A key question for understanding speech evolution is whether or not the vocalizations of our closest living relatives—nonhuman primates—represent the precursors to speech. Some believe that primate vocalizations are not volitional but are instead inextricably linked to internal states like arousal and thus bear little resemblance to human speech. Others disagree and believe that since many primates can use their vocalizations strategically, this demonstrates a degree of voluntary vocal control. In the current study, we present a behavioral paradigm that reliably elicits different types of affiliative vocalizations from marmoset monkeys while measuring their heart rate fluctuations using noninvasive electromyography. By modulating both the physical distance between marmosets and the sensory information available to them, we find that arousal levels are linked, but not inextricably, to vocal production. Different arousal levels are, generally, associated with changes in vocal acoustics and the drive to produce different call types. However, in contexts where marmosets are interacting, the production of these different call types is also affected by extrinsic factors such as the timing of a conspecific’s vocalization. These findings suggest that variability in vocal output as a function of context might reflect trade-offs between the drive to perpetuate vocal contact and conserving energy.

Understanding how human speech evolved is an enormously difficult problem (1, 2). As evolution typically tinkers with preexisting phenotypes to find workable solutions to new challenges (3), it seems logical that human speech evolved from the nonspeech vocalizations of our hominid ancestors. Since social behaviors, the brain, and other vocal production-related soft tissues do not fossilize, we are left with the comparative method—investigating the vocal behavior and associated mechanisms of extant animals—to shed light on how speech may have evolved via nonspeech vocal output. In light of this, a key question for understanding speech evolution is whether or not the vocalizations of our closest living relatives—nonhuman primates (hereafter, “primates”)—represent the precursors to speech (4, 5).

Some believe that primate vocalizations are inextricably linked to internal states like arousal and thus cannot represent precursors to human speech as such vocalizations are not volitional (6–8). Others believe that this is not the case, that many primates can use their vocalizations strategically, demonstrating a degree of volitional vocal control (9–11). For example, the presence of different predators triggers the production of different alarm calls by vervet monkeys, suggesting that external cues determine which vocalizations are produced (12). However, similar-sounding vocalizations to vervet alarm calls may be elicited by events unrelated to the presence of predators (e.g., aggression), suggesting that vocal production is instead linked to the arousal state of the animal (13). Along similar lines, affiliative “grunts” produced by female macaques and baboons can be used to signal “benign intent” (i.e., volitionally) (14, 15). For example, high-ranking baboons grunt toward lower-ranking ones to signal that no aggression is imminent (15). However, in other contexts, the acoustic structure of baboon grunts is reported to differ as a function of presumed high or low arousal states (16). It is impossible to really understand how internal states like arousal relate to the production of primate vocalizations without a direct measure of such physiological states in controlled contexts (5).

Under field conditions, it is difficult to obtain physiological measures of arousal levels (e.g., heart rate, pupil dilation, skin conductance) that may accompany vocal production. It is also challenging in such conditions to account for the many variables that could be affecting an individual’s arousal level at any moment in time. By contrast, in captivity where direct physiological measures are easier to acquire, it is difficult to elicit different types of vocalizations. This is because the number of contextual cues that trigger vocalizations is small in captive conditions relative to the wild (10, 17). In the current study, we combine a behavioral paradigm to reliably elicit different types of affiliative vocalizations from captive marmoset monkeys with simultaneous measurements of their heart rate fluctuations using noninvasive electromyography (EMG). We find that arousal levels are linked, but not inextricably, to vocal production and that different arousal levels interact with external cues in the production of different call types.

In this study, we systematically manipulated social contexts while investigating the relationship between vocal output and arousal levels in pair-bonded marmoset monkeys (Callithrix jacchus; n = 6 marmosets, or three pairs of male–female cage mates). Inspired by the finding that wild pygmy marmosets will switch among different affiliative call types according to their physical distance from their social group (18), we designed an experiment in which marmosets were separated in different ways from their social groups.

Results

In this study, we systematically manipulated social contexts while investigating the relationship between vocal output and arousal levels in pair-bonded marmoset monkeys (Callithrix jacchus; n = 6 marmosets, or three pairs of male–female cage mates). Inspired by the finding that wild pygmy marmosets will switch among different affiliative call types according to their physical distance from their social group (18), we designed an experiment in which marmosets were separated in different ways from their social groups.

**Significance**

Understanding how human speech evolved requires investigating how our closest extant relatives—nonhuman primates—use and produce their vocalizations. However, some believe that primate vocalizations are driven purely by internal states such as arousal and thus have little or no relationship with human speech. We show that, in marmoset monkeys, vocal production varies systematically as a function of social distance and—in via measures of heart rate—with arousal levels. However, arousal is not inextricably linked with primate vocal production. Rather, it interacts in complex ways with external cues, revealing the interplay between communication and energetic demands.

Author contributions: D.A.L., L.X.C., and A.A.G. designed research; D.A.L., Y.S.Z., and A.A.G. performed research; D.A.L. analyzed data; and D.A.L. and A.A.G. wrote the paper. The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Published under the PNAS license.

*To whom correspondence should be addressed. Email: asig@princeton.edu.
This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1722426115/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1722426115

*PNAS* Latest Articles | 1 of 6
Changes in context affect acoustic properties of vocal behavior. (i) A single marmoset was alone in one corner of an experiment room (Alone); (ii) a pair of marmosets were placed in opposing corners with an acoustically transparent opaque curtain blocking visual access (Occluded Far [OccFar]); (iii) a pair of marmosets were placed in opposing corners without an opaque curtain, allowing visual contact [Visible Far (VisFar)]; and (iv) a pair of marmosets were placed in the same corner with visual, but not physical, contact [Visible Close (VisClose)] (Fig. 1A). Here, “social distance” refers to physical distance from a partner, and whether the partner could be seen or not, in accord with these four experimental conditions.

We first measured how social distance affected vocal output. The percentage of time spent vocalizing by marmoset monkeys decreased as a function of decreasing social distance (Fig. 1B) \( (P = 2.15e-12) \). Even though individual marmosets differ in how voluble they are, this relationship is generally sustained (Fig. S1). The acoustic structure of vocalizations also changed systematically. On a per-session basis, we measured duration, dominant frequency, amplitude, and Weiner entropy (how noisy the calls were) (Fig. 1C and D). To avoid categorization biases (19, 20), we did not prior label calls according to ethology (21). Using a fixed-effects linear regression model, we found that, with decreasing social distance, marmoset vocalizations became shorter in duration \( (P = 2.60e-55) \), lower in dominant frequency \( (P = 7.80e-05) \), quieter \( (P = 4.66e-47) \), and noisier \( (P = 1.10e-51) \).

By subsequently classifying the marmoset vocalizations into call types (21), we found that three affiliative vocalizations made up 89.4% of all calls produced: phees, trillphees, and trills (Fig. 2A). We focused our subsequent analyses on these three vocalizations. Were these different call types uniquely associated with each context? We examined how frequently these different call types were produced across the different contexts (Fig. 2B). When alone, marmoset only make phee calls, but with decreasing social distance, the proportion of phees produced decreases. Conversely, trillphees and trills are completely absent when marmosets were alone, but production of both call types increased with decreasing social distance until trills were the majority call type produced in the VisClose condition. A multiple-regression model found a significant effect of call type on context \( (P = 4.25e-111) \) and a significant interaction between call type and context \( (P = 1.00e-11) \). For dominant frequency, we found that it decreased in phee calls but increased in trillphees and trills with decreasing social distance \( (P = 0.0009) \); call type, \( P = 1.29e-49 \); call type, \( 3.76e-55 \), lower in dominant frequency \( 4.66e-47 \), and noisier \( 1.82e-83 \). Furthermore, all the marmosets showed a similar pattern of transitioning between call types (Fig. S2). These results show that different proportions of affiliative call types are produced as a function of social distance.

We next examined whether the change in acoustic features was solely due to the change between call types across contexts or whether vocal flexibility is also present within call types. Even though the percent time spent vocalizing decreased linearly with decreasing social distance, the call rate by context did not follow the same trend (Fig. S3). We looked at the acoustic features within each call type for each context. For instance, Fig. 2C shows that the phee calls from a single marmoset exhibit duration and amplitude changes as a function of context. We thus calculated, for all marmosets, changes in duration, dominant frequency, amplitude, and entropy for the phee, trillphee, and trill calls (Fig. 2D). For duration, we found significant changes in both call type and context: vocalizations were shorter as social distance decreased (context, \( P = 1.29e-49 \); call type, \( P = 3.76e-76 \)). For dominant frequency, we found that it decreased in phee calls but increased in trillphees and trills with decreasing social distance (context, \( P = 0.0009 \); call type, \( P = 1.00e-11 \)).

---

**Fig. 1.** Changes in context affect acoustic properties of vocal behavior. (A) Schematic of room and marmoset placement in four different social contexts. In order of decreasing distance: Alone, a marmoset was alone in one corner; Occluded Far, two marmosets were at opposing corners with an opaque curtain between them; Visible Far, two marmosets were at opposing corners without a curtain; Visible Close, two marmosets were in the same corner. (B) Percent time spent calling in the session by context in order of decreasing social distance. (C) Four acoustic features were measured in the study: duration, dominant frequency, amplitude, and Weiner entropy. Vertical red lines indicate the boundaries of the identified vocalization. (D) Change in those four acoustic features over contexts. Error bars capture the SEM.
amplitude, there was a decrease in intensity with decreasing social distance (context, \( P = 8.87 \times 10^{-20} \); call type, \( P = 0.2895 \)). For entropy, each call type increased in noisiness the closer the marmosets were to each other (context, \( P = 9.99 \times 10^{-38} \); call type, \( P = 1.39 \times 10^{-07} \)). Together, these results indicate that there is a significant amount of flexibility within particular call types that is revealed when the same call type is made in different social contexts.

How are arousal levels associated with changes in overall vocal output, its acoustic features, and ultimately the switch to different call types? Heart rates were measured by recording non-invasive EMG signals using surface electrodes placed over the chest and back (\( n = 3 \) of the six marmosets). We used changes in heart rate because it is a temporally precise index of arousal state (22–24). In Fig. 3A, exemplars of the audio recording and raw physiological signal, along with the calculated heart rate, are shown. The amount of time spent calling was positively correlated with mean heart rates of the session (Fig. 3B; partial correlation with subject identity taken into account: \( \rho = 0.3928, P = 0.0086 \)). Moreover, a paired \( t \)-test showed that the heart rate 1 s before call onset was generally higher than the mean heart rate for the session (Fig. 3C; \( P = 4.25 \times 10^{-04} \)). By correlating heart rates before call onset with our four different acoustic measures, we found that as heart rates increased, duration and amplitude of the subsequent call increased (duration: \( \rho = 0.4476, P = 7.78 \times 10^{-04} \); amplitude: \( \rho = 0.3501, P = 0.0102 \)). There was a negative correlation between heart rates and entropy (\( \rho = -0.3624, P = 0.0077 \)), and no significant correlation between heart rates and dominant frequency (\( \rho = 0.1272, P = 0.3639 \)).

When vocal output and heart rate measures were collapsed across sessions, the different affiliative call types were associated with different levels of arousal just before their production (\( P = 0.023 \)) (Fig. 3E). The phee call was produced when the heart rate was highest, and the trill call was produced when heart rate was lowest. These relationships between call types and heart rate levels were not surprising given that heart rates decrease as a function of decreasing social distance (Fig. 3B) and that the proportion of different call types produced also changes as a function of social distance (Fig. 2B). However, within a particular context, there appeared to be no consistent relationship between the heart rate and the call type produced (Fig. 3F; OccFar, \( P = 0.388 \); VisFar, \( P = 0.305 \); VisClose, \( P = 0.961 \)). One possibility is that vocal production is not linked to the absolute level but to the relative level of arousal. Given that heart rate is elevated before call onset relative to the session mean (Fig. 3C), we examined whether there were differences in the change in heart rate for the different call types (Fig. S4). We found that, for all sessions, there was a significant effect of call type on the heart rate change (\( P = 0.021 \)). However, when we separated out the call types into their respective contexts, our regression model found no significant effect of call type (OccFar, \( P = 0.054 \); VisFar, \( P = 0.104 \); VisClose, \( P = 0.467 \)). Overall, these data suggest that vocal production (e.g., the type of call produced) is not intrinsically linked to levels of arousal.

To investigate this further, we measured the relationship between a 0.1-Hz autonomic nervous system rhythm known as the Mayer wave and the timing of vocal production. From our prior work, we know that this 0.1-Hz rhythm (measured via heart rate fluctuations; ref. 25) drives the timing of phee call production when the marmoset is alone: marmosets tend to produce phee calls roughly every 10 s (23). We hypothesized that decreasing social distance (i.e., increasing available external/social cues) would perturb this intrinsic arousal rhythm and consequently weaken its relationship with vocal output. We calculated the coherence between the call output and the heart rate for each session (Fig. 4A). As before (23), we found a strong peak around 0.1 Hz when marmosets were in the Alone condition. As predicted, the coherence decreased to nonsignificance in the two visible conditions. This tells us that the influence of this internal arousal rhythm decreases in the presence of extrinsic factors.

Among those extrinsic factors is, of course, what the other conspecific may be doing. We thus investigated how changes in social distance affected the timing of vocalizations produced by an individual and across pairs. A histogram of call latencies for each context is shown (Fig. 4B). We defined the call latency as the interval between the end of a previous vocalization to the start of a new one (regardless of who produced it). In the Alone condition, a peak emerged at around 10 s, consistent with the coherence analyses (Fig. 4A). However, in contexts where the marmoset was interacting with another conspecific, there was a bimodal distribution of call latencies: a sharp, fast response and a slower peak at around 10 s. We separated out those rapidly produced vocalizations made within 12 s that were in response to another marmoset’s vocalizations. Much like their relationship with increasing arousal levels (Fig. 3D), we found positive correlations between duration, frequency, and amplitude with increasing response latency (Fig. 4C); the slower the response, the longer, higher-pitched, and louder the vocalization (duration: \( \rho = 0.117, P = 5.61 \times 10^{-08} \); frequency: \( \rho = 0.241, P = 1.99 \times 10^{-29} \); amplitude: \( \rho = 0.132, P = 1.14 \times 10^{-09} \)). Similarly, there was a negative correlation with entropy: the faster the response, the noisier the call (\( \rho = -0.061, P = 0.005 \)) (Fig. 4C).

We then calculated the proportion of different call types produced as a function of response latency (Fig. 4D). The top panels belong to an exemplar marmoset and the bottom panels are the group mean [OccFar: \( n = 4 \) (two marmosets only produced phee calls in this context and were excluded); VisFar and
Fig. 3. Arousal states influence subsequent vocal production. (A) Depiction of electromyography (EMG) setup. Surface electrodes were applied to the dorsal and ventral thorax of an adult marmoset. To the right, signal exemplars were aligned to call onset—waveform for a phee call, raw EMG signal capturing cardiorespiratory activity, isolated heart rate signal. The red tick indicates call onset. (B) Mean heart rate for the entire session correlated with percent time spent calling in the session. Colored dots indicate context. All heart rates are in units of hertz (beats per second). (C) Heart rate before call onset was generally increased over the heart rate of the session mean. Gray triangle represents area where heart rate before vocalization was the equal to or less than the session mean. (D) Correlation between heart rate before call onset and acoustic features of the subsequent call (duration, frequency, amplitude, entropy). Linear trend lines are shown. Trend lines in red show significant correlations. (E) Heart rate plotted by call type collapsed across context. Scatter points are individual sessions. (F) Heart rate differences by call type are separated into their respective contexts. Error bars capture the SEM.

VisClose: \( n = 6 \). We found that when responses were quick—less than a few seconds—the call types produced were more likely to be trills or trillphees. When more time passed, the call type was more likely to be a phee call. In two contexts, the probability to produce certain call types significantly depended on the latency (fixed-effect multiple regression model; OccFar: call type by latency, \( P = 0.017 \); VisFar: call type by latency, \( P = 0.0016 \); VisClose: call type by latency, \( P = 0.566 \)).

The phee call is a louder, longer, and more tonal vocalization than the other call types and requires greater coordination of vocal biomechanics (more respiratory power and greater laryngeal tension) (24, 26). We also know from developmental studies that the timing of spontaneous vocal output relative to arousal levels determines which call types are produced, with phee calls being produced when arousal levels are highest (24). We thus hypothesized that a marmoset’s level of arousal when it hears a conspecific’s vocalization may influence the timing of its vocal response. To test this hypothesis, we calculated the heart rate before call perception (1 s before hearing a call) and the heart rate before call production (1 s before calling back) (Fig. 4E). We found that response latency is positively correlated with heart rate before call perception (\( r = 0.138, P = 0.001 \)), and even more strongly correlated with heart rate before call production (\( r = 0.2885, P = 3.08 \times 10^{-12} \)). These results show that the arousal state is lower when the marmoset responds faster upon hearing another’s call.

**Discussion**

Understanding the interplay between internal states and extrinsic factors in the production of primate vocalizations is necessary for investigating the origins of human communication. One of the difficulties in determining what role internal states may play in vocal production is that such states contain a number of related parameters including a motivational component (i.e., does the behavior fulfill a certain need), an affective component (e.g., positive or negative valence), and the arousal component related to the probability and latency to respond (5, 27). It is also worth noting that external and internal milieus may not be easily separated. External stimuli might elicit adaptive changes in the internal state that prepare the animal for appropriate action, including vocal output. For example, changes in social status and group structure as a result of aggression or predation (i.e., extrinsic factors) could lead to increased levels of stress hormones (i.e., an internal state change) (28–30). Increased levels of stress hormones can subsequently increase the probability of producing vocalizations (31). Another major problem with invoking a role for internal states in producing variation in vocal output is that it is impossible to falsify hypotheses without an actual physiological measure. In our experiment, we controlled for the first two components of internal states by creating a consistent motivation across experimental contexts (reunion with partner) and only positively valenced interactions. We eliminated the other obstacle to our understanding by directly measuring heart rate changes as function of social distance using noninvasive EMG.

We found that decreases in social distance reliably elicited changes in the acoustic parameters of vocalizations that, ultimately, translated into producing different proportions of three affiliative call types. Measures of heart rate as a function of social distance revealed that, in general, arousal levels also decreased with decreasing social distance. Thus, on average, arousal levels...
were correlated with changes in different acoustic features and the production of different vocalizations. These findings are consistent (except for noisiness) with many vocalization studies of primates that infer arousal level changes based on an animal’s behavior as a function of the presence and/or distance from a conspecific or predator. Vocalizations get longer, get louder, and are produced at a higher rate with increasing arousal (32–37). However, at social distances that could elicit all three call types (phee, trillphees, and trills), their production was not reliably associated with different arousal levels. This suggested that extrinsic factors also play a role in driving the production of vocalizations.

Consistent with this idea, an autonomic nervous system rhythm linked to the timing of spontaneous phee call production in marmosets (23) decreased in power with decreasing social distance. Conversely, extrinsic factors such as the vocalizations of partners became increasingly important in influencing the marmoset’s vocal output. As the role of arousal is essentially a means of allocating metabolic energy [i.e., preparing the body for action (38)], variability in vocal output as function of context—from changes in acoustic features to the production of different call types—might be the result of the interplay between energy usage and the facilitation of social coordination. One advantage of vocal communication is the ability to broadcast signals over long distances, and reciprocal exchanges of calls allows for maintenance of social bonds in visually challenging habitats (39). However, vocal production, particularly the production of loud, long, and tonal contact calls, is energetically demanding, eliciting high metabolic costs (40). Visual contact could supplement vocal fidelity, allowing energy savings through the production of quieter and shorter vocalizations. For marmoset monkeys, switching contexts rebalances the constraints of broadcasting social information (e.g., making longer, louder, lower entropy calls) and the energy the marmoset invests in making these vocalizations at the expense of other behaviors (e.g., foraging, infant care, etc.).

Primates must learn to produce the correct vocalizations in the correct contexts. For example, infant vervets produce “eagle” alarm calls to a very broad class of visual stimuli found in the sky above (both harmful and harmless bird species, falling leaves, etc.). Over time, they refine their alarm calls to a very limited set of genuinely dangerous raptor species (12, 41). Similarly, infant marmoset monkeys initially produce some affiliative call types in the “incorrect” alone context (19, 24), and only later in postnatal life produce them in the correct contexts via experience (42). In both species, arousal levels also influence vocal output (13, 24), and thus it seems external context cues must be correctly linked with different levels of arousal. This type of associative learning—between internal states and external cues—could be mediated by communication between the two parallel neural pathways involved vocal production. In primates (including humans), vocalizations are produced via the coordination of brain regions within a “medial” pathway involved in both arousal regulation and the motor aspects of vocal production and a “lateral” pathway that is involved in volitional control of vocal production through learning (11, 43). Recent neurophysiological data are consistent with these ideas [macaque monkeys (44, 45); marmosets (46–48)].

Although it is widely believed that primate vocalizations cannot represent precursors to human speech because their production is linked to arousal (or some other internal state) (6–8), this belief is inconsistent with what is known about speech production. Like primate vocalizations, speech production is linked
Individual marmosets were tested in four different conditions: (i) Alone (A), a single marmoset was placed in a corner of a testing room; (ii) Occluded Far (OF), a pair of marmosets were placed in opposing corners with an opaque curtain across the middle blocking visual access; (iii) Visible Far (VF), a pair of marmosets were placed in opposing corners without an opaque curtain; and (iv) Visible Close (VC), a pair of marmosets were placed in the same corner. The participants were three cage mate pairs (n = 6 marmosets, 3 males/3 females).

For each, one marmoset from each pair (n = 3 marmosets: 2 males, 1 female) was selected to record EMG signals from. We used two pairs of Ag–AgCl surface electrodes secured around the marmoset’s thorax. The electrode signals were collected by a Plexon Multichannel Acquisition Processor (MAP) data acquisition system and digitized at 1,000 Hz.

Methods are described in detail in SI Materials and Methods.

ACKNOWLEDGMENTS. We thank Weiji Ma for a set of queries that prompted this study. We thank Daniel Takahashi and Lauren Kelly for helpful discussions. This work was supported by NIH Grant R01NS054898 (to A.A.G.) and a National Science Foundation Graduate Research Fellowship (to D.A.L.).
Supporting Information
Liao et al. 10.1073/pnas.1722426115

SI Materials and Methods

Subjects. Marmoset monkeys live with pair-bonded mates in family groups and were born in captivity. They had ad libitum access to water and were fed daily with commercial marmoset diet, supplemented with fresh fruits, vegetables, and insects. The colony room was maintained at a temperature of ∼27 °C with 50–60% relative humidity and a 12:12 light/dark cycle. All experimental sessions were conducted between 0100 hours and 1800 hours. All experiments were performed with the approval of the Princeton University Institutional Animal Care and Use Committee.

Experimental Setup. Additional treats were used before each session to transfer the marmosets from their home cage into a transfer box. Marmosets were then situated in corners of a sound-attenuated experimental room (3.25 × 2.90 m) in four different configurations portraying different social contexts (Fig. 1A). These contexts were as follows: (i) Alone (A), a single marmoset was placed in a corner of the room; (ii) Occluded Far (OF), a pair of marmosets were placed in opposing corners (4.36-m separation) with an opaque curtain across the middle blocking visual access; (iii) Visible Far (VF), a pair of marmosets were placed in opposing corners (4.36-m separation) without an opaque curtain; and (iv) Visible Close (VC), a pair of marmosets were placed in the same corner (0.3-m separation) without visual occlusion. One microphone (AKG C1000S) was positioned ∼0.3 m above each marmoset, placed in Plexiglas and wire testing boxes, and vocalizations were recorded using Audacity at 96,000 Hz. The participants were three cage mate pairs (n = 6 marmosets, 3 males/3 females). Each marmoset was run in each context for a minimum of eight 30-min sessions, a maximum of 12 sessions. The order in which contexts were run was randomized for each marmoset. Sessions where no calls were made were discarded and not analyzed further. A total of 140 sessions and 9,912 calls were recorded.

Noninvasive Electromyography. In 54 sessions (Alonge: n = 14; OccFar: n = 14; VisFar: n = 13; VisClose: n = 13), one marmoset from each pair (n = 3 marmosets: 2 males, 1 female) was selected to record electromyography (EMG) signals from. Sessions with no vocalization and/or low signal-to-noise ratios were excluded from this set of data. To record the EMG signal, we used two pairs of Ag–AgCl surface electrodes (Grass Technology). The electrodes were sewn onto a soft elastic band that was secured around the marmoset’s thorax. One pair of electrodes was placed on the chest close to the heart and the second pair was placed on the back, close to the diaphragm. On a separate day before experimental sessions, we shaved the chest and back of the marmosets to improve the signal-to-noise ratio. Additionally, we applied ECI gel on the surface of each electrode before wrapping it around the marmoset. The electroderm signals were collected by a Plexon Multichannel Acquisition Processor (MAP) data acquisition system and digitized at 1,000 Hz. A third microphone was arranged close to the diaphragm. On a separate day before experimental sessions, we shaved the chest and back of the marmoset and the signal sent to a separate channel of the MAP system to help with call alignment in subsequent analyses. Recorded sessions were 30 min with sessions (n = 4; mean, 24.25 min) aborted early if the signal diminished.

Quantification and Statistical Analysis. Vocalization processing. A custom MATLAB routine automatically detected the onset and offset of any acoustic signal that differed from the background noise. Four acoustic parameters, similar to those used previously, were calculated from automatically detected calls (1). We first bandpass filtered the audio signal between 5 and 11 kHz to detect calls. The duration of the call is the difference between the offset and onset of the detected vocalization. Individual syllables are combined into a call if the interval between them is <500 ms. A spectrogram was calculated using a fast Fourier transform window of 1,024 points, Hanning window, with 50% overlap. The dominant frequency of a syllable was calculated as the average frequency at which the spectrogram had maximum power. The amplitude is calculated as the power of the sound at the identified dominant frequency. Calls with higher values have more energy throughout their dominant frequencies. The power (p) is converted using the equation: $L_p = 20 \times \log_{10}(p/p_0)$, where $p_0 = 20e-6$ Pa. We then converted to the sound power level using the equation: $L_{w} = L_p + [10 \times \log(Q/4\pi r^2)]$, where $Q$ is the directivity factor and $r$ is the distance to the sound source (e.g., marmoset is 0.3 m away from the microphone). The Wiener entropy, representing how broadband the power spectrum of a signal is, is the logarithm of the ratio between the geometric and arithmetic means of the values of the power spectrum for different frequencies. Additionally, the time spent calling per session was calculated by summing the duration of all of the calls in the session and then dividing by the length of that session.

Each automatically detected call was then manually classified based on their spectral-temporal profiles from Bezerra and Souto (2). The major call types observed were phees, trillphees, and trills. The classification of the trillphee required a minimum of 50 ms presented for each component; otherwise, it was classified as the call type that spans the majority. Other call types seen included twitter, chirp, cough, tsik, ek, and tsik-ek calls. To assess the reliability of the classification, 15 audio sessions were randomly selected (one for each marmoset, each context) and coded by a second, independent rater. The raters were 96% similar and demonstrated an interrater reliability of 92.18 (Cohen’s k).

Multiple linear regression analysis. We used the MATLAB routine (fitlm) to fit a multiple linear regression to the data. We go through our models below. The data points for these analysis were the mean values of the variable of interest for each session. To evaluate the contextual dependence of the time spent calling, we fitted the multiple linear regression model:

\[ Time\ spent\ calling = a + b \times context + c \times subject + error, \]

where context and subjects are nominal variables. Subject identity was set as a fixed variable; due to low sample size, we chose to model each marmoset individually. After running an ANOVA on the model, we reported the $P$ value for context. Similar analyses were carried out for the contextual dependence of acoustic features.

To assess the differences of call type proportions under different contexts, we fitted the multiple linear regression model:

\[ Proportion = a + b \times call\ type + c \times context + d \times subject + e \times context \times call\ type + error, \]

where call type, context, and subject are nominal variables. We reported the $P$ value on an ANOVA for call type and the interaction term of context and call type.

To determine how the acoustic features for each call type changed over contexts, we fit the following models:
Acoustic feature value = \(a + b \cdot \text{call type} + c \cdot \text{context} + d \cdot \text{subject} + \text{error}\),

where \text{call type}, \text{context}, and \text{subject} are nominal variables. After running an ANOVA on the model, we reported the \(P\) values for context and \text{call type}.

**Call rate analysis.** The call rate was calculated by taking the total number of calls and dividing by the length of the session to get calls per minute. To evaluate the contextual dependence of the call rate, we fitted the multiple linear regression model:

\[
\text{Call rate} = a + b \cdot \text{context} + c \cdot \text{subject} + \text{error},
\]

where context and subject are nominal variables. After running an ANOVA on the model, we reported the \(P\) value for context. **Heart rate data.** Heart rate signals were extracted using custom-written MATLAB code (3). The raw EMG data were first bandpassed between 5 and 39 Hz to better isolate the heartbeats. Individual heartbeats were detected using a template matching method. False positives were removed based on the average interspike intervals, and false negatives were distinguished with an adaptive local threshold. The heartbeats were binarized (2.5-ms bin width) and convoluted with a Gaussian window (\(\sigma = 0.25\) s) to get a continuous estimate of heart rates. Heart rates were averaged within each session to get a session heart rate. Heart rates from the interval 1 s before call production to call onset were averaged to get a heart rate before call. We took this heart rate interval to be a proxy for the arousal state of the animal before he/she is driven to vocalize. Calls with noisy cardiac signals were excluded from the subsequent analyses.

One concern that arises from using this method of calculating heart rate is that, at any point in time, the signal is influenced by both temporal dimensions. Thus, to check that the heart rate is not artificially elevated by the act of vocalization, we calculated the heart rate using an alternative method: the inverse of the interbeat interval (Fig. S5). The interbeat interval was calculated by taking the time interval between the peaks of consecutive heartbeats. We then calculated the reciprocal of each interval in seconds to get a heart rate value in hertz. To calculate the arousal state before call onset, the intervals from heart beats present from 1 s before call onset to call onset were selected. We then took the mean of these values. A paired sampled \(t\) test for all calls in a randomly selected session revealed that the calculated heart rates before call onset did not differ between the two methods (\(P = 0.914\)).

**Partial correlation analysis.** In the analyses of arousal effects on vocalization, to control for the effects of marmoset identity, we used partial correlation (MATLAB: partialcorr) between our variables of interest. We used this for the following analysis: (i) correlation between time spent calling in the mean heart rate for the session (\(n = 54\)); (ii) correlation between acoustic features of the call and the heart rate 1 s before call production at a session level. The \(\rho\) and \(P\) values were reported.

**Test for difference in heart rate distributions.** To test whether the heart rate before call onset was elevated relative to the mean heart rate for the session, we conducted a paired-sample \(t\) test between the session mean of the recall heart rates and the mean of the overall heart rate of the session. We reported the \(P\) value.

**Multiple linear regression for heart rate by call type.** To evaluate the correlation between heart rate and call type, we fit multiple linear regression models for all sessions and for each separate context:

\[
\text{Heart rate} = a + b \cdot \text{call type} + c \cdot \text{context} + d \cdot \text{subject} + \text{error},
\]

where call type and subject are nominal variables. We reported the \(P\) values for the effect of call type in the ANOVA of the model.

**Heart rate change by context.** The change in heart rate was calculated by taking the heart rates before call onset and subtracting the mean heart rate for the session. Positive values indicate that the heart rate was elevated before vocal production, while negative values indicate that it was suppressed. To evaluate whether there were differences between the heart rate and call types, we fit the multiple linear regression model for all sessions and for each separate context:

\[
\text{Heart rate change} = a + b \cdot \text{call type} + c \cdot \text{subject} + \text{error},
\]

where call type and subject are nominal variables. After running an ANOVA on the model, we report the \(P\) values for call type. **Coherence analysis between heart rate and vocalization.** We constructed a binary vocal pattern variable encoding the vocalization/silence states. Both vocal pattern and heart rate signals were normalized and resampled to 2 Hz. Only sessions containing more than 15 calls from the recorded marmoset were included in this analysis. For each session, we calculated the coherence between the vocal pattern and heart rate (MATLAB routine: cmtm). Mean coherence was calculated across sessions of the same context. We tested the null hypothesis that the coherence was not different from the coherence between two signals with the same power spectra but with random phase differences. We generated shuffled surrogate data by randomizing the Fourier transform phases of the signals for each session and repeated this 1,000 times. We then estimated the 95% confidence intervals. Coherence values that exceeded these bounds were significant.

**Call latencies.** We calculated call latencies to be the time interval between the end of one call and the start of the next call. A histogram of all of the latencies for each session was created to examine whether there were differences by context. The isolated context (Alone) had a peak around 10 s. All of the interactive contexts (OccFar, VisFar, VisClose) showed a second earlier peak. We then focused our analyses on the response latency, the time between the end of one marmoset’s call and the start of another marmoset’s call. In particular, only latencies shorter than 12 s were considered. Twelve seconds was chosen becomes it becomes ambiguous if calls exceeding this range are produced spontaneously or in response to the other marmoset’s call (4).

To test the dependence of call latency on acoustic features, partial correlations between the acoustic feature and the response latency were performed controlling for marmoset identity (\(n = 2,128\) calls). The \(\rho\) and \(P\) values were reported.

To examine whether different call types were made at different response latencies, we calculated the proportions of the different responded call types at each 1-s time window. We then fitted a multiple linear regression model:

\[
\text{Proportion} = a + b \cdot \text{call type} + c \cdot \text{latency} + d \cdot \text{subject} + e \cdot \text{call type} \cdot \text{latency} + \text{error},
\]

where call type and subject are nominal variables. The significance of the interaction term reveals whether the probability to respond with a certain call type depends on the response latency. \(P\) values for the interaction were reported.

To test the effect of heart rate before perception on response latency, we performed a partial correlation between the heart rate 1 s before hearing a call and the subsequent response latency with subject identity controlled. We also tested the effect of heart rate before call production on response latency with a partial correlation between the heart rate 1 s before producing a call and the previous response latency with subject identity controlled. The \(\rho\) and \(P\) values were reported.


Fig. S1. Percent time spent calling in the session by context in order of decreasing social distance for each marmoset. Error bars capture the SEM.

Fig. S2. Change in proportion of three contact call types (Phee, TrillPhee, Trill) by context for each marmoset. Error bars capture the SEM.
Fig. S3. Call rates as a function of context. Error bars capture the SEM.

Fig. S4. Heart rate changes within a session for each call type. (A) Heart rate change plotted by call type collapsed across all contexts. Scatter points are individual sessions. Error bars capture the SEM. (B) Heart rate changes by call type are shown for their respective contexts.
Fig. S5. Alternative method to calculate heart rate before call onset for one session. (A) A exemplar vocalization was selected and the corresponding waveform shown. The green tick represents the identified call onset. (B) The corresponding EMG signal is shown with identified heart beats in light green. (C) The heart rate calculated from convolving a Gaussian with binarized heart beats are shown. (D) The heart rate calculated from taking the inverse interbeat interval is shown. (E) The heart rates calculated using both methods are plotted against each other. Gray line represents where the signals would be equivalent.