NEUROSCIENCE

Vocal labeling of others by nonhuman primates

Guy Oren[1], Aner Shapira[1], Reuven Lifshitz[1], Ehud Vinepinsky[1]†, Roni Cohen[1], Tomer Fried[2], Guy P. Hadad[2], David Omer[1]*

Humans, dolphins, and elephants are the only known species that vocally label their conspecifics. It remains unclear whether nonhuman primates share this ability. We recorded spontaneous "phee-call" dialogues between pairs of marmoset monkeys. We discovered that marmosets use these calls to vocally label their conspecifics. Moreover, they respond more consistently and correctly to calls that are specifically directed at them. Analysis of calls from multiple monkeys revealed that family members use similar calls and acoustic features to label others and perform vocal learning. These findings shed light on the complexities of social vocalizations among nonhuman primates and suggest that marmoset vocalizations may provide a model for understanding aspects of human language, thereby offering new insights into the evolution of social communication.

The abilities to vocally label conspecifics and learn these labels from others are high cognitive functions in social animals. These abilities have been known to exist in humans and dolphins (*1*), and recently, they have also been observed to some extent in elephants (*2*). Vocal labeling of others requires understanding of the concept of others and may be achieved through vocal learning. It remains unknown whether nonhuman primates are capable of vocally labeling their conspecifics and learning these vocal labels from each other.

Marmosets are highly social primates that live in small family groups (two to eight animals). Marmosets rely heavily on vision but also exhibit a complex array of social calls (*3*). One such call is the phee call (*4*), a contact call, ranging from 5.5 to 10 kHz, which marmosets use to form dialogues with other group members in a turn-taking manner (*5*–*7*) when out of sight and to encode caller-related social information such as caller identity and sex (*8*, *9*). Because of the distinctive features of phee calls and the localization behavior associated with them, we hypothesized that during naturally occurring phee-call dialogues, marmoset monkeys use phee calls to label their conspecifics. Given the high degree of family cohesion in marmosets, we further hypothesized that different monkeys use similar labels to address other conspecifics and that these labels are learned among family members. We set out to test (i) whether marmoset monkeys use phee calls to vocally label their conspecifics, (ii) whether different marmoset monkeys use similar vocal labels to address other conspecifics, and (iii) whether vocal labels are learned between monkeys.

To answer these questions, we devised an experiment in which two marmoset monkeys were introduced together into the experimental room and were encouraged to naturally engage in phee-call dialogues by separating them by a visual barrier (Fig. 1C). We used a total of 10 monkeys across all experiments, members of three different family groups (Fig. 1, A and B; family groups A, B, and C; members of groups A and C were not genetically related and were all paired as mature adults; members of group B included two adult parents and their three offspring) (*10*). During the experiments, the monkeys were paired with members of both other families and their own family. The monkeys saw each other before each session started and before the placement of the visual barrier between them. Each monkey was placed in an enclosure. One enclosure (0.35 m by 0.35 m) was designed to restrict the monkey from moving in room space, whereas the other enclosure (2 m by 0.2 m) allowed the other monkey to move along the visual barrier (Fig. 1D). The monkeys were alternately positioned in either enclosure. We placed microphones in front of each enclosure to record the calls from both monkeys and used a video-based tracking system, complemented by fiducial markers on the monkeys' collars, to monitor the movements of the monkey in the long enclosure. In this setup, monkeys naturally engaged in spontaneous phee-call dialogues in a turn-taking manner (Fig. 1, E and F), which enabled us to record and compile a comprehensive and fully labeled dataset of caller and receiver interactions that amounted to 53,993 calls from various pairings (Fig. 1, A and B).

Characterizing phee calls

We correlated the monkeys' movements within the larger enclosure to the timing of their phee calls. We first identified and segmented phee calls on the basis of their time-frequ[illegible] representations (*10*). Next, we divided the [illegible] sions into 60-s-long nonoverlapping epochs and classified them into those with significant movements and those without on the basis of the monkeys' speed (epochs with significant movement were defined as epochs in which the animal's speed was above 10 cm/s in at least 10% of the epoch's duration). Epochs with movement corresponded to a significantly higher rate of phee calls (Fig. 1G; Wilcoxon rank sum test: $z = 11.97$, $P < 0.0001$; t test: $t = 11.46$, $P < 0.0001$). Additionally, we observed a notable reduction in the caller's speed around phee calls (Fig. 1H). Furthermore, on average, the phee rate decreased from 3.85 to 1.58 phees per minute within a time window of 5 min from the onset of stationary epochs (Fig. 1I; $n = 2227$ stationary epochs). These results may suggest that phee calls are used to convey the caller's location to the receiver, which is consistent with previous ethological studies that suggested that phee calls are localization calls (*11*). However, alternative explanations, such as changes in the arousal state of the caller monkey, are also possible.

Monkeys use distinct phee calls to vocally label their conspecifics

Next, we investigated whether individual monkeys use distinct phee calls when addressing different conspecifics. Because the acoustic features coding the identity of the receiver are not known, we analyzed each call's time-frequency representation (fig. S1A). This analysis allowed us to extract the frequency modulation (FM; fig. S1B, top panel) and amplitude modulation (AM; fig. S1B, lower panel) trajectories for each call, which were then normalized and resampled to a standard length, with each call embedded in an 80-dimensional features space.

To determine whether individual monkeys used a distinct phee call to communicate with each conspecific, we used random forest classifiers (*12*). These classifiers are particularly well-suited for data such as vocalizations, which are not linearly separable. This means that the different receiver classes cannot be divided by a hyperplane in high-dimensional space. For each caller monkey, we used 100 random-forest classifier models, each trained on a random balanced subset of 100 calls per receiver (*10*). The resulting confusion matrices demonstrated above-chance accuracy in identifying the intended receiver for all monkeys (Fig. 2, A and B, left panels; similar results for the remaining monkeys are displayed in fig. S3, with average accuracy, recall, and F1 scores for each monkey's classifier displayed in fig. S6). The average accuracy for each receiver monkey is shown on the diagonal of each confusion matrix and was significantly higher than shuffle distributions constructed from randomly permuted calls (*10*) (Fig. 2A, middle panel, caller

[1]Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel. [2]Benin School of Computer Science and Engineering, Hebrew University of Jerusalem, Jerusalem, Israel.
*Corresponding author. Email: david.omer@mail.huji.ac.il
†Present address: Institut de Biologie de l'ENS, Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, Paris, France.

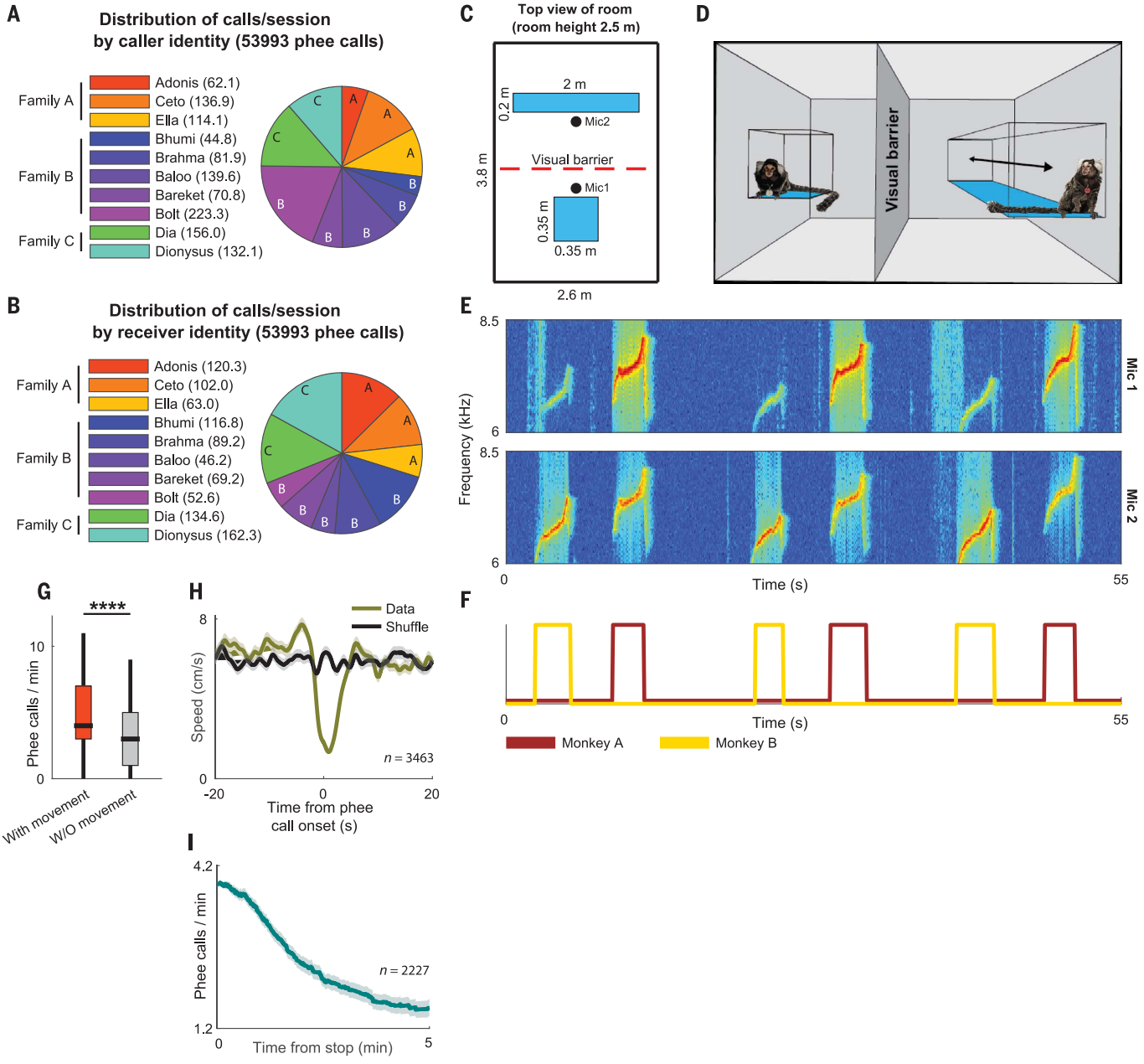

Fig. 1. Phee calls and self-location. (**A**) Pie chart showing the distribution of calls per session by caller identity (also indicated next to each caller label in parentheses in the legend). Letters indicate family group. (**B**) Pie chart showing the distribution of calls per session by receiver identity (the number of calls per session are also indicated in the legend, next to each receiver label). Letters indicate the family group. (**C**) Schematic illustration of the experimental setup, top view. Not drawn to scale. (**D**) Cartoon showing the side view of the experiment scene with small and long enclosures and the visual barrier that separated the two monkeys. Not shown to scale. (**E**) Short snippet showing an example of a phee-call dialogue between two marmoset monkeys from one of the experiments. Each panel shows a spectrogram of the audio recording. (**F**) Ethogram of the phee-call dialogue shown in (E). (**G**) Phee-call rate was significantly higher during epochs with movement than epochs without movement. Horizontal box lines indicate median, box edges indicate first and third quartiles, and whiskers indicate fifth and 95th percentiles (Wilcoxon rank sum test: $z = 11.97$, $P < 0.0001$; t test: $t = 11.46$, $P < 0.0001$). (**H**) Monkeys' velocity was significantly reduced around the occurrence of a phee call (green solid line; light green indicates SEM) as compared with shuffles (black solid line; gray indicates SEM) by triggering the monkey's velocity on random times. $n = 3463$ phee calls. (**I**) Average phee rate gradually decreased from 3.85 to 1.58 phee calls per minute within a time window of 5 min from the onset of stationary epochs (teal line; light teal indicates SEM). $n = 2227$ stationary epochs. (G), (H), and (I) suggest that phee calls are used to transmit self-location to conspecifics.

Adonis. One-tailed t test: receiver Ceto: $t = 119.38$, $P < 0.0001$; receiver Dia: $t = 73.15$, $P < 0.0001$; receiver Dionysus: $t = 87.6$, $P < 0.0001$; receiver Ella: $t = 110.3$, $P < 0.0001$. Figure 2B, middle panel, caller Ella. One-tailed t test: receiver Adonis: $t = 35.7$, $P < 0.0001$; receiver Bhumi: $t = 89.6$, $P < 0.0001$; receiver Dia: $t = 53.5$, $P < 0.0001$; receiver Dionysus: $t = 53.3$, $P < 0.0001$. A similar analysis for the remaining monkeys is shown in fig. S3). The averaged classification accuracy over all monkeys was significantly above what is expected by chance (Fig. 2E; one-tailed t test: $t = 134.08$, $P < 0.0001$). The area under the curve (AUC) of the receiver operating characteristic (ROC) curve,

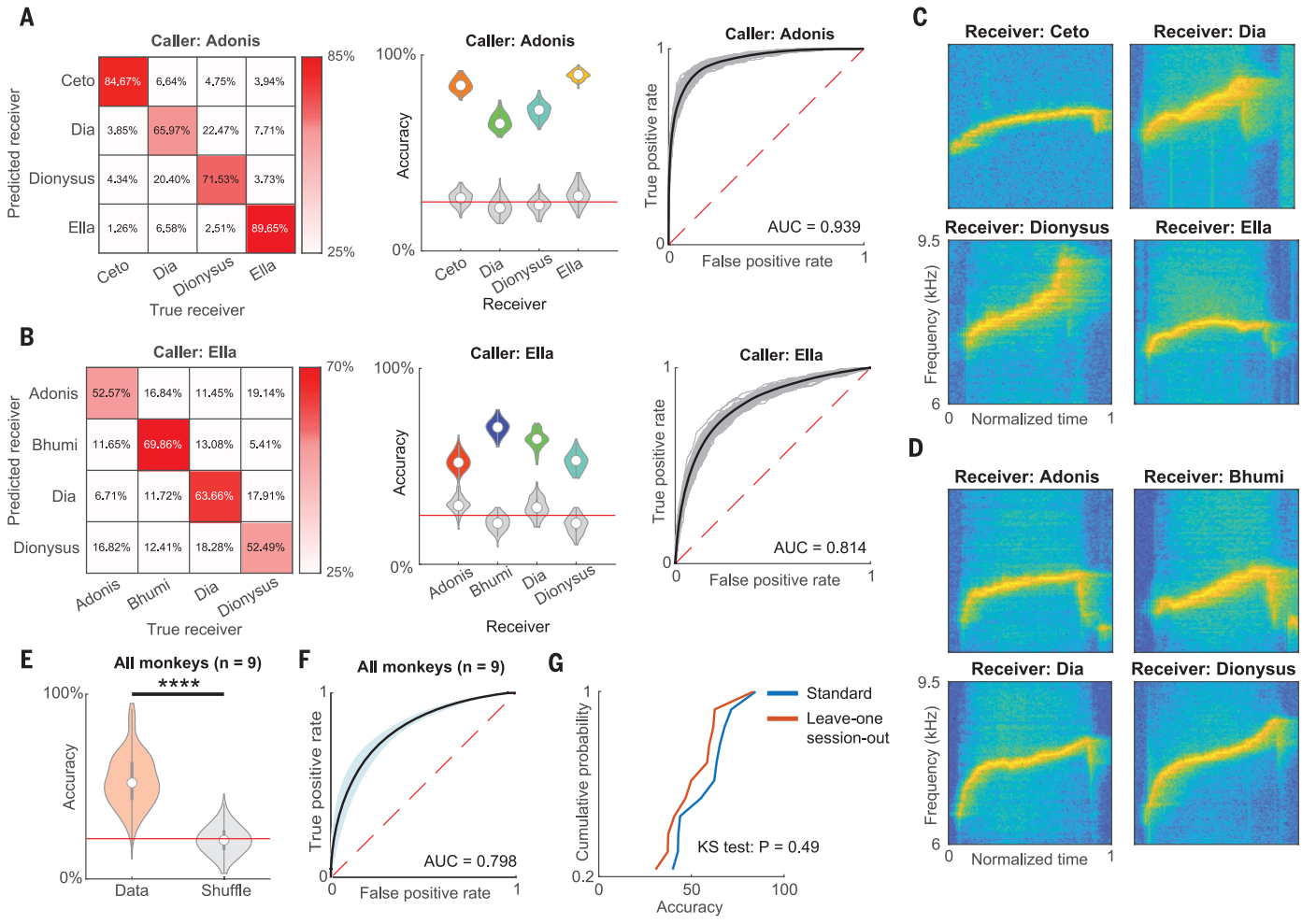

Fig. 2. Marmoset monkeys use distinct phee calls to address different conspecifics. (**A**) Average classification accuracy of 100 random-forest models trained and tested on calls from Adonis. Left panel: Mean confusion matrix for the 100 classifier models. Color code in shades of red. The lower bound of the color code is the a priori expected chance level for detection with four different receivers. Middle panel: Distribution of accuracies for each receiver identity are shown in colors (color code is the standard monkeys' color code, which is used throughout the text). Shuffle distributions are shown in gray. The red line indicates the a priori expected chance level for detection with four receiver labels. Right panel: ROC curve of each of the 100 models is shown in gray. The average ROC curve over all models is shown in a solid black line. The stippled red line indicates chance level accuracy performance. AUC = 0.939. (**B**) Same analysis as in (A), but for monkey Ella. Similar analysis for the other monkeys is shown in fig. S3. (**C** and **D**) Medoid calls, one for each of the receivers of Adonis (C) and Ella (D). (**E**) The distribution of the classifiers' accuracy across all animals is shown in pink and is significantly higher than for the same models tested on shuffled data (one-tailed *t*-test: $t = 134.08$, $P < 0.0001$). Means are indicated as white dots. The inner gray line indicates first and third quartiles. The red line indicates the a priori average chance levels across all monkeys. (**F**) Average ROC curve for all monkeys is shown in a black solid line. Light blue area indicates mean ± SD. AUC = 0.798 ± 0.065 (mean ± SD). Red stippled line indicates chance level accuracy performance. (**G**) Cumulative distributions for the standard models (blue) and the leave-one-session-out models (red) are shown. The distributions are statistically indistinguishable [Kolmogorov-Smirnov (KS) test: $P = 0.49$].

which plots the true positive rate against the false positive rate at various thresholds and is used to evaluate classification model performance, was significantly higher than 0.5 (indicating model's performance above chance level) for all monkeys (Fig. 2, A and B, right panels, Adonis: AUC = 0.939, $P < 0.0001$; Ella: AUC = 0.814, $P \leq 0.0001$; a similar analysis for the remaining monkeys is shown in fig. S3). The average AUC for all monkeys' classifiers was 0.798 ± 0.065 (mean ± SD; Fig. 2F). Spectrograms of the typical calls for each receiver varied significantly and are shown for caller monkeys Adonis and Ella in Fig. 2, C and D (each spectrogram represents the medoid call of the 100 calls for each receiver with the highest classification probability).

Examining the contribution of each set of features to the classification revealed that both AM and FM features contributed similarly across different animals with a slightly, but significantly, higher contribution by the AM features (fig. S2, A and B; $t = 5.06$, $P < 0.0001$).

The classifiers assumed that every call made by a caller in each session was intended for the conspecific behind the visual barrier. However, this assumption might not fully capture the real situation. It is likely that some calls were directed at other monkeys rather than the experimental partner. In line with this, an analysis of the classification accuracy as a function of the caller's call index within the experiment sequence, averaged across all monkeys, revealed a significant reduction in classification accuracy shortly after experiment onset (call index ~20), followed by a monotonic increase (fig. S7). This could either reflect uncertainty about the identity of the experimental partner that gradually resolved as the experiment progressed

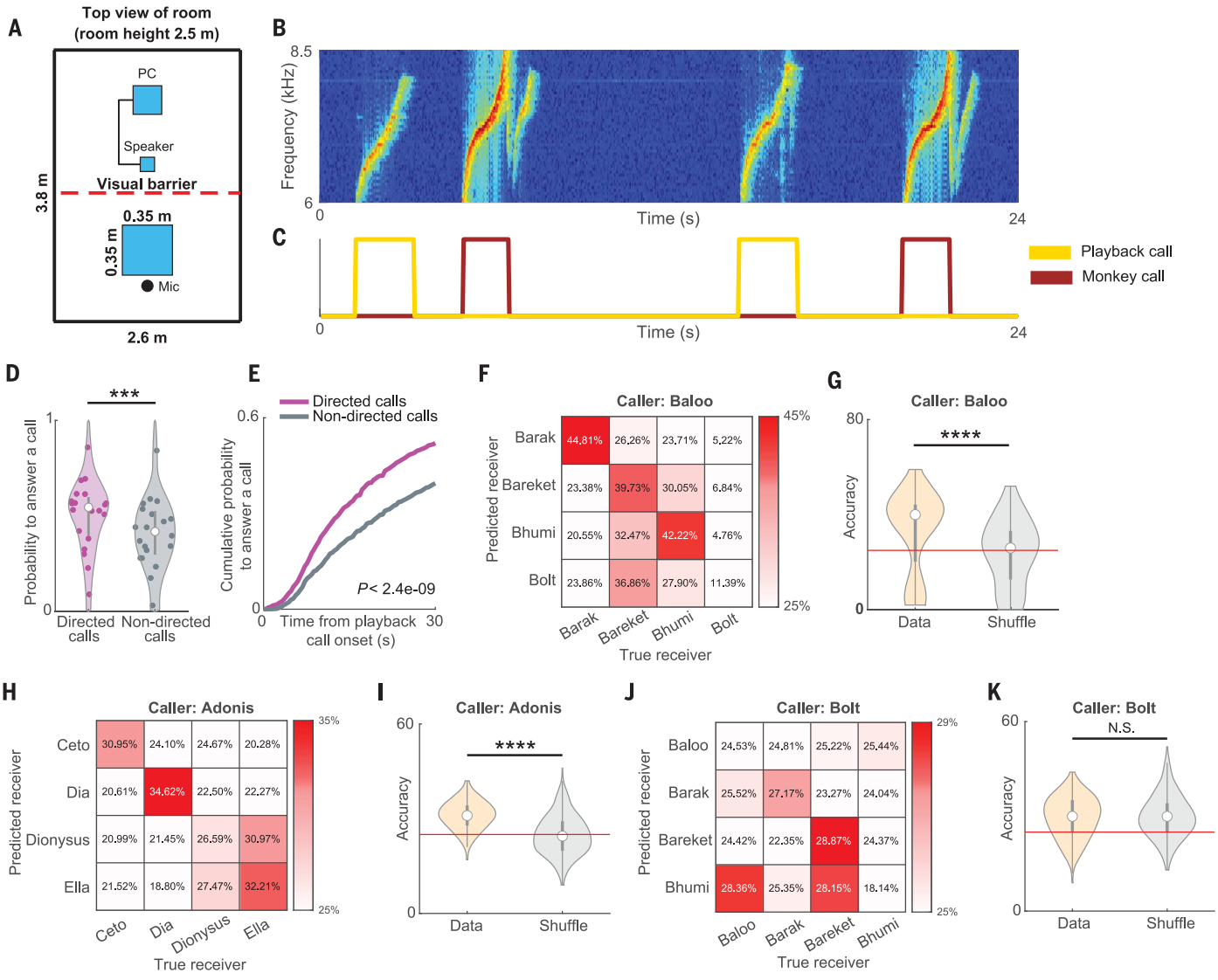

Fig. 3. Monkeys accurately perceive and respond to phee calls that are directed at them. (**A**) Schematic illustration of the experimental setup, top view. Not drawn to scale. (**B**) Snippet of the audio spectrogram showing an example of a phee-call dialogue between a monkey and the playback system. (**C**) Ethogram of the phee-call dialogue shown in (B). (**D**) Distribution of the probabilities to answer a playback directed (pink) versus nondirected (gray) calls per recording session (n = 20), pooled across monkeys. Central white dot indicates median; inner gray line indicates first and third quartiles. The average probability for answering a directed call was significantly higher (Wilcoxon signed-rank test: z = 3.88; $P < 1{\times}10^{-4}$; paired t test: t = 5.73, $P < 1.6 \times 10^{-5}$). (**E**) Cumulative distribution of probabilities to answer a playback call for the pooled data of all monkeys tested for directed (pink) and nondirected calls (gray). Cox regression test showed a significant difference between the cumulative answer probabilities for directed and nondirected playback calls (β = 1.39, $P < 2.4 \times 10^{-9}$). (**F**) Confusion matrix showing the performance of Baloo's models to detect the caller identity from Baloo's response calls to directed playback calls (χ^2 = 366.55, df = 9, $P < 0.0001$). The color code in shades of red represents the degree of accuracy. (**G**) Distribution of accuracies calculated for Baloo's calls in (F) is shown in beige against the distribution of accuracies of the shuffled data. Wilcoxon signed-rank test: z = 14.73, $P < 0.0001$. Central white dot in each violin plot indicates median; inner gray lines indicates first and third quartiles. (**H**) Same as in (F), calculated for Adonis's models (χ^2 = 1368.16, df = 9, $P < 0.0001$). (**I**) Same as in (G), calculated for Adonis. Wilcoxon signed-rank test: z = 13.73, $P < 0.0001$. (**J**) Same as in (F), calculated for Bolt's models (χ^2 = 2406.06, df = 9, $P < 0.0001$). (**K**) Same as in (G), calculated for monkey Bolt (Wilcoxon signed-rank test: z = 0.74, P = 0.5). N.S., not significant.

or a behavioral strategy that involves converging to calls addressing the experimental partner as the experiment progresses. In either case, the measured accuracy of the models is therefore likely underestimated.

Our dataset was recorded over many sessions, and the classifiers could have learned to detect differences between recording sessions rather than distinguishing between different receivers. To control for this possibility, we tested the standard classifier models, used thus far, against a new set of models trained on a leave-one-session-out basis. We constructed 100 models for each session, trained on the dataset, excluding calls from the session in question and evaluated the models' accuracy using calls from the omitted session. Subsequently, we compared the accuracy levels attributed to each receiver between the two model types. The accuracy distributions for both standard models and leave-one-session-out models

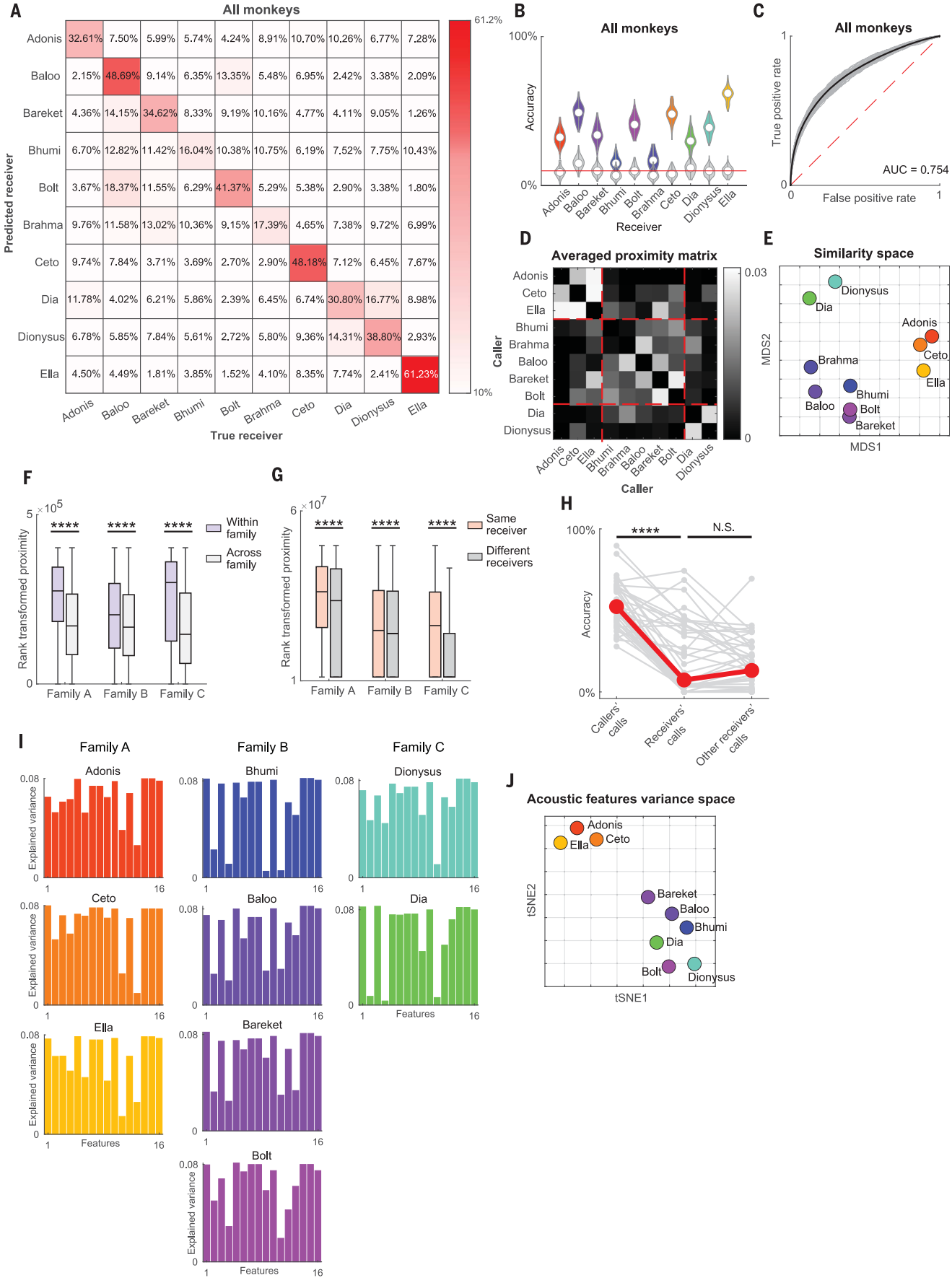

Fig. 4. Similarities of calls reveal the social structure. (**A**) Average confusion matrix over 100 models trained on calls from all monkeys to classify receivers' identity. The color code in shades of red indicates the accuracy level, with a lower bound of the a priori expected chance level of accuracy for 10 receiver identities (10%). (**B**) Distribution of accuracy levels of the 100 random forest models trained on calls from all monkeys and color coded with the standard color code of the monkeys. The distributions of shuffles are shown in gray for using the same models on a call-permuted dataset. Red line indicates the a priori chance level expected accuracy. (**C**) Average ROC curve for all 100 models shown in a black line. In gray are the ROC curves of each of the 100 models. AUC = 0.754. (**D**) Matrix showing the average proximity between calls from each pair of monkeys, using exclusively calls that were directed to the same receiver. Monkeys from the same family group tend to produce similar calls to address the same receiver. (**E**) Multidimensional scaling of the dissimilarity matrix (1 minus the similarity matrix in D). The averaged proximity is clustered into the three family groups. The standard color code for monkeys is used. MDS, multidimentional scaling. (**F**) In all three family groups, the median proximity between callers from the same group is significantly higher (purple box and whiskers plot) than between callers from different groups (gray box and whiskers plot). Horizontal box lines indicate medians. Box edges indicate first and third quartiles. Whiskers indicate the 5th and 95th percentiles. Wilcoxon rank sum test; family group A: $z = 109.9$, $P < 0.0001$; family group B: $z = 47.4$, $P < 0.0001$; family group C: $z = 75.69$, $P < 0.0001$. (**G**) In all three family groups, the median proximity between calls addressing the same receivers (orange box and whiskers plot) was significantly higher than between calls addressing different receivers (gray box and whiskers plot). Horizontal box lines indicate medians. Box edges indicate first and third quartiles. Whiskers indicate the 5th and 95th percentiles. Wilcoxon rank sum test; family group A: $z = 220.7$, $P < 0.0001$; family group B: $z = 16.9$, $P < 0.0001$; family group C: $z = 976.1$, $P < 0.0001$. (**H**) Gray lines connect between detection accuracy of the models when using the caller's calls (left), the corresponding receiver's calls to a caller (middle), and the corresponding calls by all other receivers to a caller (right). Red lines and dots indicate the median accuracy for each measurement group. Wilcoxon signed-rank test shows a significant and consistent reduction in accuracy: $z = 5.35$, $P < 4.27 \times 10^{-8}$. N.S., not significant. (**I**) Bar graphs showing the explained variance of each feature for each monkey, ordered by family groups. The standard caller monkeys' color code is used. (**J**) Projection of the explained variance vectors (16 acoustic features) for each monkey onto the acoustic features' variance 2D space, using tSNE.

were statistically indistinguishable (Fig. 2G; Kolmogorov-Smirnov test, nonparametric: $P = 0.49$), which demonstrates that phee calls directed toward the same receiver remained consistent across sessions and supports the notion that these calls serve as enduring vocal labels of others.

Monkeys accurately perceive and respond to phee calls that are directed at them

We confirmed that phee calls contain information specific to the intended receiver and proceeded to test whether the monkeys perceive and use this information. In another experiment, the monkey in the long enclosure was replaced by a closed-loop playback system (Fig. 3A; the closed-loop playback experiment). This system, which uses simple heuristics derived from conversations recorded in the "two-monkeys" experiment (*10*), succeeded in initiating and maintaining phee-call dialogues with the monkey on the other side of the visual barrier by playing back selected calls (Fig. 3, B and C). In the playbacks, we used two types of calls: directed calls, which were specifically directed at the participating monkey in the "two-monkeys" experiment; and nondirected calls, which were originally directed for any of the other monkeys in the "two-monkeys" experiment. This system allowed us to test whether monkeys respond differently to calls that were directed or nondirected at them.

Three monkeys participated in this experiment (Adonis, Baloo, and Bolt). All three monkeys exhibited a significantly higher overall average probability of answering directed calls compared with nondirected calls (Fig. 3D for pooled data; Wilcoxon signed-rank test: $z = 3.88$, $P < 10^{-4}$; paired t test: $t = 5.73$, $P < 1.6 \times 10^{-5}$; results for individual monkeys are shown in fig. S4, D to F). Additionally, the cumulative probability of vocally responding to directed calls was significantly higher than that for nondirected calls from the onset of the playback call (Fig. 3E, data pooled across all three monkeys; Cox regression test: $\beta = 1.39$, $P < 2.4 \times 10^{-9}$; results for individual monkeys are shown in fig. S4, A to C).

Next, we explored whether monkeys could accurately identify the caller when they received directed calls by responding back to the individual who originally made the call. We used each monkey's classifier models to decode the identity of the intended receiver of each response call. The classifiers' accuracy for two out of the three monkeys was significantly higher than the classification accuracy of shuffled calls (*10*) (Fig. 3, F and G, for Baloo; Wilcoxon signed-rank test: $z = 14.73$, $P < 0.0001$. Figure 3, H and I, for Adonis; Wilcoxon signed-rank test: $z = 13.73$, $P = 0.0001$. For accuracy pooled over two monkeys, we used Wilcoxon signed-rank test: $z = 20.11$, $P = 0.0001$). The average accuracy of the classification for the third monkey, Bolt, was not significantly different from chance because Bolt did not respond correctly to playback calls on average (Fig. 3, J and K; Bolt: Wilcoxon signed-rank test: $z = 0.74$, $P = 0.5$). Nevertheless, a close inspection of the confusion matrix for Bolt revealed that the classification frequencies were not randomly distributed across all receivers' labels, as well as for the other two monkeys. This was confirmed by the χ^2 test (Bolt: $\chi^2 = 366.55$, df = 9, $P < 0.0001$; and for the other two monkeys, Adonis: $\chi^2 = 1368.16$, df = 9, $P < 0.0001$; Baloo: $\chi^2 = 2406.06$, df = 9, $P < 0.0001$). Bolt showed a significant tendency to respond to playback calls from Baloo and Bareket with calls labeled as the alpha female, Bhumi, and responded correctly above chance level to some of the playback calls from Bareket and Barak. This suggests that monkeys use different behavioral strategies to respond to directed calls.

Family members use similar calls to label others and perform vocal learning

Next, we investigated whether different monkeys use similar phee calls when communicating with the same receiver. We trained random-forest classifier models on the collected calls from all monkeys to predict the intended receiver. The results, as shown by the confusion matrix in Fig. 4A that depicts the classifier's accuracy compared with shuffled data in Fig. 4B, and the ROC curve in Fig. 4C, indicated that the classifiers can predict the receiver's identity with significant accuracy when trained on calls from all monkeys, although the accuracies were lower than those achieved for individual monkeys. This observation could be attributed to one of two scenarios: (i) The significant prediction accuracy might reflect that each monkey uses a distinct call to address other monkeys and that other monkeys also use similar calls when addressing the same individual, or (ii) the accuracy may solely represent the individualized call usage by each monkey to address others, with each different monkey using a different distinct call for the same receiver. To differentiate between these scenarios, we applied a pairwise proximity measure to calls from the all-monkeys classifiers. Proximity is a similarity measure that is defined as the proportion of times that two calls ended up in the same leaf of a decision tree, which thus reflects their similarity (*10*). It ranges from 0 to 1, with higher proximity values indicating greater similarity between calls. Using proximity as a similarity measure between calls reflects their similarity based on the predicted receiver label without assuming a linear relationship between

calls. We computed the average proximity between calls from various callers by focusing each time exclusively on calls that were made by each of the two different callers to the same receiver. The resulting averaged proximity matrix in Fig. 4D reveals a pattern that mirrors the monkeys' family grouping, which suggests that calls addressing the same receiver from monkeys from the same family group tend to be similar. We then embedded the calls in the similarity space by applying multidimensional scaling on the transformed dissimilarity matrix (1 minus the similarity matrix). The embedding further corroborated this finding by illustrating the distinct clustering of callers into the three family groups (Fig. 4E). We further compared the average proximity between callers from the same family group and callers from other groups when addressing the same receiver (Fig. 4F). In all three family groups, the similarity between calls to the same receiver was significantly higher between family members than with members of the other family groups (Wilcoxon rank sum test; family group A: $z = 109.9$, $P < 0.0001$. Family group B: $z = 47.4$, $P < 0.0001$; family group C: $z = 75.69$, $P < 0.0001$).

To determine whether the similarity between calls addressing the same receiver is specific to the modulation pattern encoding the receiver's identity rather than a nonspecific convergence between family members' calls, we compared the average proximity between family members' calls addressing the same receiver and those addressing different receivers (Fig. 4G). Our results show a significantly higher similarity between family members' calls addressing the same receiver compared with those addressing different receivers (Wilcoxon rank sum test; family group A: $z = 220.7$, $P < 0.0001$; family group B: $z = 17$, $P < 0.0001$; family group C: $z = 976.2$, $P < 0.0001$).

These findings suggest that scenario one, in which monkeys from the same family group use similar calls when addressing the same individual, is the correct scenario. Furthermore, the observed higher similarity in calls among family members compared with members of other family groups for groups A and C (rank transformed proximities: group A, 2.74×10^5; group C, 2.99×10^5), none of whom are genetically related and who were all paired as mature adults, suggests that marmoset monkeys use vocal learning to adjust their calls for different recipients, which makes their calls resemble those directed toward the same receivers by other group members.

Vocal labels of others can be learned by listening to dialogues between other monkeys and imitating the calls of other callers to different receivers, or by imitating the receiver's calls during active participation in a dialogue, similar to how dolphins were reported to label their conspecifics (*13*). We tested the hypothesis that vocal labeling of others in marmoset monkeys relies on imitation of the receiver's calls. If this hypothesis is correct, we would expect each monkey's classifier to classify calls made by the caller to the receiver similarly to calls made by the receiver to the caller. In other words, we anticipate that the classification accuracy for calls made by the caller to the receiver would be comparable to calls made by the receiver to the caller. However, we observed a significant reduction in the model's classification accuracy for calls made by the receiver to the caller compared with those made by the caller to the receiver across all monkeys (Fig. 4H; Wilcoxon signed-rank test between callers calls to a receiver and receivers calls addressing the caller: $z = 5.35$, $P < 0.0001$). Furthermore, there was no significant difference in the classification accuracy between calls made by the receiver to the caller and calls made by all other receivers to the same caller (Fig. 4H; Wilcoxon signed-rank test: $z = 0.49$, $P = 0.31$). This indicates that learning to vocally label conspecifics in marmoset monkeys does not rely on the imitation of the receiver's calls during active dialogue.

Encoding receivers' identities by acoustic features

We further aimed to identify whether there are predefined acoustic features of phee calls that encode the receiver's identity—for example, mean fundamental frequency, frequency slope, and amplitude slope, among others. We defined a set of 16 acoustic features (fig. S5A) (*10*) comprising 8 features from the FM aspect and another 8 from the AM aspect of the calls (*10*). For each caller, we selected 100 calls addressed to each receiver (100 calls per receiver), focusing on those with the highest classification probability to obtain a balanced subsample of callers. Dimensionality reduction analysis revealed a distinct clustering pattern for receiver identity for all monkeys (fig. S5B), which indicates that the 16 features captured some of the encoding of the receiver's identity. We then used principal component analysis to reduce the data's dimensionality to three dimensions by concentrating on the principal components that, on average, accounted for 75% of the variance (fig. S5C; mean explained variance 75 ± 3%). Subsequently, we calculated the explained variance for each original acoustic feature for every monkey (Fig. 4I; fig. S8 shows the average explained variance for each feature, sorted by the amount of explained variance and shown for each family group). Because all 16 predefined features contributed to some extent to the explained variance across all monkeys, we concluded that no specific subset of the acoustic features solely encodes the receiver's identity.

Do monkeys from the same group use similar acoustic features to encode the receiver's identity? That is, do the explained-variance vectors of monkeys belonging to the same family group exhibit clustering? We embedded the explained variance vectors of all monkeys in a two-dimensional acoustic features space using t-distributed stochastic neighbor embedding (tSNE) (Fig. 4J) (*14*). This analysis revealed a clustering of callers into the family groups. Members of family groups A and B each demonstrated a tendency to use a similar pattern of explained-variance features for encoding the receiver's identity. By contrast, group C displayed a less defined clustering pattern in two dimensions and tended to cluster more closely with group B.

Discussion

We found that marmoset monkeys use phee calls to vocally label their conspecifics, distinguish between phee calls that were directed at them as opposed to nondirected calls, and can respond correctly to the caller's identity. These results could not be explained by a variation in calls between sessions. Furthermore, we observed that monkeys from the same family group tend to use similar calls to vocally label others and use similar acoustic features to encode the identity of others. Additionally, family members' calls addressing the same receivers were significantly more similar than calls addressing different receivers. These similarities were observed even among individuals who are not genetically related (family groups A and C), which implies that vocal learning may occur among adult members of a family group.

The natural habitat of marmoset monkeys is the densely foliated rainforest, in which visual occlusions may endanger group cohesion. Vocal labeling of others may be an evolved behavior that aids group cohesion and survival. Humans (*Homo sapiens*), bottlenose dolphins (*Tursiops truncates*) (*1*, *15*), and African elephants (*Loxodonta africana*) (*2*) are the only species that have been reported so far to vocally label their conspecifics. The finding in dolphins seems to be different from our finding in marmoset monkeys because each caller dolphin uses a distinct signature whistle, and other caller dolphins mimic each other dolphin's signature whistle to label others. By contrast, we show here that, in marmoset monkeys, vocal labels of conspecifics are not an imitation of receiver's calls (Fig. 4H).

Social vocal accommodation (*16*, *17*) is the capacity to modify existing vocalizations in social contexts and was reported to exist in both humans and nonhuman primates, including in marmoset monkeys (*18*). In contrast to previously reported cases of social vocal accommodation, the vocal labeling of others that we report here extends far beyond social vocal accommodation and reflects the capacity to modify the fine acoustic structure of a call to

encode the receiver's identity. It is distinct from all the previously reported examples of social vocal accommodation for the following reasons that are not expected from social accommodation alone. First, by using classifier models trained on calls from each monkey, we show that the modifications of the acoustic features of phee calls are distinct to each receiver (Fig. 2). Second, we demonstrate that marmosets respond to directed calls with a higher probability than to nondirected calls, which suggests that marmosets can perceive the encoded receiver's identity (Fig. 3, D and E). In addition, we show that the similarity between calls by two family members addressing the same receiver is significantly higher than the similarity between calls made by two family members to different receivers (Fig. 4G). This result supports the existence of a similarity component between calls that is specific to those addressing the same receiver, which resembles the concept of naming.

Our results also indicate that marmosets perform vocal learning by imitating other group members' calls to other receivers (Fig. 4, D to G) and by using similar patterns of acoustic features to encode other receivers (Fig. 4, I and J). Vocal learning in marmosets was reported before to occur during development and in adults (*19*–*23*). Although a developmental process might explain the similarity between calls in group B (two parents and their offspring), this is not the case for groups A and C because these animals were not genetically related and were paired as adults to form a family group. Furthermore, because vocal labeling of others depends on the recognition of conspecifics, it is likely learned among family members, as monkeys are not born with the knowledge of the vocal labels of their future social companions. It remains to be investigated in future studies whether vocal labels are learned through a foundational, step-by-step learning process or whether monkeys are born with a predefined set of labels that they then learn to associate with different individuals.

Our findings also provide new insights into the evolutionary discontinuity between human language and vocal communication in nonhuman primates. This insight is based on the accepted notion that the evolution of human brain mechanisms supporting social cognition are the cognitive precursors of language (*24*) and the continuity in brain mechanism supporting social cognition between nonhuman primates and humans (*25*). Marmoset monkeys are highly social nonhuman primates, and they diverged from our common ancestor ~35 million years ago (*26*). Nevertheless, they exhibit notable similarities to humans in their social structure. Marmosets live in small family groups of 6 to 8 members and are among the few nonhuman primates that pair-bond and show cooperative care of their young (*27*). These similarities suggest that marmosets faced similar social evolutionary challenges as our prelinguistic humanoid ancestors, which might have pressured them to develop similar brain mechanisms to support social cognition and enable them to navigate their social world.

The vocal labeling of others that we report here represents a learned, highly flexible call production that requires brain mechanisms for representations of others as discrete concepts, vocal learning, imitation, and modification of the acoustic fine structure of calls. These mechanisms might be similar to those that facilitated the evolutionary transition from nonlinguistic communication to language in our prelinguistic humanoid ancestors.

REFERENCES AND NOTES

1. V. M. Janik, L. S. Sayigh, R. S. Wells, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8293–8297 (2006).
2. M. A. Pardo *et al*., *Nat. Ecol. Evol.* **8**, 1353–1364 (2024).
3. B. M. Bezerra, A. Souto, *Int. J. Primatol.* **29**, 671–701 (2008).
4. H.-C. Chen, G. Kaplan, L. J. Rogers, *Am. J. Primatol.* **71**, 165–170 (2009).
5. D. Y. Takahashi, D. Z. Narayanan, A. A. Ghazanfar, *Curr. Biol.* **23**, 2162–2168 (2013).
6. C. T. Miller, X. Wang, *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **192**, 27–38 (2006).
7. T. Pomberger, C. Risueno-Segovia, J. Löschner, S. R. Hage, *Curr. Biol.* **28**, 788–794.e3 (2018).
8. J. L. Norcross, J. D. Newman, *Am. J. Primatol.* **30**, 37–54 (1993).
9. C. T. Miller, K. Mandel, X. Wang, *Am. J. Primatol.* **72**, 974–980 (2010).
10. Materials and methods are available as supplementary materials.
11. D. A. G. Oliveira, C. Ades, *An. Acad. Bras. Cienc.* **76**, 393–398 (2004).
12. L. Breiman, *Mach. Learn.* **45**, 5–32 (2001).
13. V. M. Janik, *Science* **289**, 1355–1357 (2000).
14. L. Van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
15. S. L. King, V. M. Janik, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13216–13221 (2013).
16. H. Ruch, Y. Zürcher, J. M. Burkart, *Biol. Rev. Camb. Philos. Soc.* **93**, 996–1013 (2018).
17. V. M. Janik, P. J. B. Slater, *Anim. Behav.* **60**, 1–11 (2000).
18. Y. Zürcher, E. P. Willems, J. M. Burkart, *Sci. Rep.* **11**, 15683 (2021).
19. D. Y. Takahashi *et al*., *Science* **349**, 734–738 (2015).
20. D. Y. Takahashi, D. A. Liao, A. A. Ghazanfar, *Curr. Biol.* **27**, 1844–1852.e6 (2017).
21. Y. B. Gultekin, S. R. Hage, *Sci. Adv.* **4**, eaar4012 (2018).
22. Y. B. Gultekin, S. R. Hage, *Nat. Commun.* **8**, 14046 (2017).
23. Y. B. Gultekin, D. G. C. Hildebrand, K. Hammerschmidt, S. R. Hage, *Sci. Adv.* **7**, eabf2938 (2021).
24. M. D. Hauser, N. Chomsky, W. T. Fitch, *Science* **298**, 1569–1579 (2002).
25. R. M. Seyfarth, D. L. Cheney, *The Social Origins of Language*, M. L. Platt, Ed. (Princeton Univ. Press, 2018).
26. Y. Shao *et al*., *Science* **380**, 913–924 (2023).
27. C. T. Miller *et al*., *Neuron* **90**, 219–233 (2016).
28. G. Oren *et al*., Zenodo; https://doi.org/10.5281/zenodo.12721811.

ACKNOWLEDGMENTS

We thank I. Nelken, Y. Yovel, Y. Grodzinsky, D. Dor, P. Schlenker, E. Chemla, and the Schlenker lab members for valuable comments and discussions. We also thank A. Mizrahi, M. London, and T. Eliav for comments on the manuscript; T. Ravins, N. Eshkol, and J. Yagil for veterinary support; E. Sarig for preliminary experiments; G. Ogen for animal handling; and O. Ozana and the ELSC administration staff for administrative support. We thank A. Ghazanfar for sharing with us the code for the close-loop playback experiments. **Funding:** This study was supported by research grants to D.O. from the European Research Council (ERC—SYG, OxytocINspace) to D.O., and the Israel Science Foundation (ISF 915/22) to D.O. **Author contributions:** Conceptualization: D.O.; Methodology: D.O., G.O., A.S., R.L., E.V., and R.C.; Investigation: D.O., G.O., T.F., and G.P.H.; Visualization: D.O. and G.O.; Funding acquisition: D.O.; Project administration: D.O.; Supervision: D.O.; Writing – original draft: D.O. and G.O.; Writing – review and editing: D.O., G.O., and E.V. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials or are deposited at Zenodo (*28*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adp3757
Materials and Methods
Supplementary Text
Figs. S1 to S8

Submitted 22 March 2024; accepted 25 July 2024
10.1126/science.adp3757