The findings suggest that these coupled oscillations, crucial for the emergence of human speech, evolved early in the primate lineage.

Risueno-Segovia & Hage, 2020, Current Biology *30*, 4276–4283
November 2, 2020
https://doi.org/10.1016/j.cub.2020.08.019

CellPress

CellPress

Current Biology

Report

Theta Synchronization of Phonatory and Articulatory Systems in Marmoset Monkey Vocal Production

Cristina Risueno-Segovia[1,2,3] and Steffen R. Hage[1,2,4,*]

[1]Neurobiology of Social Communication, Department of Otolaryngology, Head and Neck Surgery, Hearing Research Centre, University of Tübingen Medical Center, Elfriede-Aulhorn-Str. 5, 72076 Tübingen, Germany

[2]Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Otfried-Müller-Str. 25, 72076 Tübingen, Germany

[3]Graduate School of Neural & Behavioural Sciences - International Max Planck Research School, University of Tübingen, Österberg-Str. 3, 72074 Tübingen, Germany

[4]Lead Contact

*Correspondence: steffen.hage@uni-tuebingen.de

https://doi.org/10.1016/j.cub.2020.08.019

SUMMARY

Human speech shares a 3–8-Hz theta rhythm across all languages [1–3]. According to the frame/content theory of speech evolution, this rhythm corresponds to syllabic rates derived from natural mandibular-associated oscillations [4]. The underlying pattern originates from oscillatory movements of articulatory muscles [4, 5] tightly linked to periodic vocal fold vibrations [4, 6, 7]. Such phono-articulatory rhythms have been proposed as one of the crucial preadaptations for human speech evolution [3, 8, 9]. However, the evolutionary link in phono-articulatory rhythmicity between vertebrate vocalization and human speech remains unclear. From the phonatory perspective, theta oscillations might be phylogenetically preserved throughout all vertebrate clades [10–12]. From the articulatory perspective, theta oscillations are present in non-vocal lip smacking [1, 13, 14], teeth chattering [15], vocal lip smacking [16], and clicks and faux-speech [17] in non-human primates, potential evolutionary precursors for speech rhythmicity [1, 13]. Notably, a universal phono-articulatory rhythmicity similar to that in human speech is considered to be absent in non-human primate vocalizations, typically produced with sound modulations lacking concomitant articulatory movements [1, 9, 18]. Here, we challenge this view by investigating the coupling of phonatory and articulatory systems in marmoset vocalizations. Using quantitative measures of acoustic call structure, e.g., amplitude envelope, and call-associated articulatory movements, i.e., inter-lip distance, we show that marmosets display speech-like bi-motor rhythmicity. These oscillations are synchronized and phase locked at theta rhythms. Our findings suggest that oscillatory rhythms underlying speech production evolved early in the primate lineage, identifying marmosets as a suitable animal model to decipher the evolutionary and neural basis of coupled phono-articulatory movements.

RESULTS AND DISCUSSION

We measured vocal behavior in three adult marmosets (*Callithrix jacchus*), a highly vocal New World monkey species, using positive reinforcement to investigate the mechanisms underlying vocal rhythmicity. The monkeys produced a total of 3,449 calls, including phees (49.2%), chirps (35.3%), tsiks (3.2%), ekks (3.0%), and other vocalizations (9.3%). We focused on phees, a long-distance contact call, comprising one (so-called single phees), two (double phees), or more syllables [19] (Figure S1A). Long-duration, frequently uttered vocalizations provide an exceptional opportunity to characterize intrinsic vocal rhythms due to the presence of segmented and unsegmented phee syllables, a feature lacking from other calls such as twitters, chirps, tsiks, and ekks [20]. We analyzed 1,289 first phee syllables, which show great variability in call structure ranging from one segment, i.e., unsegmented phees (n = 469 calls, monkey L: 199, monkey P: 163, monkey C: 107), to segmented phees with up to 13 segments [20–22] (n = 820 calls, monkey L: 308, monkey P: 375, monkey C: 137), with a higher probability of segmented phees with fewer segments (Figures 1A and 1B). Segmented phees were recently discovered in marmosets and are produced during positive reinforcement [20] and in more naturalistic conditions, e.g., turn-taking behavior [21] and other vocal encounters in freely moving animals within marmoset colonies [22]. The vocal behavior of each monkey was recorded in daily 30-min sessions (40 sessions for monkey L, 74 for monkey P, and 15 for monkey C).

First, we analyzed segmented phees to characterize their intrinsic acoustic properties and verify that they are proper phees. As previously shown, the duration of phee segments varied according to their position in the phee call, with a significantly longer first segment [20]; the remaining segments show a rather precise and stereotyped duration distribution (Figures 1C and S1B;

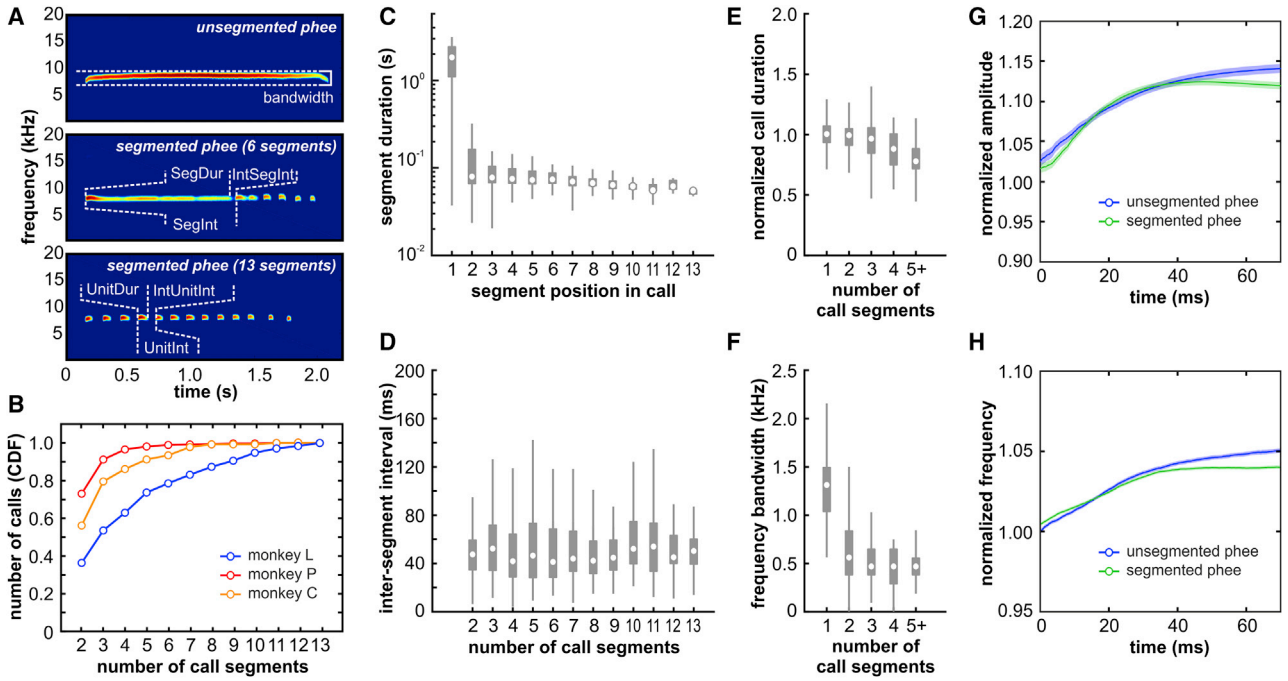


Figure 1. Normal and Segmented Marmoset Phee Calls Share Basic Characteristics

(A) Example spectrograms of degrees of phee segmentation ranging from unsegmented (top panel) to segmented phee calls (middle panel, six segments) with up to 13 segments (bottom panel). SegDur, segment duration; IntSegInt, inter-segment interval; SegInt, segment interval; UnitDur, phee unit duration; IntUnitInt, inter-unit interval; UnitInt, unit interval.

(B) Cumulative distribution function of segmented phee calls to the number of call segments within individual vocalizations for each monkey.

(C) Distribution of segment duration as a function of segment position within an individual call of all monkeys. Medians: white circle inside boxes; first and third quartiles: lower and upper margins of boxes, respectively; 0.4% and 99.6% quantile: end of whiskers below and above boxes, respectively.

(D) Distribution of inter-segment interval duration as a function of the number of segments within the corresponding call.

(E) Distribution of call duration as a function of the number of segments within the corresponding call.

(F) Distribution of call frequency bandwidth as a function of segments within the corresponding call.

(G) Mean amplitude trajectories (± SEM) for unsegmented and segmented phees.

(H) Mean frequency trajectories (± SEM) for unsegmented and segmented phees.

See also Figures S1 and S2.

n = 2,796, p = 1.0e-84; first segment with the rest, p < 2.86e-7; remaining segments among each other, p > 0.99; ANOVA with post hoc multiple comparison; first segment median = 1.63 s versus other segment medians between 53 and 76 ms). In contrast, inter-segment interval (ISI) durations remained stable, independent of the number of segments (Figures 1D and S1C; n = 1,869 segments, R = 0.047, p = 0.044; monkey L, n = 1058, R = 0.045, p = 0.143; monkey P, n = 540, R = −0.059, p = 0.179; monkey C, n = 271, R = −0.145, p = 0.017, Pearson's correlation; ISI median, 41–54 ms; while significant, the correlation coefficients were below or close to −0.1 in all animals, indicating little effect at the individual level). These findings indicate stable segmentation with precise and stereotyped vocal motor units once segmentation starts after the first segment of variable duration.

Second, we compared features between segmented and unsegmented phees to further determine whether segmented phees are proper phee calls and share similarities with unsegmented calls. The duration of segmented phees did not increase with a greater number of segments within the call, but remained rather stable and even exhibited a significant decrease in call duration relative to unsegmented phees (Figures 1E and S2A; pooled data; n = 1,289, R = −0.337, p = 1.7e-35, medians, 0.8–1.0; monkey L, n = 507, R = −0.555, p = 3.2e-42; monkey P, n = 538, R = −0.030, p = 0.492; monkey C, n = 244, R = −0.023, p = 0.725, Pearson's correlation). Next, we examined the distribution of frequency bandwidth as a function of degree of call segmentation and observed a significant decrease in calls' frequency bandwidth with increased segmentation (Figures 1F and S2B; pooled data: n = 1,289, R = −0.600, p = 1.4e-126, unsegmented phee median = 1.3kHz versus segmented phee medians, 470–560 Hz; monkey L: n = 507, R = −0.663, p = 1.6e-65; monkey P: n = 538, R = −0.643, p = 3.9e-64; monkey C: n = 244, R = −0.406, p = 4.1e-11, Pearson's correlation). While unsegmented phees generally possessed frequency bandwidths higher than 1 kHz, the frequency bandwidths of segmented phees were generally lower. Finally, we examined the trajectory of call amplitude and frequency within the first 70 ms after call onset to further determine the similarity between unsegmented and segmented phees (a window size of 70 ms was chosen to include the maximum number of segmented phees in the analysis due to the highly variable duration of the first segment, as seen in Figure 1C). We found a similar slope for call amplitude and frequency within the first 70 ms after call onset in both the pooled dataset (Figures 1G and 1H) and each individual

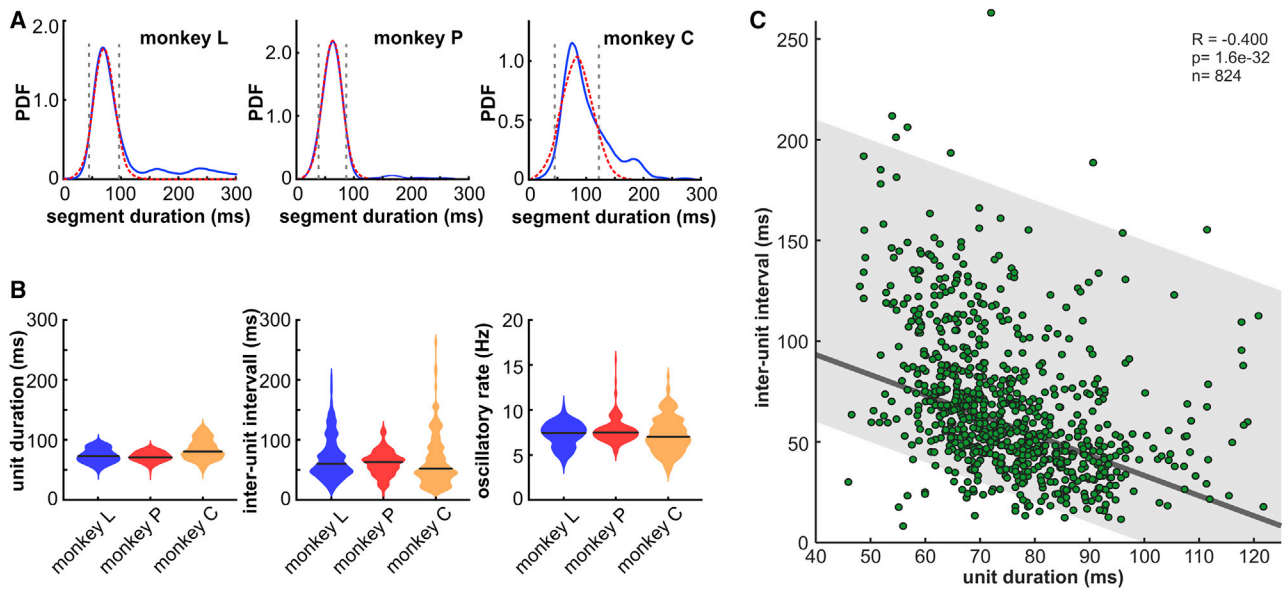


Figure 2. Phee Unit and Unit Intervals Build a Regular Oscillatory Cycle

(A) Distribution of segment durations (blue line). Phee unit duration (area between gray dashed lines) was defined for each individual monkey as the segment duration within one standard deviation from the mean of a fitted Gaussian curve (red line). PDF, probability density function.

(B) Violin plots depicting the distribution of unit duration, inter-unit interval, and oscillatory rate for each individual monkey. Black horizontal bars indicate the medians.

(C) Phee unit duration as a function of the respective inter-unit interval exhibits a significant indirect correlation for all monkeys (p = 1.6e-32, R = -0.40, n = 824, Pearson's correlation). The dark gray line indicates the combination of phee units and corresponding inter-unit intervals resulting in an oscillatory rate of 7.5 Hz. The gray area, which encompasses more than 95% of all data points, shows the combination of phee units and corresponding inter-unit intervals resulting in oscillatory rates within 4 and 10 Hz.

See also Figure S3.

monkey's dataset (Figures S2C and S2D; n for unsegmented phees, monkey L:199 calls, monkey P:163, monkey C:107; n for segmented phees, monkey L:308 calls, monkey P:375, monkey C:137). In summary, our results reveal that segmented phees are similar in basic call structure to unsegmented phees, indicating that common principles underlie their production mechanisms. Additionally, they show a highly stereotyped pattern, making them ideally suited to study the vocal rhythmicity underlying phee production.

Next, we investigated the vocal patterns underlying the production of segmented phees. We isolated vocal motor units from longer segments (mostly first call segments; see Figure 1C) based on the segment duration distribution by fitting a Gaussian. With this approach [23], we could extract proper phee units from multimodal distributions and exclude longer segments that might be a combination of more than one unit. Vocal motor units were defined as segments with durations within one standard deviation of the mean of a Gaussian curve fitted to the segment duration distribution (Figure 2A; n for phee units, monkey L: 808, monkey P: 492, monkey C: 213). Consequently, ISIs following a unit were considered inter-unit intervals (IUIs; Figure 1A). Cycle duration, i.e., unit interval, was defined as the time between the onsets of two consecutively uttered units. The duration of units and IUIs were similar between monkeys, giving rise to similar oscillation rates (Figure 2B, median unit duration: 72.9 ms for monkey L, 71.0 ms for monkey P, and 80.6 ms for monkey C; median IUI: 59.6 ms for monkey L, 63.0 ms for monkey P, and 52.4 ms for monkey C). The phee unit duration and consecutive IUI summed up to median cycle durations of 134.4 ms for monkey L, 133.5 ms for monkey P, and 142.6 ms for monkey C. From the cycle duration, the median oscillatory rate of the phonatory system was within the speech-related theta range: 7.44 Hz for monkey L, 7.49 Hz for monkey P, and 7.01 Hz for monkey C (Figure 2B). Statistical analyses revealed small yet significant differences in all three tested variables (unit duration: p = 7.9e-13; IUI: p = 0.025; oscillatory rate: p = 0.045; n = 824, Kruskal-Wallis test), indicating that marmosets exhibit a species-specific oscillatory rate during call pattern production that has small yet distinct differences between individuals.

Because of the variable distribution of unit durations and IUIs in all monkeys, we explored the capability of the monkeys to display a regular oscillatory rate between 4 and 10. Only an indirect correlation between these two variables would result in a regular, stable, quasi-periodic [11, 24–27] oscillatory rate; i.e., the presence of long units paired with short IUIs and vice versa (see Figure 1A). These two measures correlated indirectly, indicating a quasi-periodic oscillatory rate in phee production (Figures 2C and S3; pooled data: n = 824 units, R = −0.40, p = 1.6e-32; monkey L: n = 528, R = −0.60, p = 6.256e-52; monkey P: n = 137, R = −0.28, p = 1.093e-03; monkey C: n = 159, R = −0.11, p = 0.171, Pearson's correlation). Although the oscillatory rate was variable, more than 95% of data points were within 4–10 Hz, largely overlapping with the speech-related theta range (3–8 Hz) exhibited during syllable production in most human languages [1–3].

The acoustic call structure indicates that marmosets exhibit a stable, quasi-periodic oscillation during phee production. Next,

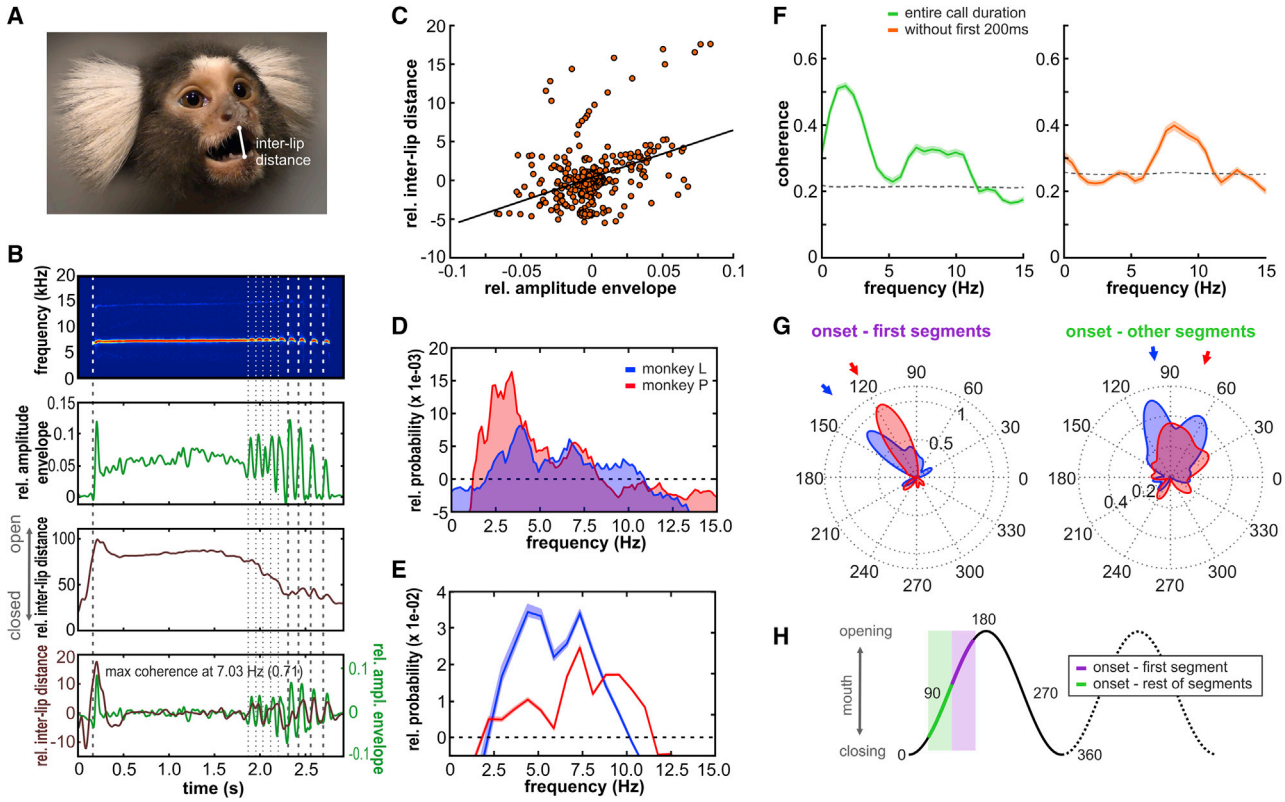


Figure 3. Coupled Phono-articulatory Movements at Theta Rhythmicity

(A) Schematic representation of inter-lip distances as a measure of jaw movement, extracted from video recordings of a vocalizing marmoset.

(B) Example spectrogram of a segmented phee of monkey P (top), relative call amplitude envelope (upper middle), relative inter-lip distance (lower middle), and overlapping call amplitude envelope and inter-lip distance, both high-pass filtered at 2 Hz (bottom). A maximum coherence of 0.71 between these two signals was found at 7.03 Hz. Segment onsets are indicated by thick vertical dashed lines. Onsets of rhythmical changes in call amplitude without full segmentation are indicated by thin dashed lines.

(C) Relative amplitude envelope power as a function of relative inter-lip distance shows a significant positive correlation (extracted from B).

(D) Detrended power spectra of vocalization-correlated jaw movements with significant spectral power in the theta range. Black dashed line denotes an $f^{-\alpha}$ fit to the data.

(E) Detrended mean power spectra (± SEM) of the amplitude envelope of recorded vocalizations with significant spectral power in the theta range. Black dashed line denotes an $f^{-\alpha}$ fit to the data.

(F) Mean coherence with 95% jackknife confidence interval between the inter-lip distance and the call amplitude envelope for monkey P. The coherence is plotted for the entire call (left panel) and without the first 200 ms (right panel), omitting the low-frequency components caused by mouth opening at call onset. Black dashed lines indicate significant levels.

(G) Distribution of phase angles at the onset of the first segment, i.e., call onset, and subsequent segments for the two monkeys. The height indicates the relative number of observations within the respective bin in the polar plot. Arrows indicate median angles. Phase angles of 0° and 360° indicate a minimum and 180° a maximum inter-lip distance within the oscillatory vocalization-correlated jaw movements.

(H) Schematic representation of the phase angles at the onset of the first segment, i.e., call onset, and at the onset of subsequent segments within an oscillatory cycle.

See also Figure S4 and Videos S1 and S2.

we investigated whether this pattern is exclusively auditory or whether the observed phonatory rhythms are synchronized with vocal-motor activity, such as orofacial movements. Therefore, we video-recorded monkeys L and P, who were trained in a primate chair during phee production, for a subset of their vocalizations. To represent articulatory movements during vocal production, we used the relative inter-lip distance, as performed previously [3] (Figure 3A). We examined exemplar phees from both monkeys to check for potential synchronizations between the inter-lip distance and acoustic structure of calls. After high-pass filtering both signals with a 2-Hz cutoff eliminating low-frequency fluctuations, distinct synchronization of call amplitude and inter-lip distance was observed, especially in the second half of phees in monkey P and last third of phees in monkey L, resulting in a maximum coherence of 0.69 at 8.20 Hz for monkey P and 0.71 at 7.03 Hz for monkey L (Figures 3B and S4A). Additionally, the call amplitude envelope and inter-lip distance directly correlated in both monkeys, indicating a direct phono-articulatory link (Figures 3C and S4B; monkey P: $p = 7.0e\text{-}15$, $R = 0.38$, $n = 388$; monkey L: $p = 3.44e\text{-}5$, $R = 0.25$, $n = 267$; Pearson's correlation). In the grouped data, the mean power spectra of the inter-lip distance and call amplitude envelope exhibited significant gains at theta rhythm frequencies (Figures 3D and 3E). We calculated the mean coherence of the call amplitude

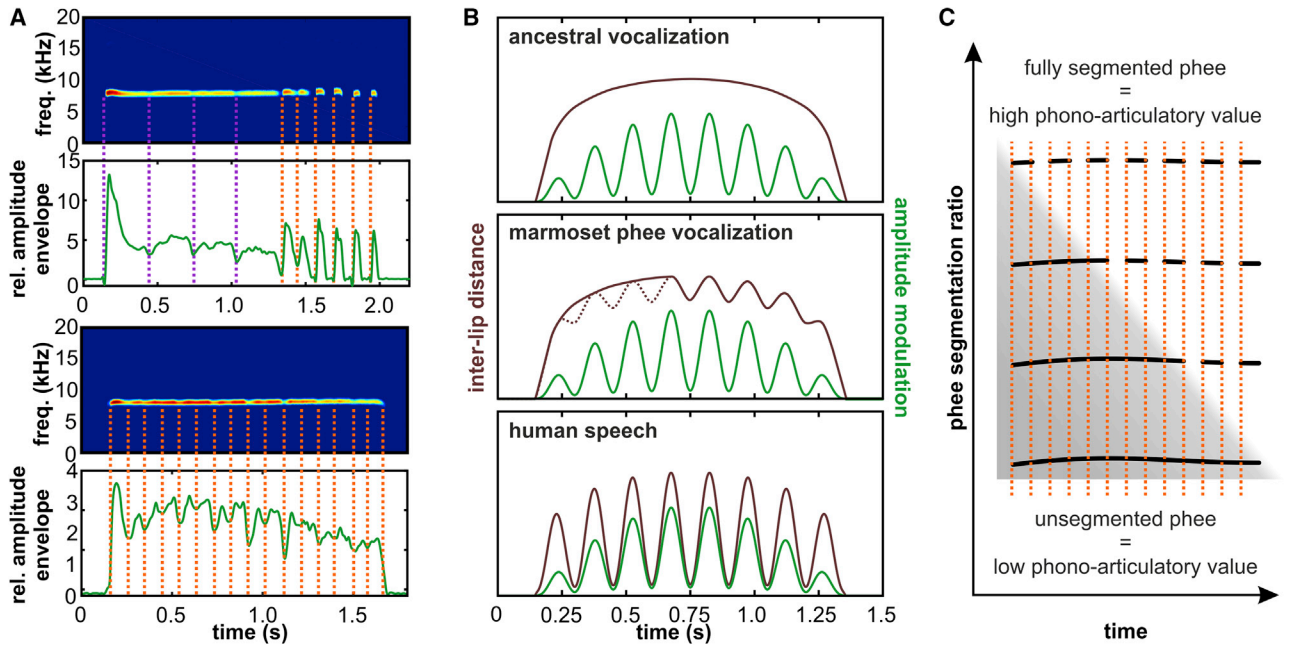


Figure 4. Hypothetical Models for the Evolution of Human Speech from Ancestral Vocalizations

(A) Spectrogram of a segmented (upper plot) and an unsegmented phee (lower plot) with rhythmical changes in call amplitude within the first segment and entire call, respectively. Orange dashed lines indicate rhythmicity at 6–8 Hz, dashed purple lines at 3–5 Hz.

(B) Hypothetical transition from an ancestral phonatory-only (green line) vocal rhythm without a facial articulatory component (brown line; top) to one that is phono-articulatory, with both speech-like mouth movements and acoustics sharing the same coupled rhythmicity (adapted from [9]). Based on our data, marmoset vocalization exhibits an intermediate evolutionary stage.

(C) Hypothetical model depicting the relationship between phee segmentation and the amount of phono-articulatory integration.

envelope and corresponding inter-lip distance throughout the call of monkey P and observed significant coherence values (in accordance with the jackknifing procedure) with peaks at approximately 2–3 and 7–8 Hz (Figure 3F; the low number of video-recorded segmented phees precluded a similar analysis for monkey L). In a second analysis, we omitted the first 200 ms of calls to investigate whether these slow oscillations were caused by slow movements exhibited during mouth opening. Without the start of calls, significant coherence values only occurred between 6–11 Hz with a peak in the speech-related theta range at approximately 7–8 Hz. Overall, these findings indicate that articulatory movements and call amplitude are coupled at frequencies similar to oscillatory rhythms present in human speech [1–3] (Figure 3F).

Our data indicate that articulatory movements and the acoustic structure of phee vocalizations correlate in marmosets (see Videos S1 and S2 for visualization). These findings suggest that certain acoustic features such as segment onsets closely correlate to distinct phases of articulatory movements. Therefore, we investigated how segment onsets related to the phase of inter-lip distances and observed a significantly uneven distribution of phase angles at call onset, with medians of 135.6° for monkey L and 119.9° for monkey P (Figures 3G and 3H; monkey P: p = 4.67e-03, n = 15, z = 5.06; monkey L, p = 6.23e-03, n = 9, z = 4.64; Rayleigh test). We also found a significantly uneven distribution of the onset of the remaining segments with medians occurring slightly earlier in the cycle at 97.7° for monkey L and 73.3° for monkey P (monkey P: p = 1.76e-02, n = 29, z = 3.97, monkey L: p = 1.65e-03, n = 17, z = 5.99; Rayleigh test). Our results support that segmentation onset is directly phase-locked to articulatory movements, with call onset initiated just before maximal opening of the mouth and the remaining segments at an intermediate mouth opening position. Interestingly, rhythmic changes can also be occasionally observed in call amplitude (Figure 4A), as well as in mouth movements (Figure 3B; Videos S1 and S2), during the production of constant long phee segments and unsegmented phees. These findings indicate that such phono-articulatory oscillations are not only present during segmented phee production and suggest an endogenous and fundamental rhythmic pattern underlying vocal production mechanisms in marmosets in general. Future studies will have to verify whether such a proposed endogenous oscillation is also exhibited in other marmoset call types.

Self-generated motor activities possess a temporal nature. Periodicity is widely present among human biological processes such as the heartbeat or breathing and is considered to be a fundamental vertebrate ability [11]. Oscillatory rhythms in the theta range, at 4–8 Hz, can be observed in the sniffing and whisking behavior of rodents [28, 29], during saccadic eye movements in monkeys [30], and during finger movements in humans [31]. Human speech is rhythmically structured in time [6, 32] and involves oscillations in the same frequency range. Furthermore, vocal production rhythms in humans may be linked to perceptual mechanisms, playing a crucial role in communication intelligibility [33], but only within the theta frequency range [11, 34]. Theta rhythmicity underlies the syllabic rate of speech across

all languages [1–3] and is assumed to derive from natural mandibular-associated oscillations [4, 35], which are tightly coupled to the periodic vibrations of vocal folds [4, 6, 7]. This phono-articulatory rhythm, considered to be absent from monkey vocalizations, has been postulated as one of the crucial precursors for the evolution of speech in the primate lineage [3, 8, 13, 36]. The present results challenge this view, suggesting that these oscillations could have already been embedded in the brain of our ancestors earlier in the primate lineage. The central pattern generator underlying orofacial movements and fine respiratory motor control could have existed in a common marmoset-human ancestor, ensuring coordination of the lips and jaw with subglottal pressure and vocal folds, respectively.

To date, speech-like rhythms have been detected in the facial expressions of Old World monkeys during teeth chattering [15] and vocal [16] and non-vocal lip smacking [13]. Furthermore, recent studies in apes also identified theta rhythms in the song phrases of gibbons [12] and voiceless clicks and faux-speech in one orangutan [17]. Moreover, the lip smacks of our closest living relatives, chimpanzees, are produced with similar temporal regulation [14]. Our results from marmosets, in combination with published work on Old World monkeys and apes, suggest a continuous evolution of rhythmic vocal motor acts in the primate lineage. We show that the acoustic call envelope and call-correlated articulatory movements in marmosets, a New World primate species that diverged from the human lineage more than 35 million years ago [37], share core timing and sequencing characteristics with humans, challenging current theories on the evolution of vocal rhythmicity. This implies that the emergence of vocalization-correlated articulatory and phonatory synchronization may have evolved early in the primate lineage and be intrinsic in primate vocal production systems.

The synchronization of phonatory and articulatory systems is a complex natural phenomenon that has presumably been preceded by more elementary modifications [9]. From the ancestral phonatory-only vocal rhythm without concomitant facial motions to human phono-articulatory rhythmicity, the marmoset phee vocalization introduces an intermediate stage in a continuous evolutionary process in the primate lineage (Figure 4B). Phees range from low to high phono-articulatory integration according to the degree of segmentation (Figure 4C). Consequently, we hypothesize that such coupled phono-articulatory oscillations should also be present in other monkey and ape species.

From a neurophysiological perspective, our model suggests the crucial role of theta rhythmicity and synchronization of articulatory and phonatory components in monkey vocal production, from which a more intricate human speech production system might have evolved. This indicates that, independent of whether this rhythmicity is a cortically controlled or brainstem-based motor act, the observed phono-articulatory theta oscillation may be an intrinsic mechanism already present in primate vocal production. Previous data indicate that monkey vocalizations are produced by a vocal-pattern-generating network situated in the lower brainstem that receives input from higher order structures [38–41]. However, the neural mechanisms underlying the observed phono-articulatory theta rhythmicity in marmosets remain unknown. Further studies should elucidate whether the observed call-related rhythmicity and synchronization is driven by intrinsic theta oscillations produced by the vocal-pattern-generating network itself or by cortical oscillations, which are thought to be crucial during speech production [34, 42–45]. Furthermore, comparative studies in other monkey species and apes will contribute to our understanding of how the observed coupled oscillations have evolved to highly variable and controllable speech-related, rhythmical mouth movements present in human speech. The investigation of rhythmic-segmented marmoset calls and the underlying synchronized articulatory and phonatory oscillations will help us decipher the evolution of basic temporal and sequential components of vocal communication signals and, ultimately, understand human speech rhythmicity and its related disorders [26, 35, 46].

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Marmosets
- METHOD DETAILS
  - Experimental setup
  - Data analysis of vocal recordings
  - Population analysis and data normalization
  - Data analysis of video recordings
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.08.019.

ACKNOWLEDGMENTS

We thank John Holmes for proofreading and Vera Voigtländer for fruitful discussion on the manuscript. This work was supported by the Werner Reichardt Centre for Integrative Neuroscience (DFG) at the Eberhard Karls University of Tübingen (CIN is an Excellence Cluster funded by the Deutsche Forschungsgemeinschaft within the framework of the Excellence Initiative EXC 307).

AUTHOR CONTRIBUTIONS

C.R.-S. and S.R.H. conceived the study, designed the experiments, created the visualizations, interpreted the data, and wrote the paper. C.R.-S. conducted the experiments and performed the data analyses. S.R.H. provided the animals, acquired funding, and supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 2, 2020
Revised: July 20, 2020
Accepted: August 5, 2020
Published: September 3, 2020

REFERENCES

1. Bergman, T.J., Beehner, J.C., Painter, M.C., and Gustison, M.L. (2019). The speech-like properties of nonhuman primate vocalizations. Anim. Behav. *151*, 229–237.
2. Coupé, C., Oh, Y., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. Science Advances *5*, https://doi.org/10.1126/sciadv.aaw2594.
3. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. *5*, e1000436.
4. Giraud, A.L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S.J., and Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron *56*, 1127–1134.
5. MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production. Behav. Brain Sci. *21*, 499–511, discussion 511–546.
6. Kotz, S.A., and Schwartze, M. (2010). Cortical speech processing unplugged: a timely subcortico-cortical framework. Trends Cogn. Sci. *14*, 392–399.
7. Ghitza, O., and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. Phonetica *66*, 113–126.
8. Ghazanfar, A.A., and Takahashi, D.Y. (2014). Facial expressions and the evolution of the speech rhythm. J. Cogn. Neurosci. *26*, 1196–1207.
9. Ghazanfar, A.A., and Takahashi, D.Y. (2014). The evolution of speech: vision, rhythm, cooperation. Trends Cogn. Sci. *18*, 543–553.
10. Bass, A.H., Gilland, E.H., and Baker, R. (2008). Evolutionary origins for social vocalization in a vertebrate hindbrain-spinal compartment. Science *321*, 417–421.
11. Kotz, S.A., Ravignani, A., and Fitch, W.T. (2018). The Evolution of Rhythm Processing. Trends Cogn. Sci. *22*, 896–910.
12. Terleph, T.A., Malaivijitnond, S., and Reichard, U.H. (2018). An analysis of white-handed gibbon male song reveals speech-like phrases. Am. J. Phys. Anthropol. *166*, 649–660.
13. Ghazanfar, A.A., Takahashi, D.Y., Mathur, N., and Fitch, W.T. (2012). Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Curr. Biol. *22*, 1176–1182.
14. Pereira, A.S., Kavanagh, E., Hobaiter, C., Slocombe, K.E., and Lameira, A.R. (2020). Chimpanzee lip-smacks confirm primate continuity for speech-rhythm evolution. Biol. Lett. *16*, 20200232.
15. Toyoda, A., Maruhashi, T., Malaivijitnond, S., and Koda, H. (2017). Speech-like orofacial oscillations in stump-tailed macaque (Macaca arctoides) facial and vocal signals. Am. J. Phys. Anthropol. *164*, 435–439.
16. Bergman, T.J. (2013). Speech-like vocalized lip-smacking in geladas. Curr. Biol. *23*, R268–R269.
17. Lameira, A.R., Hardus, M.E., Bartlett, A.M., Shumaker, R.W., Wich, S.A., and Menken, S.B.J. (2015). Speech-like rhythm in a voiced and voiceless orangutan call. PLoS ONE *10*, e116136.
18. Lieberman, P. (1968). Primate vocalizations and human linguistic ability. J. Acoust. Soc. Am. *44*, 1574–1584.
19. Agamaite, J.A., Chang, C.-J., Osmanski, M.S., and Wang, X. (2015). A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). J. Acoust. Soc. Am. *138*, 2906–2928.
20. Pomberger, T., Risueno-Segovia, C., Löschner, J., and Hage, S.R. (2018). Precise Motor Control Enables Rapid Flexibility in Vocal Behavior of Marmoset Monkeys. Curr. Biol. *28*, 788–794.e3.
21. Zhao, L., Roy, S., and Wang, X. (2019). Rapid modulations of the vocal structure in marmoset monkeys. Hear. Res. *384*, 107811.
22. Zürcher, Y., and Burkart, J.M. (2017). Evidence for Dialects in Three Captive Populations of Common Marmosets (Callithrix jacchus). Int. J. Primatol. *38*, 780–793.
23. Castellucci, G.A., Calbick, D., and McCormick, D. (2018). The temporal organization of mouse ultrasonic vocalizations. PLoS ONE *13*, e0199929.
24. Peelle, J.E., and Davis, M.H. (2012). Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. *3*, 320.
25. Giraud, A.L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. *15*, 511–517.
26. Wieland, E.A., McAuley, J.D., Dilley, L.C., and Chang, S.E. (2015). Evidence for a rhythm perception deficit in children who stutter. Brain Lang. *144*, 26–34.
27. Lee, B., and Cho, K.H. (2016). Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. Sci. Rep. *6*, 37647.
28. O'Keefe, J., and Recce, M.L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus *3*, 317–330.
29. Moore, J.D., Deschênes, M., Furuta, T., Huber, D., Smear, M.C., Demers, M., and Kleinfeld, D. (2013). Hierarchy of orofacial rhythms revealed through whisking and breathing. Nature *497*, 205–210.
30. Landau, A.N., and Fries, P. (2012). Attention samples stimuli rhythmically. Curr. Biol. *22*, 1000–1004.
31. Gross, J., Timmermann, L., Kujala, J., Dirks, M., Schmitz, F., Salmelin, R., and Schnitzler, A. (2002). The neural basis of intermittent motor control in humans. Proc. Natl. Acad. Sci. USA *99*, 2299–2302.
32. Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54*, 1001–1010.
33. Fujii, S., and Wan, C.Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. Front. Hum. Neurosci. *8*, 777.
34. Assaneo, M.F., and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. Sci. Adv. *4*, https://doi.org/10.1126/sciadv.aao3842.
35. McNeil, M.R., Pratt, S.R., and Fossett, T.R.D. (2004). The differential diagnosis of apraxia of speech. In Speech motor control in normal and disordered speech, B. Maassen, R.D. Kent, H.F.M. Peters, P. van Lieshout, and W. Hulstijn, eds. (Oxford University Press), pp. 389–413.
36. Fischer, J., and Hage, S.R. (2019). Primate vocalization as a model for human speech: scopes and limits. In Human Language: From Genes and Brains to Behavior, P. Hagoort, ed. (New York: MIT Press), pp. 639–656.
37. Miller, C.T., Freiwald, W.A., Leopold, D.A., Mitchell, J.F., Silva, A.C., and Wang, X. (2016). Marmosets: A Neuroscientific Model of Human Social Behavior. Neuron *90*, 219–233.
38. Hage, S.R., and Nieder, A. (2016). Dual Neural Network Model for the Evolution of Speech and Language. Trends Neurosci. *39*, 813–829.
39. Jürgens, U. (2002). Neural pathways underlying vocal control. Neurosci. Biobehav. Rev. *26*, 235–258.
40. Loh, K.K., Petrides, M., Hopkins, W.D., Procyk, E., and Amiez, C. (2017). Cognitive control of vocalizations in the primate ventrolateral-dorsomedial frontal (VLF-DMF) brain network. Neurosci. Biobehav. Rev. *82*, 32–44.
41. Simonyan, K., Ackermann, H., Chang, E.F., and Greenlee, J.D. (2016). New Developments in Understanding the Complexity of Human Speech Production. J. Neurosci. *36*, 11440–11448.
42. Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. Neuron *69*, 407–422.
43. Wilson, S.M., Saygin, A.P., Sereno, M.I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nat. Neurosci. *7*, 701–702.
44. Tremblay, S., Shiller, D.M., and Ostry, D.J. (2003). Somatosensory basis of speech production. Nature *423*, 866–869.

45. Poeppel, D., and Assaneo, M.F. (2020). Speech rhythms and their neural foundations. Nat. Rev. Neurosci. *21*, 322–334.

46. Ludlow, L., Connor, N.P., and Bassich, C.J. (1987). Speech Timing in Parkinson's and Huntington's Disease. Brain Lang. *214*, 195–214.

47. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (Hillsdale: Lawrence Erlbaum Associates).

48. Berens, P. (2009). CircStat: A Matlab Toolbox for Circular Statistics. J. Stat. Softw. *31*, 10.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Callithrix jacchus	German Primate Center, Göttingen, and Werner Reichardt Center for Integrative Neuroscience, University of Tübingen	N/A
Software and Algorithms		
MATLAB	MathWorks	R2019b
OpenEx	Tucker-Davis Technologies	N/A
SASLab Pro	Avisoft Bioacoustics	version 5.2.13

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Steffen R. Hage (steffen.hage@uni-tuebingen.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All data needed to evaluate the conclusions are present in the paper. The raw datasets and code supporting the current study have not been deposited in a public repository because of further analyses, but are available from the Lead Contact, Steffen R. Hage (steffen.hage@uni-tuebingen.de), upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Marmosets

We recorded 3449 calls produced by three adult common marmosets, *Callithrix jacchus*, two females (monkey C and P) and one male (monkey L), housed at the University of Tübingen. Animals were born in captivity and kept in mixed sex pairs. The facility room was maintained at approximately 26°C, 40%–60% relative humidity, and 12-h:12-h light/dark cycle. They had *ad libitum* access to water and were fed daily with standard commercial chow, fresh fruit and vegetables, mealworms, and locusts. Marshmallows and special fruit (e.g., bananas, grapes) were used to transfer the animals from their home cages to a transfer box. Experimental procedures were approved by the local authorities of Tübingen (*Regierungspräsidium*) and are in agreement with the guidelines of the European Community for the care of laboratory animals.

METHOD DETAILS

Experimental setup

Vocal recordings setup

For all three monkeys, vocal recordings were conducted in a soundproof chamber. The vocal behavior was reinforced for every uttered vocalization by administering a liquid reward (consisting of water, fruit, marmoset gum, marshmallow, and curd cheese) via a metal syringe. Monkeys L and P were trained while sitting in a primate chair and monkey C in an experimental cage (25 × 25 × 28cm). With this approach the animals produced a high number of phee calls independently of the used training procedure. Vocalizations were recorded using a microphone (MKH 8020 microphone with MZX 8000 preamplifier, Sennheiser, Germany) located at a distance of 10 cm from the monkey's head, in a soundproof chamber to avoid sound reflection. Monkey L was usually trained between 10 am and 12 pm and monkeys P and C between 11 am and 1 pm.

Vocal detection and reward presentation were synchronized and performed automatically with a custom-written program (OpenEx, Tucker-Davis Technologies, U.S.A.) running on a workstation (WS-8 in combination with an RZ5 bioamp processor and RZ6D multi I/O processor, Tucker-Davis Technologies, U.S.A.). Vocalizations were recorded using the same system with a sampling

rate of 100kHz. The behavior of the monkeys was constantly monitored using a video camera (Brio, Logitech). To ensure precise timings, vocal onset and offset times were detected offline (Avisoft-SASLab Pro 5.2.13, Avisoft Bioacoustics).

Video recording setup

We recorded video using a 4K video camera (HC-VX989, Panasonic Corporation, Japan; sampling frequency, 30fps) of the two animals that were trained to vocalize under reinforced conditions in the primate chair (four sessions each) in order to analyze the vocalization-correlated orofacial kinetics of 24 segmented phee calls (see below). Accurate tracking of the lips, i.e., the inter-lip distance, during vocal output was ensured by restricting the analysis only to vocalizations during which the monkeys kept their heads relatively still and directly facing the camera.

Data analysis of vocal recordings

In the present study, we aimed to investigate the intrinsic oscillatory mechanisms of acoustic call structure, e.g., amplitude envelope, and call-associated articulatory movements, i.e., inter-lip distance during call production. Therefore, we focused on single phees (segmented and unsegmented) and the first phee syllable of multi-syllabic phees. Long vocal utterances like phee calls (> 1 s) are required in order to calculate power spectra or coherence of the phono-articulatory coupling with high frequency resolution within theta range (3-8Hz). Shorter marmoset call types such as twitter, chirp, tsik and ekk calls (< 100ms) [19, 20] might display some limitations in that sense.

Since segmented phee calls with five or more segments were rare, we grouped these calls together to analyze call duration and frequency bandwidth distributions as a function of the number of segments per call. Segmented phee call duration was calculated as the difference between the onset of the first segment and offset of the last segment within a single vocalization. Segment duration was defined as the time between segment onset and offset, and ISI as the time between segment offset and consecutive segment onset. Segment interval was defined as the time between segment onset and the next segment onset. The frequency bandwidth was calculated as the difference between the highest and lowest peak frequency at each time point throughout a phee. A smoothing factor of 30 was applied to the trajectories of the mean frequency and amplitude of the normal and segmented phees. The amplitude envelope, more precisely the upper amplitude envelope, was extracted from the sum of the acoustic power across all frequencies at given times during the call. For a more objective and standardized analysis, phee units were defined as the segment duration within one standard deviation of the mean of a Gaussian curve fitted to the segment duration distribution for each monkey. We used the Gaussian to confine the analysis to the first component of the multimodal distributions, since we aimed to evaluate proper phee units by excluding longer segments that might be a combination of more than one unit. This approach has been used in a similar analysis of vocal behavior before [23]. For visualization purposes, a moving average was used to depict segment duration distributions, with a window size of 20ms in steps of 1 s and a smoothing factor of 30. Unit duration was the time between unit onset and offset, IUI the time between unit offset and subsequent unit onset, unit interval or cycle duration the time between unit onset and the next unit onset, and oscillatory rate the inverse of the cycle duration. The violin plots shown in Figure 2B were calculated using the MATLAB function, violin.m, using kernel density estimation with a kernel bandwidth of 4 for unit duration and IUI and a kernel bandwidth of 0.3 for oscillatory rate.

Population analysis and data normalization

Call duration was normalized by dividing call durations by the mean duration of unsegmented phees for each individual monkey. Call frequency and amplitude of normal and segmented phees were normalized by dividing their respective means by the mean values for normal phees at call onset for each animal.

Data analysis of video recordings

To determine the mouth opening or inter-lip distance, the Euclidean distance was computed between pairs of upper and lower lip's coordinates that were manually flagged offline, frame by frame, using Adobe Premiere Pro CS5 software. We performed a linear interpolation (factor 5) of the jaw movement trajectory to increase the sampling frequency to 150Hz and normalized the signal to the maximum opening of the mouth per call. The local phase angle of the inter-lip distance at segment onset was estimated over single cycles from the discrete-time analytic signal using the Hilbert transform with the MATLAB function, ksdensity.m, with a kernel bandwidth of 0.2. The mean coherence of the call amplitude envelope and corresponding inter-lip distance was calculated using the magnitude-squared coherence estimate (using Welch's overlapped averaged periodogram method). This coherence estimate is a function of frequency with values between 0 and 1 indicating how well the amplitude envelope of the acoustic signal corresponds to the inter-lip distance at each frequency. For example, a coherence function equal to 1, or maximum coherence, at a specific frequency would mean that the phases of these two signals are perfectly synchronized. For the coherence and correlation analyses of the relative call amplitude envelope and relative inter-lip distance we down-sampled the auditory signal via linear interpolation to match the 150Hz of the jaw movement signal.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using MATLAB (MathWorks, Natick). Two-way ANOVAs and post hoc multiple comparison t tests were performed to compare distributions of segment durations according to the segment position within the call. Kruskal-Wallis tests were used to compare distributions of unit durations, IUIs, and oscillatory rates between animals. Pearson's correlations were

used to identify the relationship between ISI, total call durations, and frequency bandwidths with respect to the number of call segments. Pearson's correlations were also used to study the relationship between unit duration and the duration of the subsequent IUI, as well as between the relative amplitude envelope power and relative inter-lip distance. To correct for high sample sizes in the latter test, we introduced Cohen's effect size [47]. To obtain ranges in the power spectra of the call amplitude envelope and the jaw movements exhibiting significant gains, we fit a linear regression model to the logarithm of the power and the logarithm of frequency for $f^{-\alpha}$ trends [3]. For visualization purposes the signal was detrended afterward [17]. Significant deviations from this fit are indicated by the black dashed line. A 95% jackknife confidence interval was calculated from the mean coherence signal [17] to test for frequency ranges with significant coherence values (indicated by the black dashed lines). The 95% confidence interval from the mean coherence over the 10–15kHz range was used as the baseline coherence. The Rayleigh test for non-uniformity of circular data [48] was computed for the polar histograms of the articulatory jaw phase at segment onset. In all performed tests, significance was tested at an alpha = 0.05 level.