universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Multimodal Recognition in Common Marmosets

verfasst von | submitted by

Ilsong Jeon

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 878
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Verhaltens-, Neuro- und Kognitionsbiologie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Thomas Bugnyar

Abstract auf Deutsch

Multimodale Kommunikation spielt eine zentrale Rolle in der sozialen Organisation verschiedener Arten. Die multimodale Erkennung ist ein kognitiver Prozess, bei dem unterschiedliche sensorische Modalitäten integriert werden, um eine individuelle Repräsentation zu bilden. Dieser Prozess ermöglicht Tieren möglicherweise, ihre soziale Umgebung effektiver einzuschätzen. Während bei verschiedenen Arten gezeigt wurde, dass sie Artgenossen modalitätsübergreifend erkennen können, ist über die multimodale Erkennung von artfremden Individuen noch wenig bekannt.

Ziel dieser Studie war es, festzustellen, ob Weißbüschelaffen (Callithrix jacchus) spezifische Menschen anhand zweier sensorischer Modalitäten – auditiver (Stimme) und visueller (Gesicht) Informationen – identifizieren können. Die Studie wurde mit 15 erwachsenen Weißbüschelaffen und vier menschlichen Teilnehmern durchgeführt: Zwei der Menschen waren den Affen vertraut, die anderen beiden waren ihnen unbekannt. Das Experiment umfasste vier Bedingungen: (1) vertraute menschliche Stimme und Gesicht stimmen überein, (2) vertraute menschliche Stimme und Gesicht stimmen nicht überein, (3) unvertraute menschliche Stimme und Gesicht stimmen überein, (4) unvertraute menschliche Stimme und Gesicht stimmen nicht überein.

Wir erwarteten, dass Weißbüschelaffen multimodale Erkennung zur Wahrnehmung von Menschen nutzen und je nach Kongruenz zwischen menschlichen Gesichtern und Stimmen unterschiedliche Verhaltensweisen zeigen. Darüber hinaus vermuteten wir, dass die Affen die stärkste „Schreckreaktion“ zeigen würden, wenn die Modalitäten bei den vertrauten Menschen nicht übereinstimmen. Um diese Hypothese zu testen, wurden aufgenommene Videos kodiert und mittels Heatmaps sowie Hauptkomponentenanalyse (PCA) analysiert. Die PCA ergab jedoch keine klaren Cluster der Hauptkomponenten. Daher wurden einzelne Verhaltensweisen zwischen den Bedingungen mit dem Friedman-Test analysiert und für signifikante Ergebnisse anschließend Dunn’s Post-hoc-Tests durchgeführt. Wir fanden heraus, dass sich „Blickdauer“ (Gaze Duration) und „Kopf-Schwenkfrequenz“ (Head Swaying Frequency) zwischen den Bedingungen unterschieden, wobei die Affen hauptsächlich auf die Vertrautheit der menschlichen Teilnehmer reagierten. Entgegen unseren Vorhersagen hatte die Kongruenz der audio-visuellen Stimuli innerhalb der vertrauten Kategorie keinen signifikanten Einfluss.

Diese Erkenntnisse tragen zu einem tieferen Verständnis der Fähigkeit von Weißbüschelaffen bei, individuelle Merkmale von Menschen zu erkennen, und bilden die Grundlage für zukünftige Studien zur kognitiven Interaktion zwischen Menschen und nicht-menschlichen Primaten.

Abstract in English

Multimodal communication plays a pivotal role in the social organization of various species. Multimodal recognition is a cognitive process by which different sensory modalities are integrated to form an individual representation. This process may enable animals to more effectively assess their social surroundings. While a variety of species have been demonstrated to recognize conspecifics cross-modally, little is known about the multimodal recognition of heterospecifics.

This experiment aimed to determine whether common marmosets (*Callithrix jacchus*) can identify specific humans using two sensory modalities, auditory (voice) and visual (facial) information. The study was conducted with 15 adult individuals and involved four human participants: two of them were familiar to the marmosets, whereas the other two were unfamiliar. The experiment contained four conditions: (1) familiar human face-voice match, (2) familiar human face-voice mismatch, (3) unfamiliar human face-voice match, (4) unfamiliar human face-voice mismatch.

We expected that marmosets would utilize multimodal recognition to perceive humans and exhibit behavioral differences based on congruency between human faces and voices. Additionally, we predicted them to exhibit the strongest 'startle' response when the modalities from familiar humans are mismatched. To test this hypothesis, recorded videos were encoded and analyzed using heatmaps and PCA. However, PCA did not produce clear principal component clusters; therefore, individual behaviors were analyzed across conditions using the Friedman test, with Dunn's post-hoc tests performed on behaviors showing significant Friedman test p-values. We found that "Gaze Duration" and "Head Swaying Frequency" differenced between conditions, whereby the monkeys' responded primarily to the familiarity with the human participants. In contrast to our predictions, the congruency of audio-visual stimuli within the familiar category had no significant effect.

These findings contribute to a deeper understanding of marmosets' ability to recognize individual human characteristics and lay the groundwork for future studies exploring cognitive interactions between humans and non-human primates.

1. Introduction

1.1 Background

Sensation is a crucial aspect of an organism's life, yet relying on a single sense has its limitations. For instance, auditory communication enables species to distinguish members of their own species from others (Arasaki et al., 2023). While monomodal recognition is functional, it has limitations. When sensory information is ambiguous or incomplete, organisms may compensate for these limitations by integrating multiple sensory inputs, a process known as multimodal, or cross-modal, recognition (Han et al., 2019; Chaka et al., 2018). This allows for more accurate environmental assessment, enhanced learning, and improved communication (Forceville, 2012; Potvin, 2017).

Various primates demonstrate species recognition through monomodal sensory processing. Old World monkeys such as De Brazza monkeys (*Cercopithecus neglectus*), Campbell's monkeys (*Cercopithecus campbelli*), Guereza colobus monkeys (*Colobus guereza*), and red-capped mangabeys (*Cercocebus torquatus*) can distinguish between the voices of their own species and those of other species (Candiotti et al. 2013). Similarly, the rhesus macaques (*Macaca mulatta*) can identify the voices of familiar and unfamiliar humans (Sliwa et al. 2011). Western gorillas (*Gorilla gorilla*) also exhibit different responses to sensory stimuli from familiar and unfamiliar humans. For instance, they show longer attention times, more frequent gazing, and quicker responses to unfamiliar voices or individuals who have had previous negative interactions. Conversely, voices of people deemed safe or non-threatening, who have had familiar and positive interactions, elicit relatively lower responses (Salmi et al., 2021).

Although monomodal recognition plays a basic role in perception and social interaction, its limitations underscore the significance of multimodal recognition. In animal communication, multimodal recognition enables receivers to perceive their environment more quickly and accurately. Furthermore, this enhanced ability allows species to develop unique communication strategies, facilitating the recognition of conspecifics and the differentiation of individuals. Consequently, social animals establish familiarity with others which in turn helps them recall past interactions and the contributes to formation the of social hierarchies, relationships and the stabilization of social structure (Tibbetts & Dale, 2007; Dubois et al., 1999).

For instance, domestic horses (*Equus caballus*) integrate auditory and visual cues to recognize familiar individuals. In a seminal study, it was shown that horses respond more quickly and for a longer duration when presented with mismatched auditory-visual cues, indicating their reliance on multimodal integration for individual recognition (Proops et al., 2009). Similarly, rhesus macaques (*Macaca mulatta*) also use multimodal cues for individual recognition. They rely on both facial and vocal information to recognize familiar individual, rather than using a single sensory modality (Habbershon et al., 2013). This capability not only aids in recognizing members of the same species but also enhances survival and reproduction by allowing animals to effectively compete or cooperate as necessary (Tibbetts & Dale, 2007).

Such multimodal recognition also occurs in heterospecific contexts. Russet-naped wrens exhibited strong aggressive responses to conspecific vocal and visual signals but showed lower response levels to heterospecific stimuli (Ku-Peralta & Sosa-López, 2025). Additionally, female zebra finches increased their confusion responses when exposed to mismatched vocal and visual signals from conspecifics and heterospecifics, often refraining from making a choice or exhibiting neutral reactions (Campbell & Hauber, 2010).

The common marmoset (*Callithrix jacchus*), a New World primate, generally lives in small family groups of 4 to 15 members in dense forests (da Cunha & Byrne, 2009). The group is dominated by breeding pairs. The gregarious species exhibit high levels of social cohesion and strengthen social bonds through cooperative care for offspring (Miller, 2016). They also possess excellent facial recognition abilities within a strong social structure (Schäffer et al., 2020), these cognitive and social traits, combined with their strong reliance on multimodal recognition, make them an ideal subject for this study. Additionally, the species exhibits sophisticated cognitive abilities, including facial recognition and vocal communication (Diehl & Romanski, 2013; Reiss, McCowan, & Marino, 1997), which allow researchers to compare its perception processes with those of other primates. Furthermore, the common marmoset's cooperative breeding system and complex social dynamics provide a unique opportunity to investigate the evolution of social cognition. Lastly, due to their increasing use in neuro-cognitive research, understanding the cognitive abilities of marmosets holds future implications for both comparative psychology and neuroscience, enhancing our knowledge of primate cognition and evolutionary development.

1.2 Aims & Hypotheses

This experiment investigates the common marmoset's ability to distinguish between human voices, focusing on their active reactions to familiar and unfamiliar humans. Additionally, it

aims to examine their response when a human's voice and face are presented in a mismatched manner, as impaired multimodal recognition is expected to elicit the highest level of reaction.
To verify this, we conducted an experiment with four different conditions.
First, we played the voice of a familiar human after which the corresponding individual entered, naming this condition "Familiar_match." Next, we played the voice of a familiar human followed by the entrance of a different familiar individual, naming this condition "Familiar_unmatch."
Similarly, we played the voice of an unfamiliar human followed by the corresponding individual entering, which we named "Unfamiliar_match." Lastly, we played the voice of an unfamiliar human after which a different unfamiliar individual entered, naming this condition "Unfamiliar_unmatch."
We expected that, like Old World monkeys, marmosets will also use multimodal recognition to distinguish between familiar and unfamiliar humans.
Firstly, when a human's voice and face are presented in a mismatched manner, the impaired multimodal recognition is expected to elicit the highest level of response.
Secondly, we expected marmoset to exhibit a lower 'startle' response, such as a shorter 'gaze duration', toward familiar humans, while displaying a relatively higher startle response, such as a longer 'gaze duration', toward unfamiliar humans.
We anticipated that the 'startle' response would not differ significantly between the two unfamiliar conditions because, in the case of unfamiliar human participants, marmosets are not expected to recognize the mismatch between the voice and face when perceiving them.

2. Materials & Methods

2.1 Marmoset Subjects & Human Participants

This experiment was conducted on a total of 15 adult common marmosets housed at the University of Vienna Biology Building (UBB), consisting of 9 males and 6 females.
The monkeys are housed in three keeping rooms within the UBB building, which are organized by group (App. 1), and are fed and participate in experiments according to a predetermined schedule. The temperature and humidity in each housing room are kept at stable levels, and a professional husbandry team is responsible for monitoring and maintaining the monkeys' body weight and overall health.
All individuals have ample experience with humans, through daily care taking and their participation in cognitive experiments and training programs (e.g. tunnel training and transport box training).
Human participants were recruited from Biology students. The age range for participants was 23 to 32 years. To control for possible sex effects, they consisted exclusively of women. Participants were divided into two groups, differing in the degree of familiarity to the monkeys. The Familiar Group consistend of Monkey Lab students who had been in contact with marmosets for approximately four months or longer, whereas the Unfamiliar Group consisted of general students who had no prior contact with marmosets

2.2 Experimental Design

Playbacks were recorded in an empty office using a Zoom H4n Pro and a Sennheiser ME66 microphone (+K6 power module) for each person. The playback phrases were: *"Hallo! Lange nicht gesehen, ja? Ich komme rein!"*, *"Ich bringe dein Essen! Heute gibt's etwas*

Spezielles. Es muss lecker sein!”, and *“Ich muss jetzt gehen! Gute Nacht, wir sehen uns morgen!”* Each playback phrase lasted about 5 seconds.
Two curtains were installed on both sides of the C2 cage, and two speakers were positioned between the doors and the curtains. The participants were positioned approximately 2 meters away from the C2 cage. The observer was positioned within the laboratory but as far away from the C2 cage as possible to record the experimental process and to easily see behind the two curtains (Fig. 1).
Cameras were installed above the C2 cage and near the observer to avoid obstructing the subject’s line of sight and to maximize the recording angle.
To minimize additional stress and limit potential stimuli beyond facial expressions and voice, participants followed the following criteria: all participants wore dark-colored clothing, and they tied their hair up. Since marmosets can read the facial expressions of their species (Dureux et al., 2023), human participants maintained a neutral expression until leaving the room. While it has not yet been confirmed whether marmosets can interpret human facial expressions, a neutral expression was enforced to control for this potential factors.
To prevent test subjects from becoming habituated, each marmoset was only tested once every 48 hours.

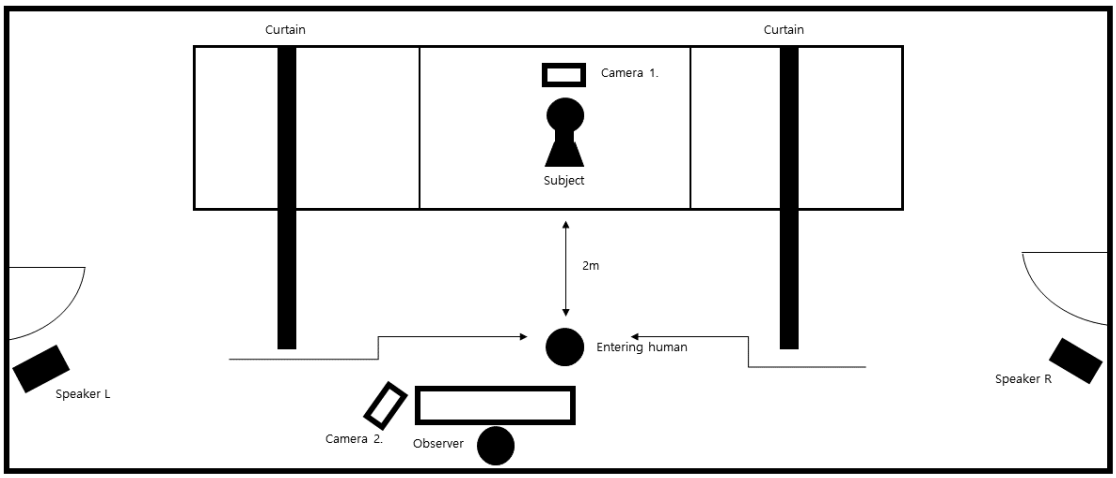


Figure 1. Set up of Experiment

2.3 Data collection

Each experimental subject was guided to the experimental room through a tunnel at the beginning of each session. Upon entry, a one-minute stabilization period took place. During this time, a snack—either a waffle or a piece of fruit—was provided.
During the stabilization phase, if an individual showed frequent startle behaviors, the experiment was immediately stopped, and the individual was returned to the group. After stabilization, the playback was presented through a speaker. After the playback, a human participant entered the experimental room, moved to a designated location, and then exited through the same route.

The video was recorded 30 seconds before the playback began and continued until 30 seconds after the participant left.
The experiment was conducted with counterbalancing, with randomized entry positions for participants, and the three different playback phrases were randomized. Each marmoset underwent a total of four sessions: familiar group match (Familiar_match), familiar group not-match (Familiar_unmatch), unfamiliar group match (Unfamiliar_match), and unfamiliar group not-match (Unfamiliar_unmatch).
Each marmoset completed two trials per day. The number of marmosets tested per day was limited to 5, and the total number of experiment days was scheduled for 12 days.
The total time each participant spent in the experiment room was fixed at 30 seconds. This duration was consistent with previous experiments conducted on Old World monkeys, in which the reaction time for the playback was also fixed at 30 seconds (Candiotti et al., 2013).

2.4 Video-Analysis

Video recording was conducted using focal sampling and continuous recording methods, focusing on a 30-second period from the moment participants entered until they exited the experiment room.
The behavioral patterns analyzed included vocalizations including isolation calls (long Phee), Tsik (single), Twitter, Phee, Ek, and Whirr/Trill.
Head movements observed included head swaying, head-cock stare, and gaze, while body movements consisted of grooming (self-scratch), freezing, and agitated locomotion.
Lastly, wariness-related behaviors were characterized by piloerection, genital display, and self-scratching.

2.5 Data Analysis

The recorded videos are analyzed with BORIS (Behavioral Observation Research Interactive Software, version 8.5.4), and the generated Excel data is subjected to statistical analysis and visualized with Python (Spyder ver. 5.5.1).
The analysis began with the creation of a heatmap (Fig. 2) to examine correlations between each behavior. Principal Component Analysis (PCA) was then performed, identifying the number of principal components (PCs) required to achieve cumulative explained variances of approximately 80%. A scree plot was generated to visualize the variance explained by each PC (Fig. 3 & Tab. 2).
Behavioral loadings were calculated for around 8 selected PCs (Tab. 3), and a 3D scatter plot was created using the 3 PCs with the highest explained variances (Fig. 4). This scatter plot was used to identify clusters among behaviors and individual behaviors were compared across different conditions.
After testing the distribution of each comparative analysis, non-followed a normal distribution. Given that the same subjects were tested under different conditions, the data were considered dependent. Therefore, the P value was calculated using the Friedman test, which determines whether there are significant differences by conditions. If the Friedman test result was significant, Dunn's test for post-hoc test was conducted to compare differences between each group. Additionally, for pairwise comparisons between the Unfamiliar and Familiar condition groups, the Wilcoxon signed rank test was performed. Dunn's test results are shown with lines and asterisks on the boxplot, where * corresponds to p-value < 0.05, ** to p-value < 0.01, and *** to p-value < 0.001. Since Dunn's test performs pairwise comparisons among four conditions (Familiar_match, Familiar_unmatch, Unfamiliar_match

and Unfamiliar_unmatch), it does not directly compare the Familiar and Unfamiliar conditions. Therefore, to exmine the effect of Familiarity, a Wilcoxon signed rank test was conducted to determine whether there was s significant difference between Familiarity.

3. Results

3.1 Heatmap & Correlation between Behaviors

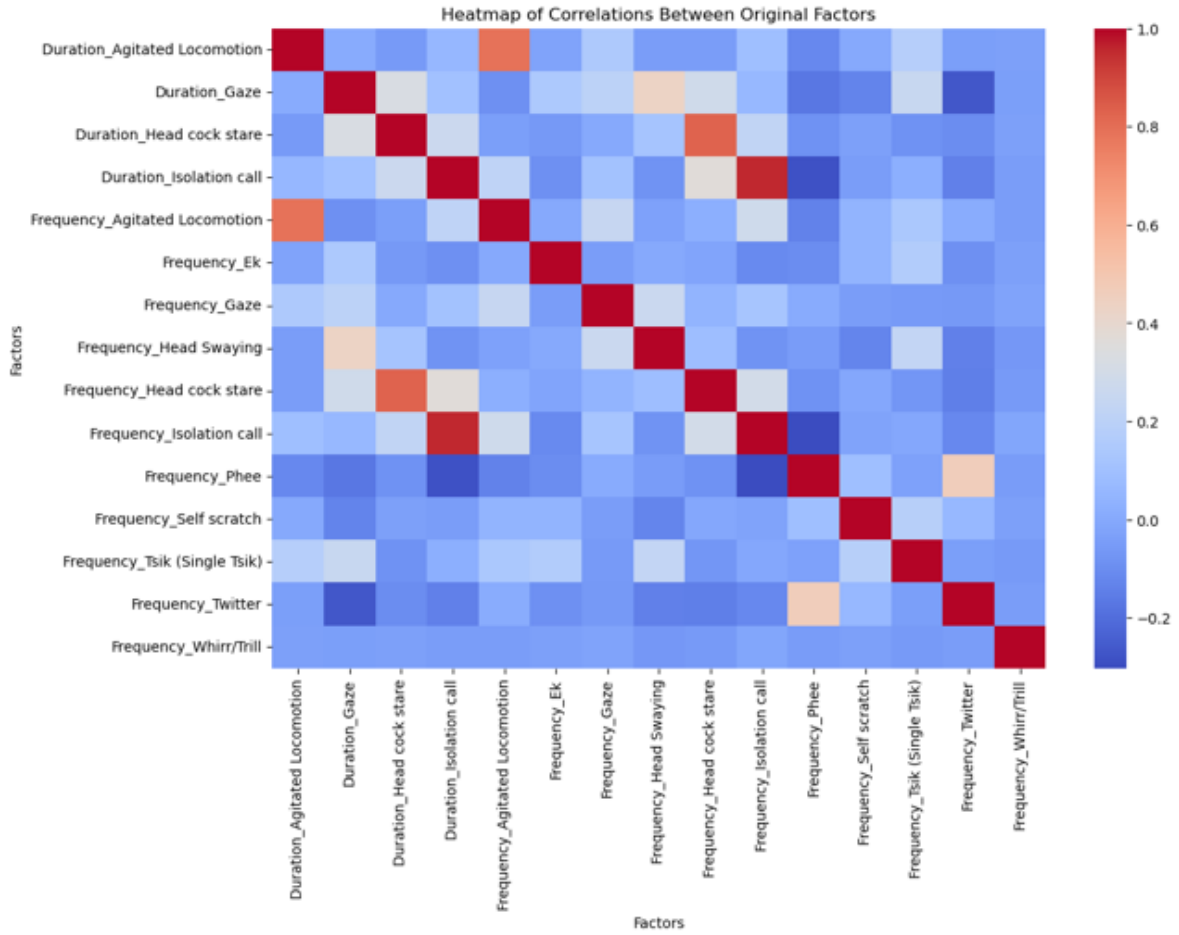


Figure 2. A heatmap of correlations between behaviors allows us to examine the correlation between head sway and gaze, as well as the correlation between Phee and twitter.

A heatmap was created to examine the correlations between behaviors.
Excluding the correlation between the frequency and duration of the same behavior, it was observed that the correlation between head swaying and gaze, as well as between Phee and Twitter, exceeded 0.4.

3.2 PC Loading & Cumulative Explained Variances

The Absence of significant correlations in the heatmap, Principal Component Analysis (PCA) was conducted to reduce dimensionality and identify underlying patterns in the data.

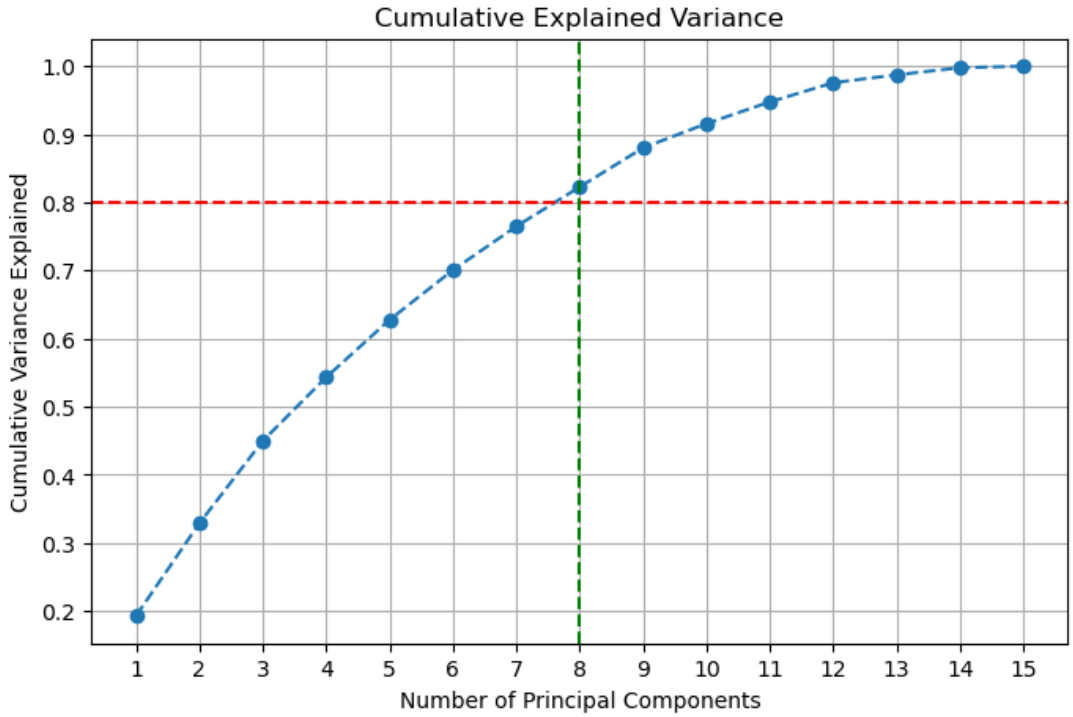


Figure 3. The number of principal components (PCs) with a cumulative explained variance of 80% is app roximately 8.

Among the 15 principal components (PCs) generated, approximately 8 PCs accounted for 80% of the cumulative explained variance (CEV) (Fig. 3). The explained variance (%) for each PC is presented in Table 1.

Table 1 The explained variance contributed by each principal component (PC)

Component	Explained Variance (%)
PC1	19.3
PC2	13.6
PC3	12.1
PC4	9.34
PC5	8.43
PC6	7.31
PC7	6.41
PC8	5.79
PC9	5.74
PC10	3.56
PC11	3.19
PC12	2.81
PC13	1.16
PC14	1.03
PC15	0.23

The individual behaviors loaded onto each PC were identified (Table 2), and titles were generated based on the behaviors with the highest proportion (Table 3).

Table 2. The behavioral loading for each principal component (PC) up to the 8th PC (Below |0.2| : weak contribution)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Agitated Locomotion duration (s)	0.13588	0.52014	0.22761	0.18597	-0.01832	-0.35571	-0.03067	-0.1169
Gaze duration (s)	0.26199	-0.28712	0.41114	0.04712	-0.00819	0.073108	0.04854	-0.11504
Head cock stare duration (s)	0.35853	-0.33182	-0.12086	0.31810	0.116634	-0.33791	0.03301	-0.02105
Isolation call duration (s)	0.46011	0.12944	-0.28279	-0.11252	0.050509	0.35721	0.00764	-0.10089
Agitated Locomotion Frequency	0.19907	0.55767	0.14116	0.19978	-0.01988	-0.22678	-0.02104	-0.02178
Ek Frequency	-0.02013	-0.04611	0.25600	-0.22156	0.386059	-0.12059	-0.29614	0.05547
Gaze Frequency	0.15048	0.10460	0.22521	0.22944	-0.44716	0.22950	0.09489	0.52655
Head Swaying Frequency	0.11072	-0.22002	0.46686	0.15397	-0.21179	0.255186	0.10213	-0.07218
Head cock stare Frequency	0.39276	-0.28616	-0.14057	0.29644	0.14728	-0.29600	0.01582	0.03899
Isolation call Frequency	0.44739	0.18239	-0.28430	-0.12214	0.01554	0.353472	0.033963	-0.08105
Phee Frequency	-0.28312	-0.05755	-0.11390	0.53403	0.007289	0.157242	0.085744	-0.10523
Self-scratch Frequency	-0.06701	0.10094	-0.03765	0.12524	0.56258	0.110161	0.37376	0.62028
Tsik (Single Tsik) Frequency	0.04509	0.09729	0.39949	0.03047	0.46516	0.282488	0.23213	-0.36309
Twitter Frequency	-0.23491	0.09723	-0.22797	0.46331	0.045593	0.200071	0.01139	-0.32444
Whirr/Trill Frequency	-0.03761	-0.00926	-0.06222	-0.25439	-0.17957	-0.27077	0.82795	-0.18823

Table 3. The dominant behavioral loading of each principal component (PC)

	Name	Top Contributing Variables	Reason
PC1	Isolation and Attention	Isolation call (duration, occurrences), Head cock stare (occurrences)	Captures behaviors indicating isolation and attentive staring.
PC2	Agitated Movement	Agitated locomotion (duration, occurrences), Head cock stare (duration)	Dominated by agitated locomotion, reflecting restless or active states.
PC3	Visual Engagement and Swaying	Head swaying (occurrences), Gaze (duration), Single Tsik (occurrences)	Highlights visual scanning and communicative swaying motions.
PC4	Vocal Communication	Phee (occurrences), Twitter (occurrences), Head cock stare (duration)	Represents verbal social interactions dominated by vocalizations.
PC5	Self-Grooming and Tsiks	Self-scratch (occurrences), Single Tsik (occurrences), Gaze (occurrences)	Reflects self-maintenance and subtle vocal communication.
PC6	Isolation and Agitation	Isolation call (duration, occurrences), Agitated locomotion (duration)	Emphasizes agitation in isolation contexts.
PC7	Trills and Scratching	Whirr/Trill (occurrences), Self-scratch (occurrences), Ek (occurrences)	Combines trilling vocalizations with self-focused actions.
PC8	Subtle Communication	Self-scratch (occurrences), Gaze (occurrences), Single Tsik (occurrences)	Indicates nuanced social interactions through self-directed and subtle signals.

3.3 3D Scatter Plot of 3 Highest PCs

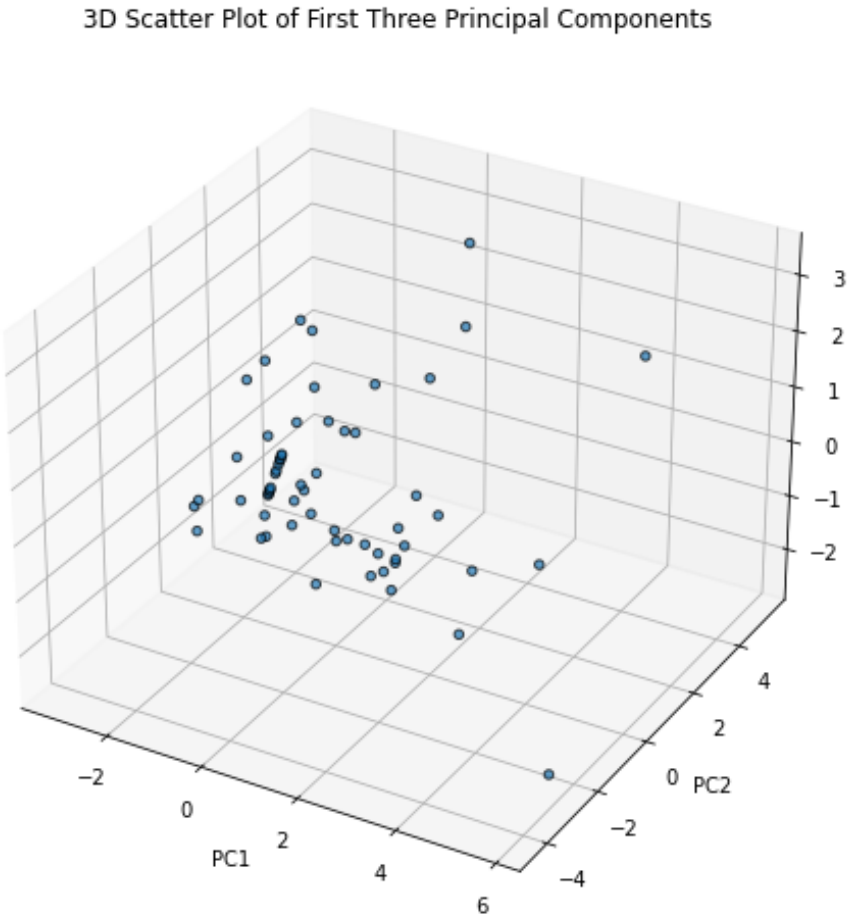


Figure 4. 3D Scatter Plot of Top 3 PCs

The top three PCs generated were visualized using a 3D scatter plot to examine clustering among the PCs. Since the data for the top three PCs did not form distinct clusters and were evenly distributed, it suggests that the composition of each PC may vary dynamically.
Given the lack of clear patterns in both the PCA results and the heatmap, statistical analysis of single variables across different conditions was conducted instead to investigate potential differences.

3.4 Single Variable Comparisons by Conditions

The Duration and Frequency of all behavioral patterns were measured, however, patterns that could not be statistically analyzed due to those small sample size were excluded from the comparisons.
Gaze duration showed a significant difference across conditions, with Friedman test result of $p = 0.00014$, indicating a clear distinction between them (Fig. 5). Based on this evidence, Dunn's test was conducted as a post-hoc test following the Friedman test. The Dunn's test result between Familiar_match and Unfamiliar_match was $p = 0.00664$, indicating a significant difference that marmosets spend more time for Gaze to Unfamiliar humans. Similarly, between Familiar_match and Unfamiliar_match was $p = 0.0000483$, showed also a significant difference (Tab. 4.).

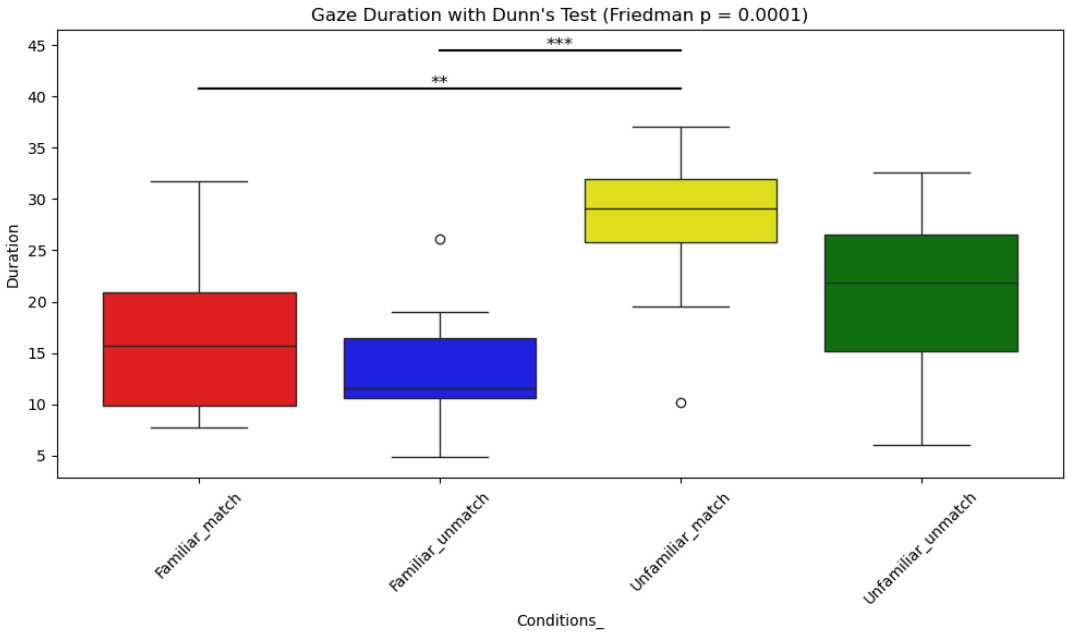


Figure 5. Gaze Duration significant differed between conditions (Friedman p = 0.000014) and 3 conditions represented significant differences within Dunn's test.

Table 4. Dunn's test result of Gaze Duration (post-hoc for Friedman test)

	Familiar_match	Familiar_unmatch	Unfamiliar_match	Unfamiliar_unmatch
Familiar_match	1.0	1.0	0.00664477085 1480999	1.0
Familiar_unmatch	1.0	1.0	4.82820286717 094e-05	0.0817083249928 3393
Unfamiliar_match	0.0066447708 51480999	4.82820286717 094e-05	1.0	0.2751167026708 557
Unfamiliar_unmatch	1.0	0.08170832499 283393	0.27511670267 08557	1.0

In the case of head swaying frequency, the Friedman test yielded a significant p-value of 0.0196. Therefore, Dunn's test was conducted to analyze the differences between conditions (Fig. 6). The Dunn's test p-value between Familiar_match and Unfamiliar_match was approximately 0.01706, indicating a significant difference as like the case of Gaze duration (Tab. 5).

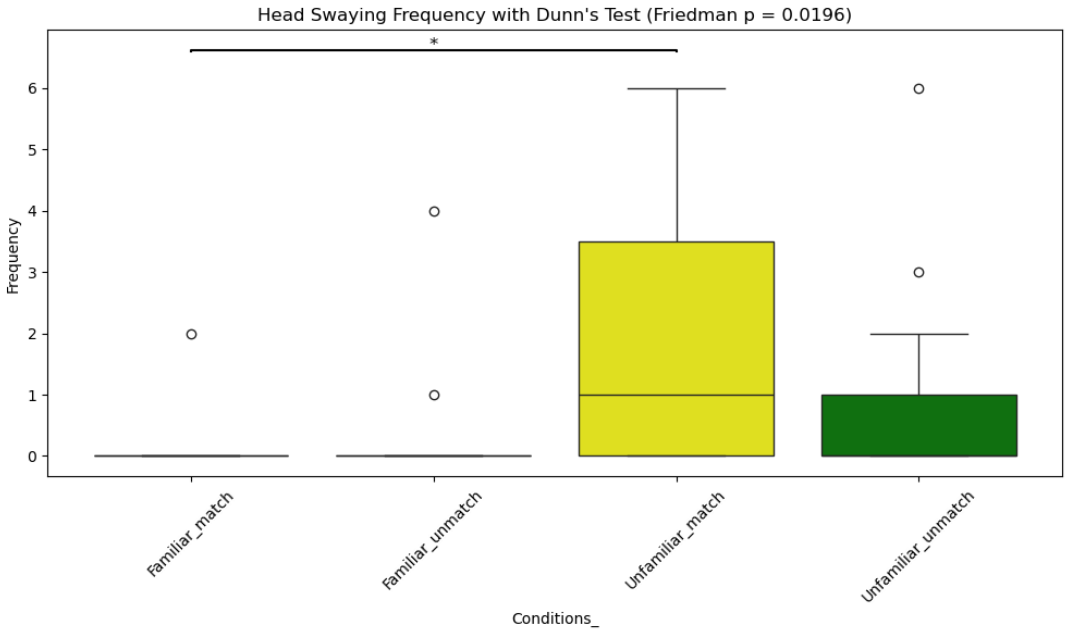


Figure 6. Head Swaying Frequency significant differed by Conditions (Friedman p = 0.01957) and Famili ar_match vs. Unfamiliar_match represented significant difference within Dunn's test.

Table 5. Dunn's test result of Head Swaying Frequency (post-hoc for Friedman test)

	Familiar_match	Familiar_unmatch	Unfamiliar_match	Unfamiliar_unmatch
Familiar_match	1.0	1.0	0.017060722755804822	0.7135552949118168
Familiar_unmatch	1.0	1.0	0.059124689516345 31	1.0
Unfamiliar_match	0.017060722755804822	0.059124689516345 31	1.0	0.9251588265864398
Unfamiliar_unmatch	0.7135552949118168	1.0	0.9251588265864398	1.0

The Friedman test for Tsik (Single Tsik) Frequency yielded a p-value of 0.0306, indicating a significant difference (Fig. 7). However, subsequent Dunn's test did not reveal any significant differences (Tab. 6).

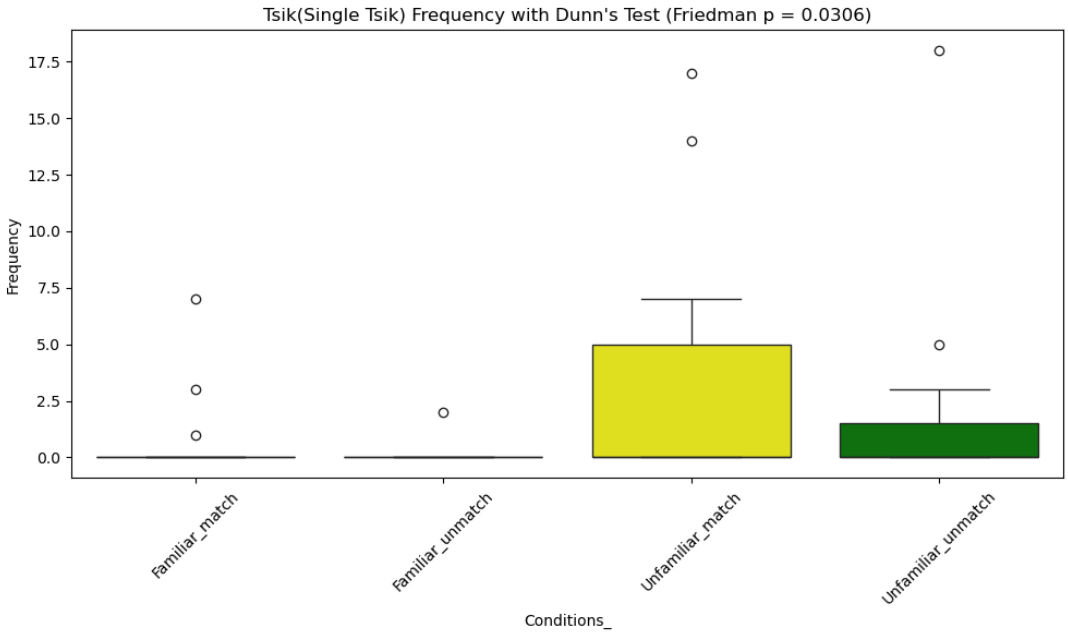


Figure 7. Tsik (Single Tsik) Frequency significant differed by Conditions (Friedman p = 0.0306), but Dunn's test did not represent significant differences.

Table 6. Dunn's test result of Tsik Frequency (post-hoc for Friedman test)

	Familiar_match	Familiar_unmatch	Unfamiliar_match	Unfamiliar_unmatch
Familiar_match	1.0	1.0	0.7510124321312807	1.0
Familiar_unmatch	1.0	1.0	0.11420876317247922	0.6189839973306132
Unfamiliar_match	0.7510124321312807	0.11420876317247922	1.0	1.0
Unfamiliar_unmatch	1.0	0.6189839973306132	1.0	1.0

For isolation calls, neither the frequency of occurrences nor duration showed significant differences across conditions. Notably, both familiar conditions exhibited nearly identical values (Fig. 8 & Fig. 9).

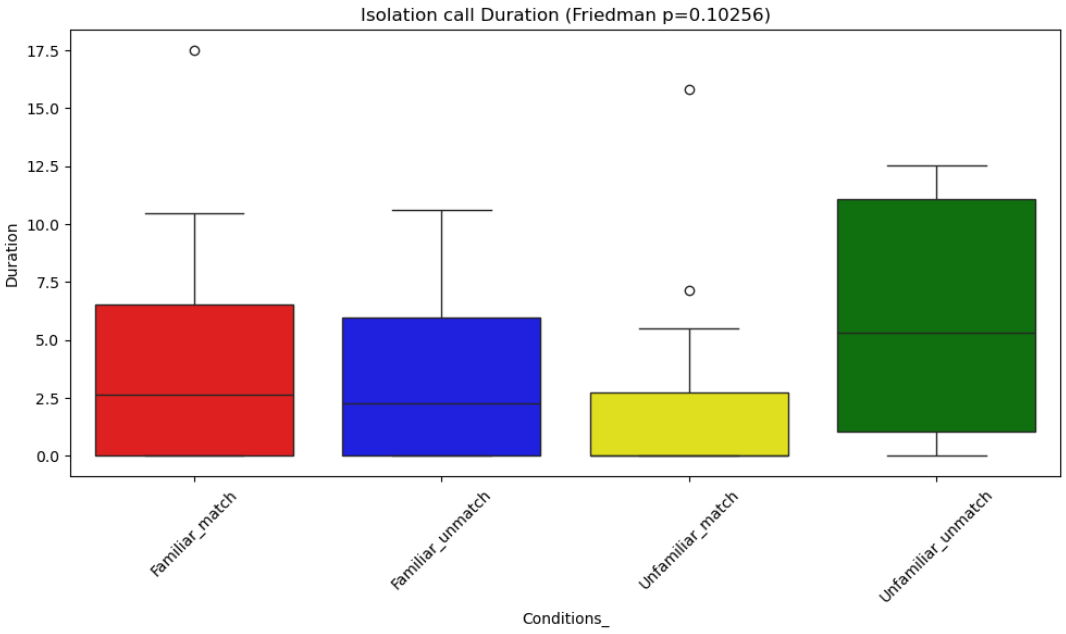


Figure 8. Isolation Call Duration represented no significant differences by Conditions (Friedman p = 0.10256)

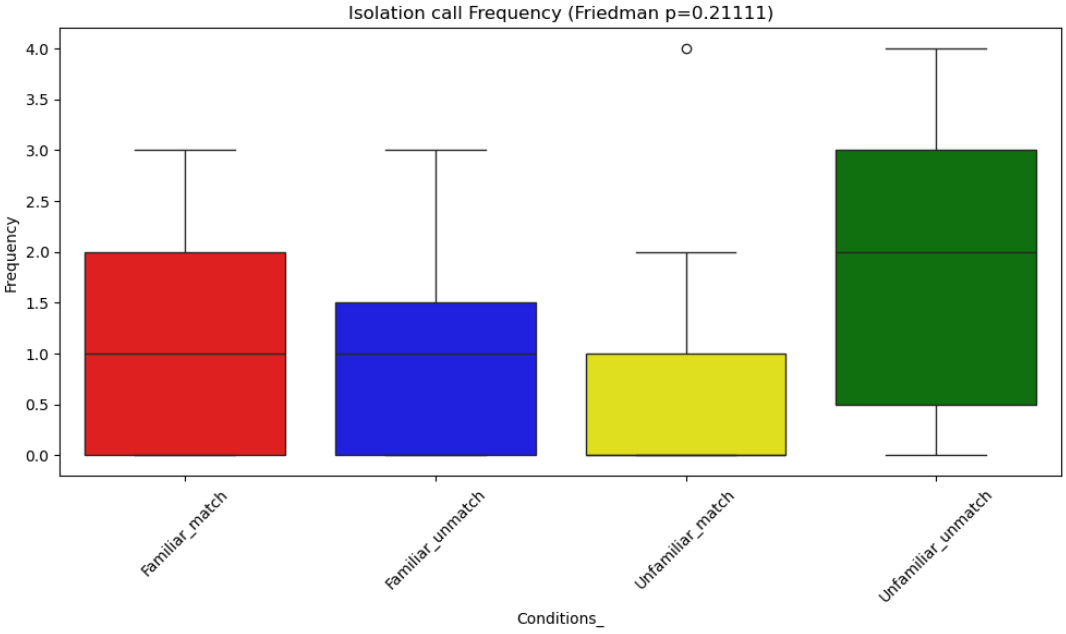


Figure 9. Isolation Call Frequency represented no significant differences by Conditions (Friedman p = 0.21111)

For Head Cock Stare (App. 3 & 4), Agitated Locomotion (App. 5 & 6), and Phee (App. 7), no significant differences were observed across conditions.

3.5 Gaze Duration, Head Swaying Comparison by Familiarity

Behaviors that represented significant differences in both Friedman and Dunn's test as post-hoc (Gaze Duration and Head Swaying Frequency), were further analyzed using Wilcoxon signed rank test, while not exhibiting within the same Familiarity conditions. This analysis aimed to determine whether there were behavioral differences influenced by Familiarity.

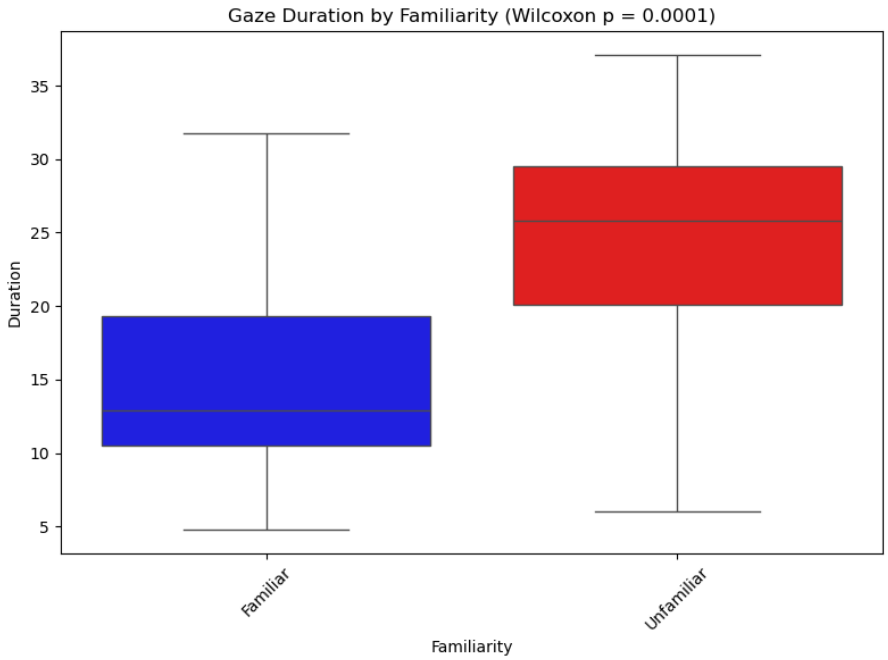


Figure 7. Gaze Duration represented significant difference by Familiarity (Wilcoxon signed rank test p value = 0.0001)

The Wilcoxon signed rank test for Gaze duration revealed a significant difference between the Familiar and Unfamiliar conditions (P = 0.0001), indicating that subjects exhibited shorter gaze durations toward familiar stimuli.

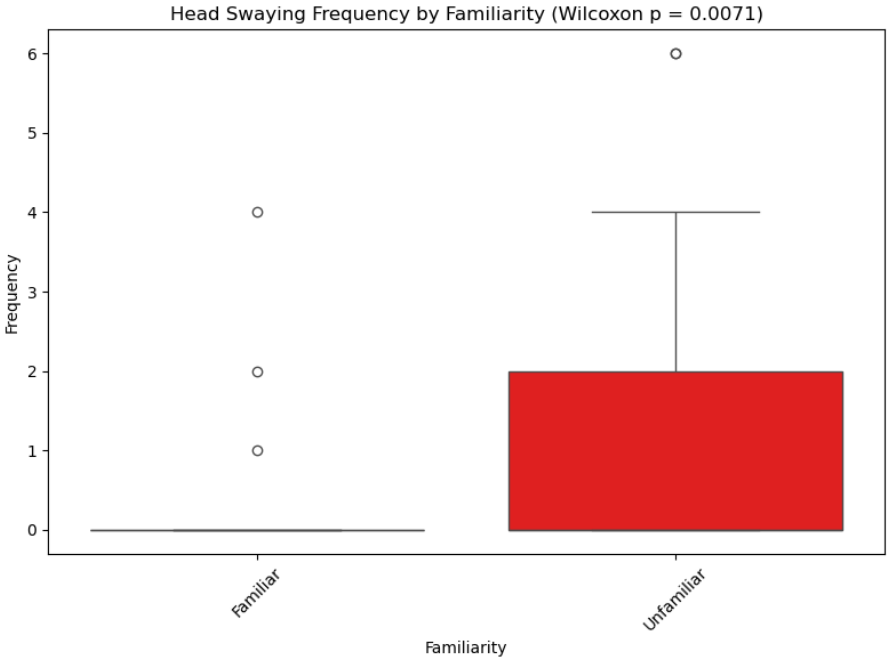


Figure 8. Head Swaying Frequency represented significant difference by Familiarity (Wilcoxon signed rank test p value = 0.0071)

For Head Swaying frequency, the Wilcoxon signed rank test yielded a p-value of 0.071, indicating a significant difference. This behavior occurred more frequently in the Unfamiliar conditions.

These results indicate that familiarity influences multi-modal recognition in marmosets

4. Discussion

4.1 Research Aims and Summary of Findings

The aim of this study was to examine whether marmosets utilize multimodal recognition to perceive humans in the same way they recognize conspecifics. Therefore, it was hypothesized that marmosets would show behavioral differences depending on the congruency between human faces and voices.

The results of this study indicate that congruency or incongruency between human facial and vocal cues did not significantly influence marmosets' behavior. Instead, differences in responses were primarily driven by familiarity with the human participants. In other words, marmosets prioritize familiarity over the congruency of audiovisual stimuli when recognizing individual human characteristics.

4.2 Comparison and Discussion with previous Studies

The results of this study showed some discrepancies compared to previous research.

According to Neiworth et al. (2007), when tamarins, another species of new world monkey, were shown 2D images of unfamiliar human faces, they exhibited higher attention levels compared to familiar human faces; however, such results were not obtained in this study. A possible reason for this difference might be that the previous study used standardized 2D images, possibly providing more consistent visual information. Based on this, it can be inferred that new world monkeys are capable of distinguishing individual humans by recognizing their unique facial features. But this study involved real human participants, which could have introduced greater variability in visual features, potentially making individual facial recognition more challenging for marmosets.

Kemp et al. (2013) reported that marmosets are sensitive to subtle facial expressions of conspecifics. Given this sensitivity, it is possible that unconscious facial expressions of human participants were not completely controlled in this study, potentially influencing the behavioral responses of the marmosets.

4.3 Limitations

In retrospect, this study might have faced some limitations. Firstly, external environmental stimuli like external personnel entering the experimental area and noise were not adequately controlled during the experiment. Secondly, the small sample size in this study limited the number of behavioral categories available for statistical analysis and likely increased the potential for statistical error. These limitations may have restricted the generalizability of the experimental findings.

4.4 Suggestions for Future Research and Final Remarks

Future studies should be designed to increase sample size and enhance experimental repetition to improve reliability. Additionally, it is recommended that future experiments control facial expressions more rigorously and minimize visual variability further.

The present study broadens our understanding of the marmosets' ability to recognize individual human characteristics and provides a foundation for further research into cognitive interactions between humans and non-human primates.

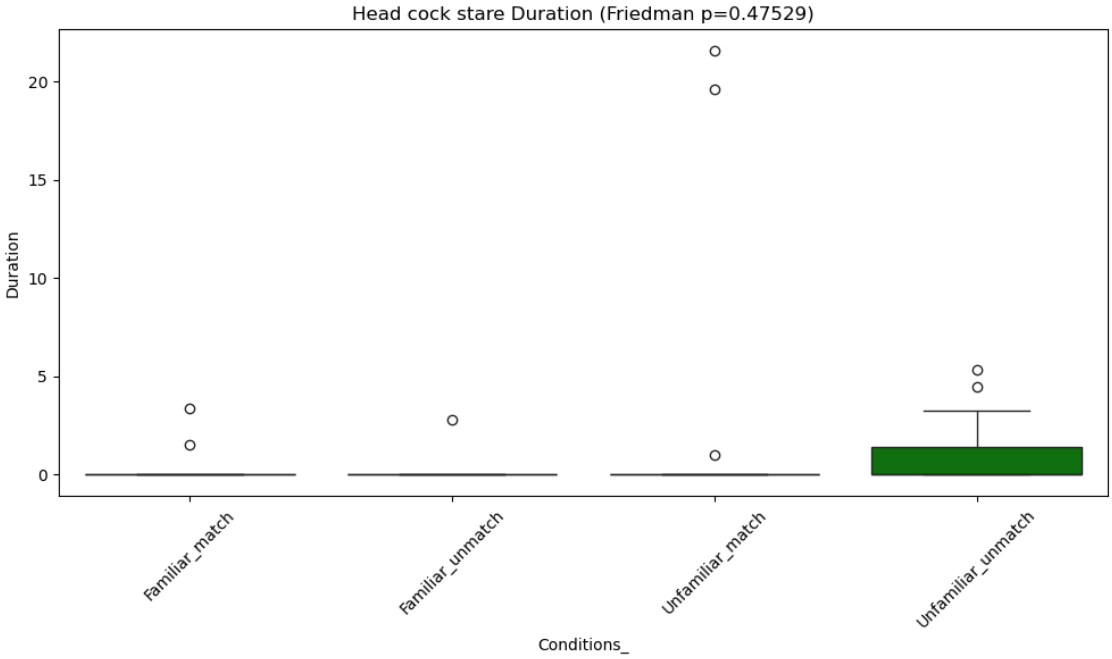
5. Appendix

Appendix 1. Information of Marmosets in Experiment

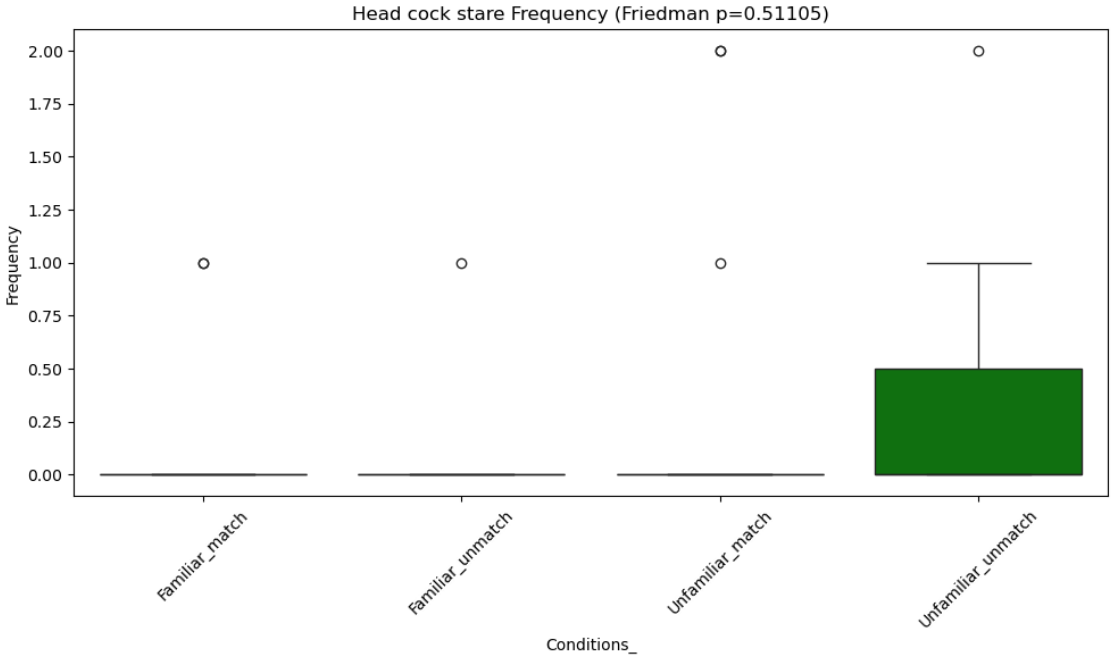
Lab	Group Nr.	Group Name	Individual Name	Code	Sex	Date of Birth	Age (Years)
1	1	Cleli Boys	Blinky Bill	BB	m	21.9.15	8
1	1		Bambi	BA	m	7.6.16	7
1	1		Maui	MU	m	9.10.17	6
1	2	Jack & Viana	Jack	JA	m	23.3.06	18
1	2		Vaiana	VA	f	9.10.17	6
1	3	Aurora	Lokri	LO	m	13.8.03	20
1	3		Aurora	AR	f	15.10.12	11
2	4	Feline	Feline	FE	f	7.6.16	7
2	4		Kobold	KO	m	11.4.05	19
2	5	V Group - Romans	Melvin	ME	m	19.7.14	9
2	5		Matilda	MA	f	19.7.14	9
3	6	Sprichtel	Nala	NA	f	28.1.16	8
3	6		Hr. Nilsson	MN	m	2.1.19	5
3	6		Kiara	KI	f	14.11.20	3
3	6		Smart	SM	m	4.11.09	14

Appendix 2. The Time Plan for Experiment

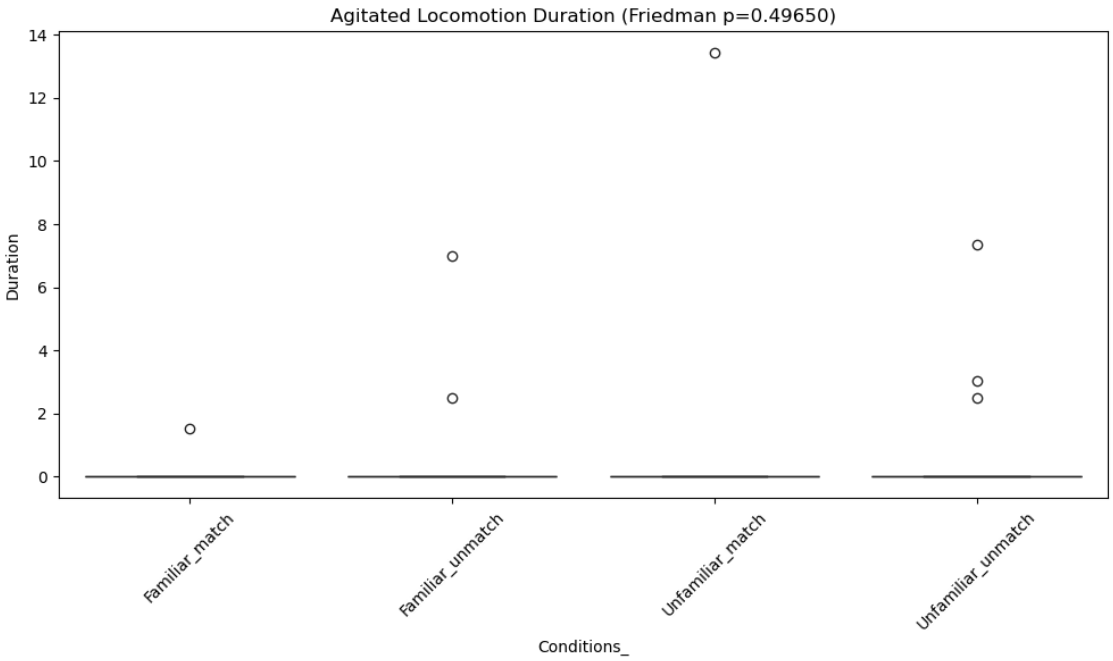
Day	Marmosets	Session
Day 1	5 marmosets	familiar match
Day 2	5 marmosets	familiar not-match
Day 3	5 marmosets	unfamiliar match
Day 4	5 marmosets	unfamiliar not-match
Day 5	5 marmosets	familiar match
Day 6	5 marmosets	familiar not-match
Day 7	5 marmosets	unfamiliar match
Day 8	5 marmosets	unfamiliar not-match
Day 9	5 marmosets	familiar match
Day 10	5 marmosets	familiar not-match
Day 11	5 marmosets	unfamiliar match
Day 12	5 marmosets	unfamiliar not-match



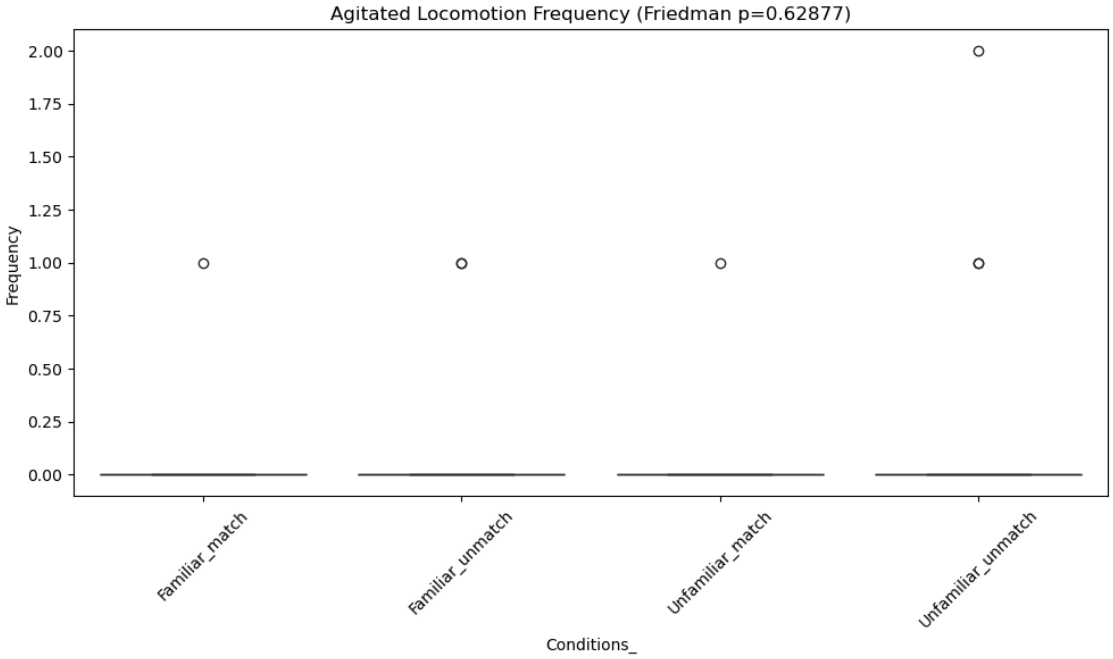
Appendix 3. Head Cock Stare Duration represented no significant differences by Conditions (Friedman p = 0.47529)



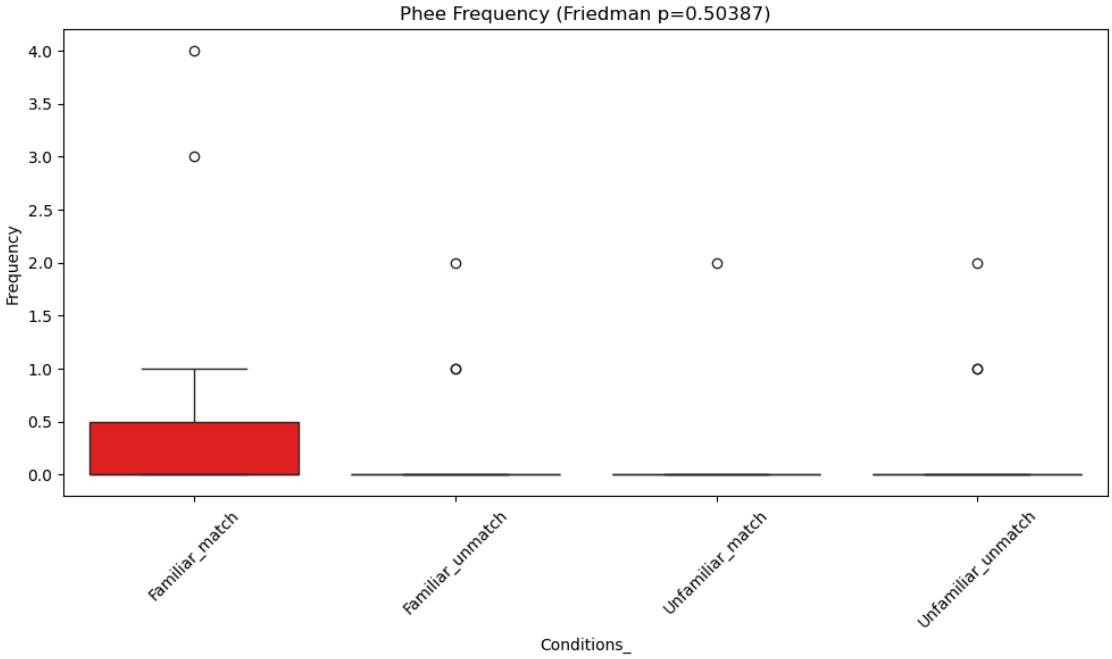
Appendix 4**. Head Cock Stare Frequency represented no significant differences by Conditions (Friedman p = 0.51105)**



Appendix 5. Agitated Locomotion Duration represented no significant differences by Conditions (Friedman p = 0.49650)



Appendix 6. Agitated Locomotion Frequency represented no significant differences by Conditions (Friedman p = 0.62877)



Appendix 7**. Phee Frequency represented no significant differences by Conditions (Friedman p = 0.50387)**

6. References

Arasaki, M. O., Campos e Silva, J. V., Faresin e Silva, C. E., & Vidotto-Magnoni, A. P. (2023). Interspecific association between two primate species in an urban park. *Semina: Ciências Biológicas e da Saúde, 44(2),* 137-144. https://doi.org/10.5433/1679-0367.2023v44n2p137

Chaka, K., Le-thanh, N., Flamary, R., & Belleudy, C. (2018). Performance comparison of the KNN and SVM classification algoriths in the emotion detection system EMOTICA. International Journal of Sensor Networks and Data Communications, 7, 1-9.

Campbell, D. L. M., & Hauber, M. E. (2010). Behavioural correlates of female zebra finch (Taeniopygia guttata) responses to multimodal species recognition cues. Ethology Ecology & Evolution, 22(2), 167–181. https://doi.org/10.1080/03949371003707885

Candiotti, A., Zuberbühler, K., & Lemasson, A. (2013). Voice discrimination in four primates. Behavioural Processes, 99(1), 67-72. https://doi.org/10.1016/j.beproc.2013.06.010

Da Cunha, R. G. T., & Byrne, R. W. (2009). The use of vocal communication in keeping the spatial cohesion of groups: Intentionality and specific functions. In P. A. Garber, A. Estrada, J. C. Bicca-Marques, E. W. Heymann, & K. B. Strier (Eds.), South American primates: Developments in primatology: Progress and prospects (pp. 341-363). Springer. https://doi.org/10.1007/978-0-387-78705-3_13

Diehl, M. M., & Romanski, L. M. (2013). Representation and integration of faces and vocalizations in the primate ventral prefrontal cortex. In H. Ogawa & M. Nakamura (Eds.), *Primate Origins of Human Cognition and Behavior* (pp. 45–69). Springer. https://doi.org/10.1007/978-1-4614-3585-3_3

Dubois, S., Rossion, B., Schiltz, C., Bodart, J. M., Michel, C., Bruyer, R., & Crommelinck, M. (1999). Effect of familiarity on the processing of human faces. NeuroImage, 9(3), 278-289. https://doi.org/10.1006/nimg.1998.0409

Forceville, C. (2012). Relevance Theory as model for analysing visual and multimodal communication. In E. Waßner & R. Alexander (Eds.), Relevance Theory and its applications (pp. 51-70). De Gruyter Mouton. DOI: 10.1515/9783110255492.51

Han, J., Zhang Z., Ren, Z., & Schuller, B. (2019). Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embedding. IEEE Transactions on Affective Computing, 12(3), 553-564.

Habbershon, H. M., Ahmed, S. Z., & Cohen, Y. E. (2013). Rhesus macaques recognize unique multimodal face-voice relations of familiar individuals and not of unfamiliar ones. Brain, Behavior and Evolution, DOI: 10.1159/000351203.

Reiss, D., McCowan, B., & Marino, L. (1997). Communicative and other cognitive characteristics of bottlenose dolphins. Trends in Cognitive Sciences, 1(5), 140–145. https://doi.org/10.1016/S1364-6613(97)01046-2

Kemp, C., & Kaplan, G. (2013). Facial expressions in common marmosets (Callithrix jacchus) and their use by conspecifics. Animal Cognition, 16(6), 773–788. DOI: 10.1007/s10071-013-0611-5

Ku-Peralta, W., & Sosa-López, J. (2025). Asymmetrical responses to multimodal signals in two Neotropical sister species of songbird. Animal Behaviour, 221. https://doi.org/10.1016/j.anbehav.2024.12.007

Neiworth, J. J., Hassett, J. M., & Sylvester, C. J. (2007). Face processing in humans and new world monkeys: The influence of experiential and ecological factors. *Animal Cognition, 10*(2), 125–134. https://doi.org/10.1007/s10071-006-0045-4

Miller, C. T., Freiwald, W. A., Leopold, D. A., Mitchell, J. F., Silva, A. C., & Wang, X. (2016). Marmosets: A neuroscientific model of human social behavior. Neuron, 90(3), 219-233. https://doi.org/10.1016/j.neuron.2016.03.018

Proops, L., McComb, K., & Reby, D. (2009). Cross-modal individual recognition in domestic horses (Equus caballus). Proceedings of the National Academy of Sciences, 106(3), 947-951. https://doi.org/10.1073/pnas.0809127105

Potvin, D. A. (2017). Coping with a changing soundscape: Avoidance, adjustments, and adaptations. Animal Cognition, 20(1), 9-18. https://doi.org/10.1007/s10071-016-0999-9

Salmi, R., Jones, C. E., & Carrigan, J. (2021). Who is there? Captive western gorillas distinguish human voices based on familiarity and nature of previous interactions. Animal Cognition, 1-3. https://doi.org/10.1007/s10071-021-01543-y

Schaeffer, D. J., Selvanayagam, J., Johnston, K. D., Menon, R. S., Freiwald, W. A., & Everling, S. (2020). Face selective patches in marmoset frontal cortex. Nature Communications, 11(1), Article 4856. https://doi.org/10.1038/s41467-020-18692-2

Seyfarth, R. M., & Cheney, D. L. (2003). Signalers and receivers in animal communication. Annual Review of Psychology, 54, 145–173. https://doi.org/10.1146/annurev.psych.54.101601.145121

Sliwa, J., Duhamel, J.-R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice–face identity matching by rhesus monkeys for familiar conspecifics and humans. Proceedings of the National Academy of Sciences, 108(4), 1735-1740. https://doi.org/10.1073/pnas.1008169108

Spadacenta, S., Dicke, P. W., & Thier, P. (2022). A prosocial function of head-gaze aversion and head-cocking in common marmosets. Primates, 63(5), 535–546 doi: 10.1007/s10329-022-00997-z

Tibbetts, E. A., & Dale, J. (2007). Individual recognition: it is good to be different. Trends in Ecology & Evolution, 22(10), 529–537. https://doi.org/10.1016/j.tree.2007.09.001