Abstract

Communication sounds are a subset of sounds that are characterized by their unique acoustic structures, biological relevance and specificity for social interaction. While there is difference between the complexity of human language and animal vocalizations, animals have been found to share similar vocal behaviors such as representative signaling and categorical perception with human. In addition, cortical organizations are similar between human and non-human primates. Therefore, it is believed that revealing the neural basis of primate vocalizations can give us insight on how human perceive speech.

Over the past decades, several experiments have been conducted to elucidate the computational principles of neural representations of species-specific vocalizations in the auditory cortex of non-human primates. However, no consistent picture has emerged. Whether vocalizations are encoded by highly specialized cortical neurons through a specific pathway is still debated. This is partly due to the lack of consideration of statistical structures of vocalizations and the behavioral relevance of their acoustic features.

In this thesis, I developed an automatic call classification and feature measurement algorithm to analyze a large number of marmoset vocalization recordings from a captive colony maintained by Dr. Xiaoqin Wang’s laboratory at Johns Hopkins University since 1996. These vocalizations were recorded from two groups of previous unfamiliar monkeys that were later placed in the same colony room (refer to hereafter as merging). We found that the call features of the current colony members (refer to hereafter as the current group) had similar distributions to those of the former members of this colony measured 18 years ago (refer to hereafter as the previous group). It is also found that after merging, the dynamics of convergence/divergence were observed in their vocalization acoustic features.
Besides acoustic features, we also examined the percentage of different call types in each session and the temporal sequences between and within the two groups of marmosets. Call type percentage was found to be different between the current and previous groups. While marmoset Trill calls were found to occur more frequently than Twitter calls in the previous group, Twitter calls occurred more often than Trill calls in the current group. The call type percentage of the current group remained relatively stable across seven months sampling period. It is found that temporal correlation between the call timing increased during the formation of a new social environment, and went back to the average a few months later. These findings suggest possible vocal plasticity and how temporal structures could vary under different social situations.

In the second part of my thesis, I used a synesthetic vocalization stimulus design to systematically probe the neural coding of vocalizations in marmoset auditory cortex with consideration of statistical structures. We found that A1 neurons showed relatively low selectivity to vocalizations and the selectivity increased in lateral belt neurons. Receptive field of these neurons also indicated vocalization features were processed with a distributed code in A1. How spatial location and background noise affected neuronal selectivity and tolerance of vocalizations was also explored.

Overall, this thesis focuses on a new approach to study monkey vocalization structures, their communication behaviors, and computational principles of vocalization processing in the auditory cortex. Preliminary data was presented in neural recordings.

Readers: Xiaoqin Wang (Advisor), Eric Young, Kechen Zhang
Acknowledgements

This thesis represents two years of hard work from experimental design, technique development, electrophysiological recordings, to data analysis. During this scientific journey, I got tremendous support from several groups of people. Here I would like to show my gratitude to those who have made the completion of this thesis possible.

First and foremost, I would like to thank my thesis advisor Xiaoqi n Wang. When I was considering a career transition from electrical engineering to neuroscience, he was willing to give me a chance to join his lab as a master student. I deeply appreciate his understanding and patience at the times when new approaches were tried and tested. He constantly urged me to pursue significant questions while allowing me the freedom to ask those interested me the most. I learned a lot from his insights on experimental design and data interpretation, as well as his integrity and cautiousness in every phase of scientific work. I am also grateful to my advisor for his generous support for many conferences and a summer school which broadened my horizon in neuroscience. I feel fortunate to work with a true scientist like him.

I am also indebted to my course instructor Eric Young. It is Models of the Neuron that first ignited my interest in computational neuroscience, and Auditory/Vestibular Brain that expanded my knowledge in auditory field. Besides intellectual guidance from class, I also learned a lot from his precise comments for every manuscript presented in journal club. Moreover, he was extremely helpful and patient in offering his advice for my experimental design and data analysis. I deeply appreciate his support and suggestions for my academic development as a graduate student.
I would also like to thank members of the Wang Lab for a wealth of intellectual, technical and mental support. Lei Feng has been a great mentor for her patience in teaching me how to do single-unit recordings. Michael Osmonski shared with me his perspectives in animal behavior. Yi Zhou and Christopher Dimattina were generous to share with me their programs, which became the foundations to develop my virtual vocalization synthesis tool. Jennifer Wang helped a lot in the physiological recordings in the last three months and shared with me her perspectives in spatial effect. Nate Satuyo and Shanequa Smith have been very helpful in assisting with animal care. I especially would like to thank Kai Yuen Lim, Lingyun Zhao, Darik Gamble, Xindong Song, Lixia Gao, Seth Koehler for their generosity in sharing with me their knowledge and advice on research career development.

I am also grateful to my summer research advisors, Krishna Shenoy and Teresa Meng. It is the summer research in my junior year that led me to a career transition from electrical engineering to neuroscience. I learned a lot from their passion in research and was inspired to be a great engineer and scientist like them. I deeply appreciate their continuing support for my academic development.

Last but not least, I would like to thank my family for their endless love. Their support and encouragements have been the source of my strength whenever there are frustrations. I deeply appreciate their understandings for the career path I chose and their unconditional care over the years in my growth.
List of Contents

1. Introduction ......................................................................................................................... 1
   1.1 Vocal behavior of non-human primates ................................................................. 1
   1.2 Neural coding strategies in auditory cortex ............................................................ 2
   1.3 Objectives of this essay .............................................................................................. 4

2. Automatic Vocalization Classification and Feature Analysis Algorithm ............... 6
   2.1 Introduction .................................................................................................................. 6
   2.2 Recording setup .......................................................................................................... 7
   2.3 Automatic call segmentation .................................................................................... 9
      2.3.1 Call detection .................................................................................................... 9
      2.3.2 Source localization ............................................................................................ 10
   2.4 Automatic call classification ..................................................................................... 13
      2.4.1 Training set selection ....................................................................................... 13
      2.4.2 Feature extraction ............................................................................................. 14
      2.4.3 Supervised model construction ......................................................................... 16
   2.5 Call feature measurement ......................................................................................... 19
      2.5.1 Broadband call feature measurement .............................................................. 20
      2.5.2 Narrowband call feature measurement ............................................................ 20
   2.6 Discussion .................................................................................................................... 21
      2.6.1 Feature selection in the algorithm .................................................................... 21
      2.6.2 Advantages and limitations of the algorithms ............................................... 22

3. Social Interaction Effect on Marmoset Conspecific Vocalization Structures .... 71
   3.1 Introduction .................................................................................................................. 71
   3.2 Results .......................................................................................................................... 73
      3.2.1 Call counts dynamics after formation of a new social environment ............ 73
      3.2.2 Call percentage dynamics after formation of a new social environment ...... 75
      3.2.3 Call power dynamics after formation of a new social environment ............ 75
      3.2.4 Comparison of feature distributions between two generations .................. 77
      3.2.5 Feature convergence and divergence between two groups ......................... 80
      3.2.6 Temporal sequence of call events between groups and within groups ....... 82
3.3 Discussion .............................................................................................................................................. 85
  3.3.1 Social interaction revealed by changes in vocalization structures .............................................. 85
  3.3.2 Implications in behavior relevance and motor control mechanisms ....................................................... 88

4. Social Interaction Effect on Marmoset Conspecific Vocalization Structures ........................................... 135
  4.1 Introduction ......................................................................................................................................... 135
  4.2 Theoretical hypothesis ...................................................................................................................... 137
  4.3 Stimuli design ..................................................................................................................................... 142
  4.4 Experimental setup ............................................................................................................................ 144
    4.4.1 Electrophysiology procedure ...................................................................................................... 144
    4.4.2 Acoustic stimuli ........................................................................................................................... 144
  4.5 Results .................................................................................................................................................. 145
    4.5.1 Neuronal response to call categories ............................................................................................. 146
    4.5.2 Spatial effect on neuronal selectivity to vocalizations ................................................................. 147
    4.5.3 Noise masking effect on neuronal selectivity to vocalizations .................................................... 147
    4.5.4 Neuronal response to call features ............................................................................................... 148
  4.6 Discussion ............................................................................................................................................ 149
    4.6.1 Optimal stimuli approach for future directions ........................................................................... 149

Reference ....................................................................................................................................................... 170
List of Tables

Table 2.1 Parameters for automated call segmentation algorithm ........................................... 24
Table 2.2 Feature vector for classification algorithm ............................................................... 25
Table 2.3 Parameters for feature extraction for classification algorithm ............................... 26
Table 2.4 Development set for classification training model ....................................................... 27
Table 3.1 Overall call counts detected .................................................................................. 89
Table 3.2 Measured feature distribution of Twitter ................................................................. 90
Table 3.3 Measured feature distribution of Phee ................................................................. 91
Table 3.4 Measured feature distribution of Trill ................................................................. 92
Table 3.5 Measured feature distribution of Trillphee ........................................................... 93
Table 3.6 Measured call type number across time ............................................................... 94
Table 3.7 Twitter feature distribution difference between sub-groups ............................... 95
Table 3.8 Phee feature distribution difference between sub-groups ................................... 96
Table 3.9 Trill feature distribution difference between sub-groups ....................................... 97
Table 3.10 Trillphee feature distribution difference between sub-groups ........................... 98
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Block diagram of Utah project</td>
<td>28</td>
</tr>
<tr>
<td>2.2</td>
<td>Schematic diagram of vocalization data collection</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>Automated call segmentation algorithm framework</td>
<td>30</td>
</tr>
<tr>
<td>2.4</td>
<td>Example of automated call segmentation algorithm</td>
<td>31</td>
</tr>
<tr>
<td>2.5</td>
<td>Automated call segmentation algorithm performance</td>
<td>33</td>
</tr>
<tr>
<td>2.6</td>
<td>Automated call classification algorithm framework</td>
<td>34</td>
</tr>
<tr>
<td>2.7</td>
<td>Example of twitter calls</td>
<td>35</td>
</tr>
<tr>
<td>2.8</td>
<td>Example of narrowband calls</td>
<td>38</td>
</tr>
<tr>
<td>2.9</td>
<td>Example of compound calls</td>
<td>41</td>
</tr>
<tr>
<td>2.10</td>
<td>Example of other call types and noise</td>
<td>44</td>
</tr>
<tr>
<td>2.11</td>
<td>Call type manual label agreement chart</td>
<td>47</td>
</tr>
<tr>
<td>2.12</td>
<td>Feature extraction for classification algorithm framework</td>
<td>48</td>
</tr>
<tr>
<td>2.13</td>
<td>Example of feature extraction for classification algorithm</td>
<td>49</td>
</tr>
<tr>
<td>2.14</td>
<td>KNN machine learning algorithm framework</td>
<td>50</td>
</tr>
<tr>
<td>2.15</td>
<td>KNN machine learning algorithm performance</td>
<td>51</td>
</tr>
<tr>
<td>2.16</td>
<td>Feature weights in optimal KNN model</td>
<td>53</td>
</tr>
<tr>
<td>2.17</td>
<td>Decision tree machine learning algorithm framework</td>
<td>54</td>
</tr>
<tr>
<td>2.18</td>
<td>Decision tree machine learning algorithm performance</td>
<td>55</td>
</tr>
<tr>
<td>2.19</td>
<td>Tree structure in optimal decision tree model</td>
<td>57</td>
</tr>
<tr>
<td>2.20</td>
<td>SVM machine learning algorithm framework</td>
<td>59</td>
</tr>
<tr>
<td>2.21</td>
<td>Parameter selection for optimal SVM model</td>
<td>60</td>
</tr>
<tr>
<td>2.22</td>
<td>SVM machine learning algorithm performance</td>
<td>61</td>
</tr>
<tr>
<td>2.23</td>
<td>Comparison of three machine learning algorithm performance</td>
<td>62</td>
</tr>
</tbody>
</table>
Figure 2.24 Feature measurement algorithm framework ............................................. 63
Figure 2.25 Example of Twitter feature measurement algorithm .................................. 64
Figure 2.26 Example of Phee feature measurement algorithm ........................................ 66
Figure 2.27 Example of Trill feature measurement algorithm ........................................... 68
Figure 2.28 Example of Trillphee feature measurement algorithm ................................. 70
Figure 3.1 Overall classified call type percentage .......................................................... 99
Figure 3.2 Overall call count and signal power dynamic across time ............................ 100
Figure 3.3 Call count dynamic across time of different call types .................................. 101
Figure 3.4 Call percentage dynamic across time of different call types ......................... 102
Figure 3.5 Signal power dynamic across time of different call types ............................ 103
Figure 3.6 Overall measured feature distribution of Twitter ......................................... 104
Figure 3.7 Overall measured feature distribution of Phee ............................................. 109
Figure 3.8 Overall measured feature distribution of Trill ............................................. 112
Figure 3.9 Overall measured feature distribution of Trillphee ....................................... 115
Figure 3.10 Dynamics of Twitter feature distribution between two groups .................... 119
Figure 3.11 Dynamics of Phee feature distribution between two groups ....................... 123
Figure 3.12 Dynamics of Trill feature distribution between two groups ......................... 125
Figure 3.13 Dynamics of Trillphee feature distribution between two group ................... 127
Figure 3.14 Inter-call-interval distribution across different periods ............................ 130
Figure 3.15 Temporal interaction across different periods ............................................ 131
Figure 3.16 Target proportion in temporal interaction ................................................. 132
Figure 3.17 Cross-correlation of temporal sequence .................................................. 133
Figure 3.18 Auto-correlation and time shift predictor of temporal sequence ................. 134
Figure 4.1 Possible neural coding strategies for vocalization ...................................... 150
Figure 4.2 Neural selectivity and tolerance to a vocalization feature model .................. 151
Figure 4.3 Neural discriminability for call categories model........................................ 152
Figure 4.4 Topographic map in right hemisphere ......................................................... 153
Figure 4.5 Examples of neural response to call categories............................................ 154
Figure 4.6 Examples of spatial effect on neural response to vocalizations ..................... 158
Figure 4.7 Examples of noise masking effect on neural response to vocalizations......... 162
Figure 4.8 Examples of neural response to features in synthesized calls..................... 163
Figure 4.9 Examples of temporal coding for synthesized vocalizations ....................... 168
Figure 4.10 Examples of U-shaped coding for synthesized vocalizations ..................... 169
CHAPTER 1:

Introduction

1.1 Vocal behavior of non-human primates

Vocal communication is a fundamental component of animal behavior. Each vocal repertoire has its specific spectra-temporal acoustic features, which are bound together into perceptually stable units to convey information (Hauser 2001). Recent studies have shown that primates use vocalizations in similar tasks to those in which human speech is employed. While specific words can represent particular environmental items in human speech, vervet monkeys were found to use three distinctly different alarm calls to indicate the presence of a python, leopard, or hawk (Seyfarth and Cheney 1982, Cheney and Seyfarth 1988). It is also found that tamarins and rhesus macaques use specific calls to under different situations (Gouzoules et al. 1995). Besides external referents, non-human primates are also found to utilize vocalizations to identify and contact other individuals (Snowdon and Hodun 1981, Schrader and Todt 1993, Kubertuhler 2000, Ghazanfar et al. 2002) or alarm predators (Seyfarth et al. 1980, Blumstein 1995, Manser et al. 2002).

Besides these simple representational functions observed in animal communicative behavior, vocal exchanges, contextual variation in semantic, and temporal sequential in the dialogue were also observed in non-human primates (Winter et al. 1966, Epple 1968, Snodon and Cleveland 1984, Maurus et al. 1985, Masataka and Biben 1987, Suiguiura 1993, Elowson and Snowdon 1994).
These findings suggest vocalizations of non-human primates exhibit similar utilities shared with human speech, including antiphonal calling paradigm, categorical perception, representative signaling, and vocal learning (Petersen 1982, Seyfarth and Cheney 1984, Moody et al. 1990, Seyfarth and Cheney 1992).

Although there is difference between the complexity of human language and animal vocalizations, vocal behaviors shared between human and non-human primates together with similar cortical organizations suggested some common brain mechanisms processing communication sounds. Therefore, revealing the neural basis of vocalizations is believed to provide insight into the brain mechanisms of human speech perception. Considering the vocal behavior, evolutionary closeness to human, and similar but flat cortical organizations, we decided to choose common marmosets as our animal model to study vocal behaviors and their neural representations.

1.2 Neural coding strategies in auditory cortex

How are acoustic signals transformed into the perception of sounds we experience? How does auditory cortex encode the spectral and temporal features carried in acoustic signals? Among all kinds of acoustic sounds, communication sounds are a subset of sounds that are characterized by their unique acoustic structures, biological relevance and specificity for social interaction. How vocalizations sounds are processed in cortex is of our interest. There are three possible candidates for communication sounds processing strategies in auditory cortex: dual stream hypothesis, a vocalization specific pathway, and distributed population coding.
While visual information is carried through two streams: ventral object and dorsal spatial processing streams, it is believed that similar dual streams, i.e. a rostral “what” pathway and a caudal “where” pathway, exist in the auditory system based on anatomical evidence (Romanski et al 1999, Kaas et al 1999, Tian et al 2001). Moreover, neurons in lateral belt regions are found to prefer vocalizations over pure tones or band-passed noise compared to strong responses to pure tones in primary auditory cortex (Tian et al 2001, Rauschecker et al 1995, Romanski and Averbeck 2009). Yet, whether spatial information is separated from auditory object information in cortex is still unknown.

Another hypothesis is that vocalization processing has a unique processing pathway in cortex different from other acoustic sounds processing strategies (Wang 2000) because of its biological importance. It can either follow the rostral pathway in dual streams model, or a topographic pathway that specifically processes vocalization signals. Such pathway assumes a hierarchical structure in auditory cortex that categorical perception and identity recognition are generated by highly specialized neurons (Winter and Funkenstein 1973, Mergoliash 1986, Cohen et al. 2007, Petkov et al. 2008, Lucia et al. 2010).

A third model states that vocalization is processed through distributed coding in cortex and there is no specific hierarchical pathway for vocalization processing. This model was supported by several studies that single units are often found to respond to more than one call or to multiple features (Newman and Wollberg 1973, Manley and Muller-Preuss 1978, Recanzone 2008). It is unknown whether neural sensitivity to specific spectral-temporal patterns can fully reflect the neural responses to certain call types. To test these hypothesis, a systematic experiment involving statistical structure of vocalizations and a complete examination of auditory cortex organization will be required.
1.3 Objectives of the essay

Perception and production of communication sounds are one of the most important behaviors in humans and highly-vocal non-human primates. Understanding how cortex processes vocalizations will give us insights on how speech perception was formed in human brain. The ultimate goal was to connect neural coding of conspecific vocalizations to behavioral relevance of vocal repertoires.

To investigate neural coding of vocalizations, we need to define the boundary of vocal repertoires first. While marmoset perceptual boundaries were unknown, the distributions of call acoustic features would be intuitive to start with. While previous studies (Jones 1993, Norcross and Newman 1993, Agamaite 1997) applied quantitative methods such as linear discriminant analysis to analyze vocalizations, call categories were defined based on visual inspection of spectrograms based on an early study (Epple 1968). However, it would be time-consuming to manually label call types with a lot of recording data. An algorithm to quantify massive recording data efficiently will be required for any further analysis.

While statistical variations define the boundary of vocal repertoires within individuals, distribution difference can also be found between individuals (Mitani and Brandt 1994). Such variations not only originate from the anatomical difference in vocal production structures and limitations in motor control mechanism (Hammerschmidt and Fischer 2008), but also come from the plasticity to adjust social distance (Boughman 1998). How call structures are influenced by social interactions and vary across different timescales will be needed to get a whole picture of their communication sounds.
With the knowledge of statistical structure of vocalizations, we would like to examine how cortical neurons encode these vocalizations. While vocalizations are thought to be encoded by highly specialized neurons (Winter and Funkenstein 1973, Mergoliash 1986, Cohen et al. 2007), single units are often found to respond to more than one call or to multiple features (Newman and Wollberg 1973, Recanzone 2008). A systematic tool to vary features in vocalizations would be required to find neural computation principles.

Two main objectives guide the experiments in this thesis: 1) to explore how social interaction affect vocal structures; 2) to explore a systematic stimulus design approach for studying neural response to vocalizations and find possible pathways in auditory cortex.

There are three questions that I have attempted to address:

1. How to automatically classify vocal repertoires and extract acoustic features from recording data with high signal-to-noise ratio (SNR)? (Chapter 2)

2. How are vocal acoustic features and temporal structures influenced by social interactions and varied across different timescales? (Chapter 3)

3. Is there a hierarchical pathway in cortex to process statistical spectral-temporal structures in vocalizations? (Chapter 4)
CHAPTER 2:

Automatic Vocalization Classification and Feature Analysis Algorithms

2.1 Introduction

A major goal in bioacoustics research is determining vocal repertoires for different species and correlating them with behavior (Cleveland and Snowdon 1982, Berg 1983, Sjare and Smith 1986, Remsen and Robinson 1990, Schrader and Todt 1993). In previous studies (Epple 1968, Stevenson and Poole 1976, Wang et al. 1995, Nagarajan et al. 2002), marmoset call types were categorized based on visual inspection of spectrograms produced from analog recordings. Although some studies (Norcross and Newman 1993, Jones 1993, Schrader and Todt 1993, Agamaite 1997) applied statistical methods such as discriminant analysis to support vocalization categories to reduce the bias from a human observer, manual classification is time-consuming while there are a large number of recordings. To handle massive data in our experiments, developing automatic marmoset call classification algorithm is needed to advance social interaction studies.

To automatically classify call types, there are two different methods: unsupervised clustering and supervised classification. While clustering maximizes similarity within the same group and differentiation between different groups (Wood et al. 2005), it requires a prior knowledge on group number. Therefore, a supervised classification would be more appropriate considering occurrence frequency of call types differ from session to session.
Different supervised classification systems were implemented in other animal studies. While artificial neural network (ANN) has been used to classify sea lion calls with spectral inputs (Campbell et al. 2002) and insect calls with temporal features (Chesmore 2001), linear spectrogram correlation incorporated with principle components analysis (PCA) has been used in various comparative studies (Potter et al. 1994, Mellinger and Clark 2000). Dynamic time warping (DTW) and hidden Markov model (HMM) with frequency-shifted Mel-frequency cepstral coefficients (MFCC) have been widely used in speech recognition and bioacoustics studies (Class et al. 1990, Anderson 1999, Novak et al. 2004).

Considering that MFCC was originally designed based on human speech properties, and PCA does not capture features well in data with high signal-to-noise ratio (SNR), we present an image-based method to extract marmoset vocalization features and compare the performance with three different classifiers. In addition to automatic classification, we also introduce a robust quantitative feature measurement method for four main simple call types. A block diagram of the full algorithm is shown in Figure 2.1. A GUI application which implements the algorithm is uploaded to the lab server.

2.2 Call Recording Setup

Two groups of marmosets previously unfamiliar with each other were housed together in a climate controlled colony room (80 F, 50% humidity). The colony group has forty marmosets living in the colony room. The Utah group has ten marmosets originally from Utah. Before housing these two groups of monkeys together in the colony, Utah group monkeys were housed in an isolated room with the same climate condition. Adults were housed individually in metal cages, and juveniles were housed together in a family cage.
Target animals were further chosen from both colony and Utah group for recordings. Eight adult marmosets (2 males and 6 females) were chosen in colony group while six adult marmosets (2 males and 4 females) and two breeding pairs were chosen in Utah group. We started recording five days after Utah monkeys were transported to Baltimore and housed in an isolation room. Recordings were typically conducted four hours per day from 11 am to 2 pm over a 7-month period. On a recording day, there were two sessions, and each consisted of four target animals. After a week, Utah monkeys were then housed in the same room with colony group monkeys, and each session consisted of two target animals from colony group and two from Utah group. Different groups of target animals were rotated between sessions to ensure each target has same amount of recording time within a period. On the other hand, recordings obtained from 18 years ago were from 8 adult marmosets (4 males and 4 females) over a 15-month period.

Recordings were taken with four directional microphones (Sennheiser ME66) which were placed 15 cm in front of the target animals. Microphone output signals were amplified with two dual microphone amplifiers (Symetrix 302) and recorded via a 4-channel data acquisition device (NI 9237) sampling at 50 kHz. All recordings were conducted with marmosets located in their home cages, while several recordings obtained from 18 years ago were performed on marmosets temporarily moved to a plastic mesh cage encapsulated in Sonex foam. Figure 2.2a shows the recording set up in one session in one day after two monkey groups were housed in the same colony room. All the Utah monkeys were spatially located near each other in the room, so that it was more robust to identify speaker source between two groups. Figure 2.2b shows an example of 4-channel recording in one session.
2.3 Automatic Call Segmentation

After data collection, how to quickly scan through a long recording and detect the segments that contain marmoset vocalizations became the first challenge. Figure 2.3 shows the overall automatic call segmentation algorithm framework. Recordings in each channel were high-pass filtered to eliminate low-frequency noise below 3 kHz, and notch filtered to reduce computer noise during recording. After filtering, possible call events were detected based on intensity and duration criteria, which were described in Section 2.3.1. After detecting possible call events, speaker source was further identified between four channels, which was described in Section 2.3.2. The last step was to group compound calls and to add buffer time to avoid call ends being cut out, as described in Section 2.3.3.

2.3.1 Call detection

A time window was selected and analyzed for each channel at a time, and would slide over the whole recordings with a step size. Figure 2.4a shows an example of two-channel recording. The intensity of data in the analyzed channel was then calculated and shown as a solid contour line in the top panel of Figure 2.4b. An intensity threshold was defined as follows, and shown as a gray dash line in the top panel of Figure 2.4b.

\[ I_\mu = \frac{\int_{t_0}^{t_0 + t_{\text{win}}} I(t) \, dt}{t_{\text{win}}} \]

\[ I_\sigma = \sqrt{\int_{t_0}^{t_0 + t_{\text{win}}} \left( t - I_\mu \right)^2 I(t) \, dt} \]

\[ I_{\text{threshold}} = I_\mu + thr \cdot I_\sigma \]
A duration threshold was further applied for segments that passed intensity threshold, as shown in the bottom panel of Figure 2.4b, where three possible events between red dash lines were then segmented.

2.3.2 Source localization

As Figure 2.4b shows, same calls might be picked up by more than one microphone if the call intensity was large enough. Besides, the calls picked up by one microphone did not necessarily come from the target animal in front of that microphone due to a crowded animal density in the recording environment. To localize the signal source, an algorithm using time differences of arrival (TDOA) and maximum likelihood estimator (MLE) was applied to identify the speaker source and assign the call to nearest channel.

There are 4 microphones distributed in three-dimensional space regarding where the target animals are. Let the observations at microphone I be

\[ u_i(k) = s(k - T_i) + n_i(k), \quad i = 1, 2, 3, 4 \]

where \( s(k) \) is the signal reached to the microphone from the source, \( T \) is the time delay associated with receiver, and \( n(k) \) is the noise assumed to be zero mean stationary Gaussian random process, which makes the noise covariance same as time delay covariance.

Time delay of arrivals between three other microphones and one chosen microphone can be computed with the delay estimation error \( n \) as

\[ d_{i,4} = T_i - T_4, \quad i = 1, 2, 3 \]

\[ d_{i,4} = d_{i,4}^0 + n_{i,4}, \quad i = 1, 2, 3 \]
Let the speaker source at unknown location \((x, y, z)\) and the microphones at locations \((x_i, y_i, z_i)\). The squared distance between the source and microphone \(i\) is computed as

\[
r_i^2 = (x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2, \quad i = 1, 2, 3, 4
\]

Distance difference between other microphones and one chosen microphone can be further computed with sound propagation speed \(c\)

\[
r_{i,4} = c \cdot d_{i,4} = r_i - r_4, \quad i = 1, 2, 3
\]

The source location would be the intersection of these three-dimensional hyperbolic surfaces. A simplified schematic diagram projected on a two-dimensional plane with three microphones is shown on Figure 2.4c. In the presence of noise, these equations will not meet at the same point. To find the answer that fits these nonlinear equations best, there are several methods such as Taylor-series expansion linearization and iteration (Foy 1976), minimizing least-square error by spherical-interpolation (Smith and Abel 1987). However, the convergence is not guaranteed in the first method, and the weighting matrix in second method is difficult to determine.

Here we applied a non-iterative realization of maximum likelihood estimator method (Chan and Ho 1994) to solve these nonlinear equations. Let the proper answer be \((x, y, z)\) and the distance to the chosen microphone is \(r\)

\[
\mathbf{v} = [x, y, z, r]^T
\]

To estimate \(\mathbf{v}\), we first assume that positions and the distance are independent with each other, and then they can be solved by least-square error (LS). The second step applies the known relationship between positions and the distance with another least-square error.
This two-step process is a maximum likelihood estimator approximation, which can be written as

\[ \mathbf{v} = (\mathbf{G}^T \psi^{-1} \mathbf{G}^T)^{-1} \mathbf{G}^T \psi^{-1} \mathbf{h} \]

where

\[ \psi = \mathbf{h} - \mathbf{G} \mathbf{v}^0 \]

\[ \mathbf{G} = \begin{bmatrix} x_{1,4} & y_{1,4} & z_{1,4} & r_{1,4} \\ x_{2,4} & y_{2,4} & z_{2,4} & r_{2,4} \\ x_{3,4} & y_{3,4} & z_{3,4} & r_{3,4} \end{bmatrix} \]

\[ \mathbf{h} = \frac{1}{2} \begin{bmatrix} r_{1,4}^2 - K_1 + K_4 \\ r_{2,4}^2 - K_2 + K_4 \\ r_{3,4}^2 - K_3 + K_4 \end{bmatrix} \]

\[ K_i = x_i^2 + y_i^2 + z_i^2, \quad i = 1, 2, 3, 4 \]

By applying this algorithm, we can assign the call signal to the nearest channel so that each call signal only appears on one channel without repeating. As shown on Figure 2.4f, events were assigned to different channels after source localization.

After each call signal was assigned to the target channel properly, we grouped call signals whose inter-call-interval on one channel was less than 500 msec as a compound call. Finally, we added buffer time before and after the segmented events to avoid the call ends being cut out. A compound call example was shown in Figure 2.4g. The evaluation of the algorithm performance with different threshold parameters was shown in Figure 2.5. Based on the performance with highest F-score, we chose parameters as listed in Table 2.1.
2.4 Automatic Call Classification

After possible call events were segmented, we split our data into two sets: prediction set and development set. Development set was further subdivided into a training set and a validation set for cross-validation in training, as shown in Figure 2.6a. We then applied a supervised classification algorithm that involved feature extraction and a machine learning classifier. The overall framework is shown in Figure 2.6b.

2.4.1 Training set selection

We first randomly selected 10 sessions out of 150 sessions that covered periods before housing two monkey groups in a same room, the date when new social environment was formed, and periods after several months.

We then categorized the segmented events in the selected sessions into seven groups based on classification scheme described in previous studies (Epple 1968, Agamaite 1997). The examples in Twitter group were shown in Figure 2.7. Twitter calls have different number of phrases, different number of linear fractions, different slopes of linear fractions, and different variations in the first phrase. Examples in Phee, Trill, and Trillphee groups were shown in Figure 2.8. Starting frequency and ending frequency in these narrowband calls have different variations. Transition between three call types is continuous. Examples of Compound group were shown in Figure 2.9. Other call types and overlapped calls were put in Other group, as shown in Figure 2.10. Events in Noise group included any sounds that were not marmoset vocalizations. The manual categorization for training set were labeled and agreed between three human observers, as shown in Figure 2.11.
2.4.2 Feature extraction

After we labeled segments for development set, we extracted the features based on their spectrograms. The algorithm framework is shown in Figure 2.12. Spectrograms were obtained with a 75% overlapping Hamming window by doing 512-point FFTs, resulting in a 2.6-msec temporal resolution and 97-Hz frequency resolution. The analyzing parts were further limited between 4 kHz and 16 kHz because the spectrums of the main call types were mostly within this range.

After the spectrograms were obtained, characteristic time-frequency traces were then extracted with the following steps. Noise in the spectrograms either introduced by human activities in the colony or produced by marmosets striking the metal cages were removed by a Butterworth notch filter and a local-smoothing Gaussian bilateral filter with kernel coefficients computed from spatial closeness and the intensity similarity within the defined area (Tomasi and Manduchi 1998, Mallawaarachchi 2007).

To do the smoothing, we defined the neighborhood center \( x \) and a nearby pixel within the local neighborhood \( y \) first. A bilateral filter is then written as

\[
h(x) = \frac{1}{k(x)} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y) \cdot c(x, y) \cdot s(f(x), f(y)) dy
\]

\[
k(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(x, y) \cdot s(f(x), f(y)) dy
\]

\[
c(x, y) = e^{-\frac{1}{2}(d(x,y))^2}, \quad d(x, y) = \|x - y\|
\]

\[
s(x, y) = e^{-\frac{1}{2}(\frac{\Delta(f(x), f(y))}{\sigma_r})^2}, \quad \Delta(f(x), f(y)) = \|f(x) - f(y)\|
\]
where \( c(x, y) \) defines the spatial closeness between \( x \) and \( y \), and \( s(f(x), f(y)) \) measures the intensity similarity between \( f(x) \) and \( f(y) \), \( d(x, y) \) is Euclidean distance between \( x \) and \( y \), \( \Delta(f(x), f(y)) \) is the intensity difference between \( f(x) \) and \( f(y) \), \( k(x) \) is the normalization factor, \( \sigma_d \) is a fraction of the window size, and \( \sigma_r \) is a fraction of the range of intensity levels.

After smoothing, we normalized the pixel intensity to a gray image (0-255), and then set the intensity threshold with a fixed pixel intensity difference. All the spectrograms were assumed to have similar signal-to-noise ratio (SNR), so the time-frequency traces could be mostly extracted. After a spectrograms was applied with an intensity threshold function \( T \), the spectrogram \( I(x,y) \) became a binary images \( J(x,y) \).

\[
J(x,y) = \begin{cases} 
1 & \text{if } I(x,y) \geq T \\
0 & \text{else} 
\end{cases}
\]

To compensate the discontinuous energy in traces, morphological operation was applied to connect parts that were originally connected and remove the transient noise. Two morphological operators were utilized with several iterations (Adams 1993, Soille 1999). Opening operation removes any isolated groups that have fewer than 10 pixels on the image. Closing operation is a dilation followed by an erosion, using the same structuring element (SE) for both processes, which fills the small gaps between groups on the binary image.

As a final step in feature extraction, harmonic structures were removed based on an assumption that the frequency difference between maximum fundamental frequencies and the minimum harmonic frequencies within a same time interval was larger than 1 kHz. The examples of how feature extraction works in each step were shown in Figure 2.13. The parameters used in this algorithm were listed in Table 2.2.
Features were then measured. A feature vector was generated for each spectrogram. Table 2.3 shows two different components of a feature vector. Four different development sets for classification training were created using two different feature vectors and different number of calls in each call type. Note that the features measured here were only used for classification because notch filtering and morphological operation would bias the feature distributions. The algorithm for future feature measurement is described in Section 2.5.

2.4.3 Supervised classifier construction

After extracting the feature vectors, the machine learning model was constructed with these vectors and the labels. Here we compared three different supervised learning methods including K-nearest neighbors (KNN), decision tree, and support vector machine (SVM), and chose the one with best performance to classify all the data set.

We first randomly split our labeled data from 10 sessions into two sets: 70% was training set and the rest 30% was validation set. After the model was constructed with the training set, it would be used to classify the validation set, which gave a cross-validated accuracy. This process repeated 10 times, and two sets were randomly chosen at each time. Such 10-fold cross-validation could avoid over-fitting and was used in all three models.

K-nearest neighbors (KNN) method assigned data to a class based on the majority vote of K nearest neighbors in feature space (Cover and Hart 1967). Distance calculated in high-dimensional feature space was normalized Euclidean distance. Example was shown in Figure 2.14b, where unlabeled data (gray) was assigned to the estimated class (red).
The KNN algorithm performance across different K number was compared between four different development sets defined in Table 2.4. Figure 2.15a shows the classification accuracy for both training set and validation set. Figure 2.15b shows the cross-validated accuracy. Development set 2 and set 3 reached higher accuracy than the other two sets in general. K number did not have much effect on cross-validated accuracy. Classification sensitivity and precision for four main call types (Twitter, Phee, Trill, and Trillphee) were shown in Figure 2.15c to Figure 2.15j. To maintain high sensitivity and high precision, we found that 6-nearest neighbors with development set 2 would be a more proper model among all the KNN algorithms. Feature ranks and weights were shown in Figure 2.16.

Another method using a non-parametric decision tree (Coppersmith et al. 1999) was described as follows. A tree was formed by a set of rules based on variables in the modeling data set. Rules were selected to get best split that maximized posterior probability of the observations in each branch. Each observation would fall into one and only one terminal node, and each node was uniquely defined by a set of binary rules. Trees could be pruned to reduce the node number while accuracy was maintained at a certain level. Example was shown in Figure 2.17b, where feature space was projected onto a two-dimensional plane. Different colors denote different classes. Rules were shown beside the nodes.

The decision tree algorithm performance across different minLeaf number (minimum number of observations per tree branch) was compared between four different development sets defined in Table 2.4. Figure 2.18 shows the tree node number, the cross-validated accuracy, sensitivity and precision for four call types. To reduce node number and maintain high sensitivity and precision for call types, development set 1 with tree (minLeaf = 5) was chosen as a proper model. The complete tree structure was shown in Figure 2.19.
The last method using support vector machines (SVM) was described in the following. Separation between two classes was achieved by a hyperplane that had the largest distance to the nearest training data point (margin). While one-against-all strategy was widely used in multi-class classification (Hastie and Tibshirani 1998), the pairwise one-against-one strategy was shown to be more stable (Wu et al. 2004). Therefore, we applied C-supported vector classification method (Cortes and Vapnik 1995, Chang and Lin 2011) with pairwise one-against-one strategy for multi-class classification, as shown in Figure 2.20b.

Given training feature vectors \( \mathbf{x} \) with length \( m \) and a label vector \( \mathbf{y} \) such that \( y_i \) belongs to one of the seven classes, C-SVC solves the optimization problem below.

\[
\min_{\omega, b, \xi} \frac{1}{2} \omega^T \omega + C \cdot \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i(\omega^T + \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, m \)

where \( \phi(x) \) maps \( x \) into a high-dimensional space and \( C \) is the hyper-parameter. Due to the high dimensionality of the vector \( \mathbf{w} \), we solved the following problem with LIBSVM tool.

\[
\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - \mathbf{v}^T \alpha
\]

\( Q_{i,j} = y_i y_j \cdot K(x_i, x_j) = y_i y_j \cdot \phi(x_i)^T \phi(x_j) = y_i y_j \cdot e^{-\gamma |u-v|^2} \)

subject to \( \mathbf{y}^T \alpha = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, m \)

where \( \mathbf{v} = [1, \ldots, 1]^T \) is the vector of all ones, \( Q \) is a positive semi-definite matrix, and \( K \) is the radial basis kernel function.
After the above equation was solved, the optimal $\omega$ satisfied

$$\omega = \sum_{i=1}^{m} y_i \alpha_i \phi(x_i)$$

To select proper parameters $\gamma$ for kernel function and the hyper-parameter C, a grid search was applied with four different development sets, as shown in Figure 2.21. The summit of the contour suggested the parameter pair for maximal cross-validated accuracy. Later on, the model with selected parameter pair was used to predict the development sets. The classification accuracy with default parameter pairs and selected parameter pairs was shown in Figure 2.22a. Selectivity and precision for four main call types were also shown in Figure 2.22b and Figure 2.22c respectively. We could find that a proper SVM model would be using development set 1 with parameter pairs ($\log_2 C = 7, \log_2 \gamma = -5$).

Finally, algorithm performance between three selected machine learning models was compared in Figure 2.23. SVM obtained the higher sensitivity and precision over the other two methods; hence, we decided to use SVM with selected parameter pairs for the rest of the marmoset vocalization classification.

2.5 Call Feature Measurement

To reduce bias in feature distributions of four main call types, a more sophisticated feature measurement algorithm was developed according to the classified call type. While the major framework between broadband and narrowband calls were similar, there were difference in the noise removal process, as shown in Figure 2.24.
2.5.1 Broadband call feature measurement

For Twitter call type, data was applied with a high-pass filter at 3 kHz first. Then the spectrogram intensity edge was detected by deducting the background intensity. We then removed the harmonic structures using same assumption described in Section 2.4.2. After that, we extracted the trace with maximal pixel intensity in each temporal column. PC noise was further removed. Note that we did not apply a notch filter here. Instead, we assumed that the continuous horizontal components in Twitter calls were less than a certain time interval, so that we could avoid information loss in that frequency band.

We further assumed that the frequency drop between phrases were large, and that the frequency rise in middle phrases before knee point had at least 1 kHz. After defining the periods for each phrase, peaks and troughs were searched with spectrogram intensity edge detection, and the maximal time-frequency trace between peak in current phrase and trough in next phrase was discarded to remove the echo effect.

The final step was to smooth the time-frequency trace using the trajectory prediction with the position of each pixel on the spectrogram image. We then marked the important features such as peaks (maximal frequency in each phrase), troughs (minimum frequency in each phrase), knees (point where slope changes in each phrase), and dominant (point where maximal energy occurs in each phrase). Examples of Twitter feature measurement algorithm was shown in Figure 2.25.

2.5.2 Narrowband call feature measurement

For narrowband calls, a different strategy for noise removal was applied to reduce the information loss since the slope of a time-frequency trace was more flat.

PC noise was assumed to be constant at around 6.4 kHz, plus an extra assumption on the discontinuity from noise to the trace. Any temporal isolated group would be removed if its duration was less than the duration of other groups. Frequency change between a pair of neighboring pixels were assumed to be less than 0.5 kHz.

The final step was to smooth the time-frequency trace using the trajectory prediction with the position of each pixel on the spectrogram image. Examples of Phee, Trill, and Trillphee feature measurement algorithms were shown in Figure 2.26, Figure 2.27, and Figure 2.28 respectively.

2.6 Discussion

This chapter describes the image-based algorithms designed for automatically detect, classify and analyze marmoset vocalizations with high signal-to-noise ratio (SNR). These algorithms could be applied in future studies related to marmoset vocalizations. Here we will explain the logic behind feature selection in the algorithm, and discuss the advantages and limitations of these algorithms.

2.6.1 Feature selection in the algorithm

The main factor for how classification algorithm performs is how feature vectors are extracted and selected. Instead of using traditional spectrum values or envelope amplitude values as feature vectors, we used spectro-temporal features which have been shown to have higher impacts on discrimination between different call types and individual identity discrimination within the same call type (Agamaite 1997). Besides, these spectro-temporal features compared to PCA and MFCCs were more robust for data with high SNRs.
2.6.2 Advantages and limitations of the algorithms

There are several advantages of the algorithms we applied here. First of all, the automatic process enhances the efficiency for data processing. Secondly, the call detection with source localization algorithm provided a method that compensated the insufficient performance of directional microphone in a crowded environment. Besides, image-based algorithm offered an intuitive and robust way of analyzing data with high SNRs without worrying about the alignment between various call durations. Furthermore, the feature measurement algorithm is robust for data collected under different conditions.

However, there are some limitations in this study. First of all, the call types in the Other group can be very different. A classification with more precise classes such as Peep, Chatter, and Egg would help people understand the call behavior better. Similar situations can be applied for compound call group. The fractions of each call type and the temporal sequence within a compound call might carry important information in communication. Since the occurrence frequency of four main call types was much higher compared to other simple call types in marmoset vocalizations as well as the known quantitative modeling for these four main call types, we focused on analyzing the four main call types in this thesis.

Secondly, the overlapped calls were not separated in these algorithms, but just directly categorized as Other group. This was due to a higher number of animals than the number of microphones, which limited us from using independent component analysis (ICA) to separate overlapped signals. Although this problem could be solved by using call type recognition, which allowed multiple call types detected on the same spectrogram, it would require a long time to scan through the spectrogram with various templates especially when temporal size within one call type varies a lot.
Moreover, the reverberation in the colony room was large which affects the accuracy in source localization to identify individual marmoset identity. Fortunately, two groups of monkeys were put in different geometry locations in the colony room, which made the differentiation between two groups possible.

We would suggest using a parabolic microphone or a collar microphone in the future colony recording experiments to reduce the reverberation noise, and enhance the speaker identification accuracy.

Furthermore, the features measured here were based on the discriminability in acoustic properties, which might suggest different roles in these features but did not necessarily represent perception preference in marmosets. In the future, video cameras would benefit analyzing the marmoset behavior correlation in these vocalizations.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Window Length (sec)</td>
<td>30</td>
</tr>
<tr>
<td>Sliding Step Size (sec)</td>
<td>24</td>
</tr>
<tr>
<td>High-Pass Cutoff Frequency (Hz)</td>
<td>3000</td>
</tr>
<tr>
<td>Notch Filter Cutoff Frequency (Hz)</td>
<td>6348</td>
</tr>
<tr>
<td>Minimum Intensity Amplitude (Std)</td>
<td>2</td>
</tr>
<tr>
<td>Minimum Event Duration (msec)</td>
<td>300</td>
</tr>
<tr>
<td>Maximum Inter-Event Time (msec)</td>
<td>500</td>
</tr>
<tr>
<td>Pre- and Post- Buffer Time (msec)</td>
<td>400</td>
</tr>
</tbody>
</table>

Table 2.1 Parameters for automated call segmentation algorithm
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of FFT</td>
<td>512</td>
</tr>
<tr>
<td>Overlapping Percentage for STFT (%)</td>
<td>0.75</td>
</tr>
<tr>
<td>Re-Scaling Starting Frequency (Hz)</td>
<td>4000</td>
</tr>
<tr>
<td>Re-Scaling Ending Frequency (Hz)</td>
<td>16000</td>
</tr>
<tr>
<td>Notch Filter Frequency (Hz)</td>
<td>6348</td>
</tr>
<tr>
<td>Butterworth Filter Order</td>
<td>2</td>
</tr>
<tr>
<td>Gaussian Bilateral Filter Window Size</td>
<td>5</td>
</tr>
<tr>
<td>Spatial-Domain SD for Closeness</td>
<td>0.3</td>
</tr>
<tr>
<td>Intensity-Domain SD for Similarity</td>
<td>0.5</td>
</tr>
<tr>
<td>Pixel Intensity Scale</td>
<td>255</td>
</tr>
<tr>
<td>Global Pixel Intensity Difference</td>
<td>50</td>
</tr>
<tr>
<td>Local Pixel Intensity Fraction (%)</td>
<td>90</td>
</tr>
<tr>
<td>Closing (‘square’) Pixel Size</td>
<td>2</td>
</tr>
<tr>
<td>Opening (‘remove’) Pixel Size</td>
<td>10</td>
</tr>
<tr>
<td>Closing Iteration Times</td>
<td>5</td>
</tr>
<tr>
<td>Opening Iteration Times</td>
<td>1</td>
</tr>
<tr>
<td>Frequency Difference Threshold (Hz)</td>
<td>1000</td>
</tr>
<tr>
<td>Analysis Window Size (msec)</td>
<td>100</td>
</tr>
<tr>
<td>Sliding Window Size (msec)</td>
<td>20</td>
</tr>
<tr>
<td>Minimum Inter-Phrase-Interval (msec)</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 2.2 Parameters for feature extraction for classification algorithm
Table 2.3 Feature vector for classification algorithm

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>#</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Duration (sec)</td>
<td>2</td>
<td>Dominant Frequency (kHz)</td>
</tr>
<tr>
<td>3</td>
<td>Minimum Frequency (kHz)</td>
<td>4</td>
<td>Maximum Frequency (kHz)</td>
</tr>
<tr>
<td>5</td>
<td>Start Frequency (kHz)</td>
<td>6</td>
<td>End Frequency (kHz)</td>
</tr>
<tr>
<td>7</td>
<td>End Frequency at 1st Phrase (kHz)</td>
<td>8</td>
<td>Time Fraction of Min Frequency</td>
</tr>
<tr>
<td>9</td>
<td>Time Fraction of Max Frequency</td>
<td>10</td>
<td>Beginning Mean Bandwidth (kHz)</td>
</tr>
<tr>
<td>11</td>
<td>Beginning Bandwidth Std (kHz)</td>
<td>12</td>
<td>Beginning Max Bandwidth (kHz)</td>
</tr>
<tr>
<td>13</td>
<td>Middle Mean Bandwidth (kHz)</td>
<td>14</td>
<td>Middle Bandwidth Std (kHz)</td>
</tr>
<tr>
<td>15</td>
<td>Middle Max Bandwidth (kHz)</td>
<td>16</td>
<td>Ending Mean Bandwidth (kHz)</td>
</tr>
<tr>
<td>17</td>
<td>Ending Bandwidth Std (kHz)</td>
<td>18</td>
<td>Ending Max Bandwidth (kHz)</td>
</tr>
<tr>
<td>19</td>
<td>Mean Time Gap (sec)</td>
<td>20</td>
<td>Time Gap Std (sec)</td>
</tr>
<tr>
<td>21</td>
<td>Max Time Gap (sec)</td>
<td>22</td>
<td>Number of Phrases</td>
</tr>
<tr>
<td>23</td>
<td>Mean Frequency Gap (kHz)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>#</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Event Duration (sec)</td>
<td>2</td>
<td>Contour Existence (logical)</td>
</tr>
<tr>
<td>3</td>
<td>Wide Contour Time Gap Number</td>
<td>4</td>
<td>Narrow Contour Time Gap Number</td>
</tr>
<tr>
<td>5</td>
<td>Number of Phrases</td>
<td>6</td>
<td>Periodicity (logical)</td>
</tr>
<tr>
<td>7</td>
<td>Beginning High Freq Hit (logical)</td>
<td>8</td>
<td>Middle High Freq Hit (logical)</td>
</tr>
<tr>
<td>9</td>
<td>Ending High Freq Hit (logical)</td>
<td>10</td>
<td>Minimum Frequency (kHz)</td>
</tr>
<tr>
<td>11</td>
<td>Maximum Frequency (kHz)</td>
<td>12</td>
<td>Event Interval (sec)</td>
</tr>
<tr>
<td>13</td>
<td>Beginning Regression MSE</td>
<td>14</td>
<td>Ending Regression MSE</td>
</tr>
<tr>
<td>15</td>
<td>Frequency Gap Existence (logical)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3 Feature vector for classification algorithm

a. Feature vector candidate 1
b. Feature vector candidate 2
<table>
<thead>
<tr>
<th>Development Set</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source from Utah</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Source from M9606</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature Length</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>Total Call Number</td>
<td>875</td>
<td>2443</td>
<td>2289</td>
<td>2289</td>
</tr>
<tr>
<td>Twitter Number</td>
<td>125</td>
<td>537</td>
<td>537</td>
<td>537</td>
</tr>
<tr>
<td>Phee Number</td>
<td>125</td>
<td>759</td>
<td>759</td>
<td>759</td>
</tr>
<tr>
<td>Trill Number</td>
<td>125</td>
<td>127</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>Trillphee Number</td>
<td>125</td>
<td>147</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Compound Number</td>
<td>125</td>
<td>151</td>
<td>151</td>
<td>151</td>
</tr>
<tr>
<td>Other Number</td>
<td>125</td>
<td>136</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>Noise Number</td>
<td>125</td>
<td>587</td>
<td>587</td>
<td>587</td>
</tr>
</tbody>
</table>

Table 2.4 Development set for classification training model
Figure 2.1  Block diagram for Utah project
Figure 2.2  Schematic diagram of vocalization data collection

a. Set up for recording marmoset vocalizations
b. Example of 4-channel recording data in Utah S18
Figure 2.3  Automated call segmentation algorithm framework
Figure 2.4  Example of automated call segmentation algorithm

c.  Example of 2-channel recording data in 9606 S138

d.  Contour threshold of channel 1

e.  Source identification simplified schematic diagram projected on 2-D. Open circle is source. Filled circles are microphones. Delay between each time of arrival pairs generates a hyperbola. Intersection of these curves estimates the source location.
Figure 2.4  Example of automated call segmentation algorithm

d.  Raw envelope and spectrogram of channel 1 in 9606 S138

e.  Filtered envelope and spectrogram of channel 1 data

f.  Segmented event for channel 1 after source localization

 g.  Segmented event for channel 2 after source localization
Figure 2.5  Automated call segmentation algorithm performance

a. Detection sensitivity and precision with different amplitude thresholds.
b. Detection F-score with different amplitude thresholds.
c. Detection sensitivity and precision with different duration thresholds.
d. Detection F-score with different duration thresholds.
a. Data was split into two sets. Development set was further subdivided into a training set and a validation set for cross-validation in training.

b. Automatic supervised classification block diagram. During training, feature vectors were extracted for constructing the classifier model. During prediction, the classifier model was used to predict the call types in all data set.
Figure 2.7  Example of twitter calls

(a. b. c.) Twitter call can have different number of phrases. (c. d. e. f.) Twitter call can have different number of linear fractions, as well as different time fraction of knee point. (g. h. i. j. k. l.) First phrase of a twitter call can have many variations, which eventually lead to a compound call.

a. S18 channel 1 4th call. Twitter call has three phrases.
b. S18 channel 1 35th call. Twitter call has sixteen phrases.
c. S18 channel 1 5th call. Twitter call has about five phrases. The time fraction of knee point is extremely small that it almost has only one linear fraction.
d. S57 channel 4 45th call. Twitter call usually has two linear fractions and the time fraction of knee point is in the middle of phrases.
e. S57 channel 1 65th call. Twitter call has two linear fractions and the time fraction of knee point is in the end of phrases.
f. S18 channel 2 33rd call. Twitter call has four linear fractions for the middle part.
g. S40 channel 1 45th call. First phrase of twitter call has a flat linear fraction before knee point, and the starting frequency is higher than the other phrases.
h. S40 channel 2 8th call. First phrase of twitter call has an arc before knee point, which can also be classified as a SaPeep-Twitter call.
i. S40 channel 1 26th call. First phrase of twitter call has a rising linear fraction before knee point, and the starting frequency is higher than the other phrases.
j. S40 channel 2 4th call. First phrase of twitter call has frequency modulation before knee point, and the starting frequency is lower than the other phrases.
k. S38 channel 1 1st call. First phrase of twitter call has a rather long flat linear fraction before knee point, which can also be classified as a Phee-Twitter call.
l. S40 channel 2 5th call. Fist phrases of twitter call has a sinusoidal frequency modulation before knee point, which can also be classified as a Trill-Twitter call.
Figure 2.8  Example of narrowband calls

(a, b.) Phee call can have different slow frequency modulation rates. (c, d.) Phee call can have different variations in the ending frequency. (e, f.) Phee call has different variations in the starting frequency. (g, h.) Trillphee call can have different trilling fraction lengths. (i, j.) Trill call can have different frequency modulation depths and durations.

a.  S18 channel 2 2nd call. Phee call has flat frequency contour.
b.  S18 channel 1 53rd call. Phee call has slow frequency modulated contour.
c.  S38 channel 4 7th call. Phee call has falling frequency in the end.
d.  S18 channel 3 13th call. Phee call has rising frequency in the end.
e.  S57 channel 1 51st call. Phee call has rising frequency in the beginning.
f.  S14 channel 2 10th call. Phee call has falling frequency in the beginning.
g.  S38 channel 2 2nd call. Trillphee call has short trilling fraction.
h.  S40 channel 2 66th call. Trillphee call has long trilling fraction.
i.  S18 channel 2 7th call. Trill call has large frequency modulation depth.
j.  S18 channel 1 24th call. Trill call has small frequency modulation depth.
Figure 2.9  Example of compound calls

(a. b.) Compound call consists of a short peep and a phee call. (c. d. e.f.) Compound calls consists of multiple phrases of same narrowband calls. (g. h.) Compound call consists of multiple short peeps. (i. j.) Compound call consists of different call types.

a.  S111 channel 2 31\textsuperscript{st} call. Phee-Peep.
b.  S18 channel 4 8\textsuperscript{th} call. Peep-Phee.
c.  S14 channel 3 24\textsuperscript{th} call. Phee-Phee.
d.  S111 channel 2 25\textsuperscript{th} call. Phee-Phee. First phee has frequency modulated end.
e.  S40 channel 2 25\textsuperscript{th} call. Phee compound. Each phrase has frequency modulation in the beginning and ending.
f.  S32 channel 3 81\textsuperscript{st} call. Trill-Trill. Each trill has falling frequency contour.
g.  S38 channel 2 2\textsuperscript{nd} call. Tpeep compound.
h.  S40 channel 2 66\textsuperscript{th} call. Ppeep compound.
i.  S18 channel 2 7\textsuperscript{th} call. Twitter-Phee compound.
j.  S18 channel 1 24\textsuperscript{th} call. Tsik-Ek (or Tsik-Egg) compound.
<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Frequency (kHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

![Waveform and spectrogram for S18 Ch1](image)

![Waveform and spectrogram for S18 Ch4](image)

![Waveform and spectrogram for S15 Ch2](image)

![Waveform and spectrogram for S40 Ch2](image)
Figure 2.10  Example of other call types and noise

a. S40 channel 2 33\textsuperscript{rd} call. Chirp (or dTwitter).
b. S40 channel 2 24\textsuperscript{th} call. Chirp with frequency variations, which can also be classified as Peep-String.
c. S129 channel 1 3\textsuperscript{rd} call. Chatter.
d. S18 channel 1 18\textsuperscript{th} call. Squeal (or infant cry).
e. S57 channel 2 49\textsuperscript{th} call. Peep.
f. S40 channel 2 45\textsuperscript{th} call. Scream.
g. S32 channel 3 1\textsuperscript{st} call. Phee overlapped with twitter.
h. S32 channel 1 25\textsuperscript{th} call. Chatter overlapped with modulated phee.
i. S57 channel 4 42\textsuperscript{nd} event. Non-transient noise.
j. S40 channel 2 69\textsuperscript{th} event. Transient noise.
Figure 2.11  Call type manual label agreement chart

Three people manually labeled call type percentage of randomly selected twelve sessions from all the data set. The compound calls and other call types were ignored here. The model was constructed based on CJC labels.
Figure 2.12  Feature extraction for classification algorithm framework
Figure 2.13  Example of feature extraction for classification algorithm

a. Spectrogram of raw S32 channel 3 48th call data.
b. Smooth filtered spectrogram
c. Pixel intensity normalized to gray image (0-255)
d. Binary image after pixel intensity thresholding
e. Binary image after morphological operation
f. Binary image after harmonic removal
Figure 2.14 KNN machine learning algorithm framework

a. 10-fold cross-validation was used to avoid over-fitting. Within each fold, a KNN model was constructed and used to classify the validation set, which gave the CV classification accuracy. This process repeats 10 times.

b. The unlabeled data was assigned to a class based on the majority vote of K nearest neighbors in feature space. Distance in high-dimensional feature space was normalized Euclidean distance.
Figure 2.15  KNN machine learning algorithm performance

KNN machine learning algorithm performance across different K number was compared between four different development sets defined in Table 2.4.

a. Overall classification accuracy.
b. Cross-validation classification accuracy.
c. Twitter classification sensitivity.
d. Twitter classification precision.
e. Phee classification sensitivity.
f. Phee classification precision.
g. Trill classification sensitivity.
h. Trill classification precision.
i. Trillphee classification sensitivity
j. Trillphee classification precision.
Figure 2.16 Feature weights in optimal KNN model

The optimal KNN model (set 2, K = 6) was selected based on the performance in Figure 2.15. To maximize the classification accuracy, feature weights are calculated for Euclidean distance measurement.

a. There are 24 features extracted in set 2 for classification model training. Red line shows mean(weights) + 0.5 std(weights)

b. Features are sorted based on their rank in feature weights.
Figure 2.17  Decision tree machine learning algorithm framework

a. 10-fold cross-validation was used to avoid over-fitting.
b. The decision tree model was constructed to maximize posterior probability and the observations in each branch split with CART algorithm. Each observation falls into one and only one terminal node, and each node is uniquely defined by a set of binary rules.
Figure 2.18  Decision Tree machine learning algorithm performance

Decision tree machine learning algorithm performance across different minLeaf number was compared between four different development sets.

a. Original tree node number.
b. Cross-validation classification accuracy for original tree.
c. Node number for pruned trees.
d. Cross-validation classification accuracy for pruned trees
e. Twitter classification sensitivity.
f. Twitter classification precision.
g. Phee classification sensitivity.
h. Phee classification precision.
i. Trill classification sensitivity.
j. Trill classification precision.
k. Trillphee classification sensitivity
l. Trillphee classification precision.
Figure 2.19  Tree structure in optimal decision tree model

The optimal decision tree model (set 1, minLeaf = 5, pruned node number = 97) was selected based on the performance shown in Figure 2.18. Rule for each node was described in the decision process. The terminal node shows the class type and the estimated probability of calls correctly classified.

a.  Half of the tree structure with mean BW -M (feature 13) smaller than 0.8486.
Figure 2.19  Tree structure in optimal decision tree model

b. Half of the tree structure with mean BW -M (feature 13) larger than 0.8486.
Figure 2.20 SVM machine learning algorithm framework

a. 10-fold cross-validation was used to avoid over-fitting. In each fold, an optimal SVM model was selected based on inner cross-validation.

b. Pairwise one-against-one strategy was used for multi-class classification.
Figure 2.21 Parameter selection for optimal SVM model

Grid search was applied for selecting hyper-parameter C and gamma in SVM kernel within each cross-validation fold. Colorbars show the cross-validated classification accuracy. The summit of the contour suggests the parameter pair for maximal classification accuracy. Development sets were from Table 2.4.

a. Set 1 grid search accuracy contour.
b. Set 2 grid search accuracy contour.
c. Set 3 grid search accuracy contour.
d. Set 4 grid search accuracy contour.
Figure 2.22 SVM machine learning algorithm performance

After selecting the optimal hyper-parameter pair (C and r) for each development set during cross-validation, the SVM performance was evaluated and compared between four development sets.

a. Original shows overall accuracy with parameters (C = 1, r = 0.5). Best CV shows the best cross-validation accuracy in Figure 2.21. Parameterized shows overall accuracy with selected parameters.
b. Four main call type classification sensitivity between four sets.
c. Four main call type classification precision between four sets.
Optimal models selected from three machine learning algorithms were compared. KNN (set 2, K = 6). Tree (set 1, minLeaf = 5). SVM (set 1, $\log_2 C = 7$, $\log_2 \gamma = -5$).

a. Original shows overall accuracy. Cross-validated shows CV accuracy.
b. Four main call type classification sensitivity between four sets.
c. Four main call type classification precision between four sets.
Figure 2.24 Feature measurement algorithm framework
Figure 2.25  Example of Twitter feature measurement algorithm

a. Spectrogram of a twitter call (S40 Ch1 26th).
b. Edge detection and harmonic removal.
c. Spectrotemporal trace of maximal pixel intensity.
d. Computer noise (constant noise at around 6.4 kHz) removal.
e. Wrap each phrase with edge detection.
f. Measure main features after Kalman prediction. (Peak (blue star): max frequency for each phrase; Trough (green star): min frequency for each phrase; Knee (red circle): knee point where trace slope changes; Dominant (purple star): frequency with largest energy.)
Figure 2.25  Example of Twitter feature measurement algorithm

g. Spectrogram of a twitter call (S23 Ch3 33rd).
h. Edge detection and harmonic removal.
i. Spectrotemporal trace of maximal pixel intensity.
j. Wrap each phrase with edge detection after computer noise removal.
k. Kalman prediction smoothed the trace by moving the outliers.
l. Measure main features. (Peak (blue star): max frequency for each phrase; Trough (green star): min frequency for each phrase; Knee (red circle): knee point where trace slope changes; Dominant (purple star): frequency with largest energy.)
Figure 2.26  Example of Phee feature measurement algorithm

a. Spectrogram of a phee call. (S18 Ch2 2nd)

b. Spectrotemporal trace of maximal pixel intensity.

c. Computer noise (constant noise at around 6.4 kHz) removal.

d. Discontinuous terminal regions removal.

e. Kalman prediction smoothed the trace by moving the outliers.

f. Measure main features. (Max (blue star): maximal frequency; Min (red star): minimum frequency; Dominant (purple circle): dominant frequency
Figure 2.26  Example of Phee feature measurement algorithm

g. Spectrogram of a phee call. (S14 Ch2 10th)
h. Spectrotemporal trace of maximal pixel intensity.
i. Discontinuous terminal regions removal after computer noise removal.
j. Measure main features after trace smoothed by Kalman predictor. (Max (blue star): maximal frequency; Min (red star): minimum frequency; Dominant (purple circle): dominant frequency)
Figure 2.27  Example of Trill feature measurement algorithm

a. Spectrogram of a trill call. (S18 Ch1 24th)
b. Spectrotemporal trace of maximal pixel intensity
c. Computer noise (constant noise at around 6.4 kHz) removal.
d. Discontinuous terminal regions removal.
e. Kalman prediction smoothed the trace by moving the outliers.
f. Measure main features. (Max (purple circle): maximal frequency; Min (red circle): minimum frequency; Peak (blue star): sinusoidal frequency peak; Trough (green star): sinusoidal frequency trough.)
Figure 2.27  Example of Trill feature measurement algorithm

g. Spectrogram of a trill call. (S18 Ch1 22nd)
h. Spectrotemporal trace of maximal pixel intensity
i. Computer noise (constant noise at around 6.4 kHz) removal.
j. Discontinuous terminal regions removal.
k. Kalman prediction smoothed the trace by moving the outliers.
l. Measure main features. (Max (purple circle): maximal frequency; Min (red circle): minimum frequency; Peak (blue star): sinusoidal frequency peak; Trough (green star): sinusoidal frequency trough.
Figure 2.28  Example of Trillphee feature measurement algorithm

a. Spectrogram of a trillphee call. (S38 Ch2 2nd)
b. Spectrotemporal trace of maximal pixel intensity
c. Computer noise (constant noise at around 6.4 kHz) removal.
d. Temporal discontinuous terminal regions removal.
e. Frequency discontinuous terminal regions removal.
f. Measure main features after Kalman prediction. (Max (purple circle): maximal frequency; Min (red circle): minimum frequency; Trans (brown circle): trill-phee transition point; Peak (blue star): sinusoidal frequency peak; Trough (green star): sinusoidal frequency trough.)
CHAPTER 3:

Social Interaction Effect on Marmoset Conspecific Vocalization Structures

3.1 Introduction

Vocal communication is a fundamental component of animal behavior. Each vocal repertoire has its specific spectra-temporal acoustic features, which are bound together into perceptually stable units to convey information (Hauser 2001). These acoustic features have statistical variations in calls between individuals (Mitani and Brandt 1994) and within individuals (Seyfarth and Cheney 1984). Such variability in vocalization features does not only originate from the anatomical difference in vocal production structures and limitations in motor control within each individual (Hammerschmidt and Fischer 2008), but also come from the learning ability to adjust social distance (Giles et al. 1991, Boughman 1998).

Modification to call features in response to social interaction was described in two different ways in previous studies: vocal convergence (Baker and Cunningham 1985) and dialectal divergence (Beecher and Stoddard 1990). Vocal convergence, a process which vocal similarity between individuals increased, has been found during vocal exchanges (Sugiura 1998), and social bonding (Lemasson and Hausbergr 2004). On the contrary, dialectal divergence, a process which vocal distinctiveness between individuals increased, has been found during neighboring groups (Ford and Fisher 1983, Crockford et al. 2004), and social isolation (Masataka and Symmes 1986).
While evidence for social influences on nonhuman primate vocalization features has been shown in several studies (Elowson and Snowdon 1994, Jorgensen and French 1998), how social interaction affects features of common marmoset vocalizations is still unknown. Besides, previous studies only examined the contact call and its basic features like duration, peak frequency, and bandwidth. Information about how other call features are modulated was missing. Moreover, some features were found to remain individually distinct and fixed over several years in a socially stable environment (Symes et al. 1979, Butynski et al. 1992). Therefore, understanding the variations in call repertoire in different timescales would be important to reveal the balance between motor control and social interaction.

On the other hand, communicative information is not limited in call types and their acoustic features. Behavior cues such as temporal sequence between calls play significant roles in communication as well. Antiphonal calling paradigm (Miller et al. 2009) was used to study vocal interaction between common marmosets. However, whether the temporal exchange dynamics in communication differs under different social environment remains unknown. Furthermore, the temporal sequence in other call types was not studied yet.

In this chapter we present the dynamics of marmoset call counts, call percentage, call power, and call features between two groups of marmosets previously unfamiliar with each other to study their social interaction. We also compare the feature distributions between marmosets in current colony and their ancestors to understand the flexibility of call feature variations in different generations. Finally, we analyze the temporal sequence of call events to reveal the social interaction process and the potential behavioral relevance of each call type and their temporal relationships.
3.2 Results

Recording experiment procedure was described in Section 2.2. The call types and their features were automatically classified and measured as described in Chapter 2. Overall call number of each call type was shown in Table 3.1. There were 28020 calls detected in total, while Utah monkeys contributed more calls (n = 16808) than monkeys originally living in the colony did (n = 11212). The percentage of each call type among all the calls was shown in Figure 3.1. Events classified into the Noise group was not taken into the consideration here. Twitter (28.4 %) and Phee (38.2 %) calls were the majority call types, while Trill (9.55 %) and Trillphee (3.92%) calls were much less than the percentage observed in the experiment done 18 years ago (Agamaite 1997). The call type distribution in two different monkey groups was further compared in Figure 3.2. Although Utah monkeys made more calls than the other group, the percentage of each call type was similar between two groups.

3.2.1 Call counts dynamics after formation of a new social environment

For days when recordings were conducted, there were four sessions. Each session took about one hour. Before Utah monkeys were placed together with the Colony monkeys to form a new social environment, channels in each session were targeted to monkeys from the same group. After two monkey groups were placed together in the same colony room, a new social environment was formed. Half of the channels was targeted to Utah group and the other half was targeted to Colony group. The recording time of each target was balanced through rotation between days. Since each session had recording time varying from forty minutes to ninety minutes, the call count in a session was normalized to one hour. At the same time, target channel number for each group was normalized to four in a day.
Normalized call counts of both Colony and Utah groups across different days were shown in Figure 3.2a. The merging day, when two monkey groups were placed together, was marked in a red dash line. Normalized call counts per hour across six periods were shown in Figure 3.2b. \(1W.b\) represents the one-week period before merging day. \(Merge\) represents the one-week period after merging day. \(1W.a\) represents the second week after merging day. \(1M.a\) represents the period after a month from merging day. \(2M.a\) represents the period after two months from merging day. \(HY.a\) represents the period after a half year from merging day. Each period had 16 sampled sessions. Call counts in the Colony group increased significantly during \(Merge\) period (Wilcoxon signed-rank test, \(p < 0.05\)) while call counts in Utah group also increased significantly (Wilcoxon signed-rank test, \(p < 0.01\)). Call counts in both groups quickly decreased back to the average on the second week after merging day. While call counts in both groups fluctuated in similar rates, Utah monkeys made more calls than Colony monkeys in each period.

Normalized call counts of each call type were further shown in Figure 3.3. For Utah monkeys, call counts of each call type increased significantly during \(Merge\) period except for Phee calls. Calls in each call type all decreased quickly during \(1W.a\) period except for Phee calls. Among these call types, Twitter, Trill, and Other calls were significantly more in later periods compared to \(1W.b\) period. Phee, Trilphee, and Compound calls remained unchanged between later periods and \(1W.b\) period. On the other hand, for Colony monkeys, calls in each call type generally remained statistically unchanged across different periods except that Twitter and Other calls increased significantly during \(Merge\) period. All the statistical tests were done with Wilcoxon signed-rank test.
3.2.2 Call percentage dynamics after formation of a new social environment

While call counts in each call type have been shown to change across different periods, how the relative percentage between these call types changed was examined in Figure 3.4. For Utah monkeys, the percentage of Twitter, Trill, and Other calls increased significantly during and after Merge period. Phee call percentage, in contrast, decreased significantly during and after Merge period. For Colony monkeys, the percentage of all the call types remained statistically unchanged across six periods except that Compound call percentage decreased significantly after IW.a period. The statistical tests were done with Wilcoxon signed-rank test.

3.2.3 Call power dynamics after formation of a new social environment

Given the benefits of louder calls on information transmission in a noisy environment, it has been shown that animals increased their vocalization sound levels while interacting with the playback (Brumn and Todt 2004). Besides, changes in vocal level of one subject have been found to produce a corresponding change in the vocal level of the other subject during social interaction, and the level matching degree could be predicted by the social desirability of the individual (Natale 1975). While call counts of both Colony and Utah monkey groups increased significantly during Merge period, we then investigated whether marmosets adjusted their call power. Call power P was defined as the mean-square value.

\[
\bar{f}(t) = f(t) - \frac{1}{T} \int_0^T f(x)dx
\]

\[
P = \frac{1}{T} \int_0^T \left(\bar{f}(t)\right)^2 dt
\]
Call signal power (mean-square value) average across different days was shown in Figure 3.2c. Average call signal power of both Colony and Utah groups in different periods was shown in Figure 3.2d. Call signal power in the Colony group increased significantly during Merge period (Wilcoxon rank-sum test, p < 0.001) while call signal power in the Utah group decreased significantly (Wilcoxon rank-sum test, p < 0.001) after merging day. Such decrease might result from the recording environmental size difference which made the reverberation problem more serious in the isolation room. The call power dynamics of Colony group did not totally follow the call count dynamics of Colony group, and the call power of both groups converged during and after 2M.a period.

To understand whether call power of different call types differ, we examined the call power of different call types as shown in Figure 3.5. For Colony monkeys, call power in most of the call types increased significantly after merging day. While the call power peak of Phee, Trill, and Compound happened during Merge period, Twitter and Other calls had the power peak delay for about one week. This showed a positive correlated trend between call counts and call power, and the increase of call power was slower than the increase of the call counts. For Utah monkeys, call power could not be compared between 1W.b and later periods because of different recording environment.

The convergence of call power between two groups was not obvious in most of the call types. The overall call power between two groups were matched eventually. However, this phenomena might come from the acoustic environment difference, as background noise signal power also showed a similar trend, as shown in Figure 3.5g. All the statistical tests were done with Wilcoxon rank-sum test.
3.2.4 Comparison of feature distribution between two generations

Besides changes in call counts and call signal power after a new social environment was formed, we would like to examine the feasibility of vocal plasticity in acoustic features of each call type. To understand this short-term process, we first investigated how these call features might have changed in a long-term process by comparing the overall feature distributions between two marmoset generations. All the calls were applied with the same feature measurement algorithm. Calls with too much noise that our feature measurement algorithm failed to analyze were excluded in analysis here.

Considering the non-Gaussian feature distributions and large sample size, we applied three statistical methods to compare the difference between two generations. The effect size was shown as Hedges’ g and auROC to complement for Mann-Whitney U-test.

Hedges’ g is a measure correcting the bias in Cohen’s d (Hedges and Olkin 1985) to quantify the distance between two distribution means. $g = 0.2$ means small effect size, $0.5$ means medium effect size, and $0.8$ means large effect size.

$$g = J(n_1 + n_2 - 2) \cdot \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

$$s^* = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$$

$$J(a) = \frac{\Gamma \left( \frac{a}{2} \right)}{\sqrt{\frac{a}{2}} \cdot \Gamma \left( \frac{a - 1}{2} \right)}$$

The area under receiver operating characteristic curve (auROC) is a measure for testing the overlapping degree between two distributions. $1$ means totally separated, and $0.5$ means fully overlapped.
For Twitter calls, there were 24 features being compared between the current monkey group (n = 1672) and the past (n = 963). Each call was separated into three parts: the first phrase (b), the average of the middle phrases (m), and the last phrase (e). Nphr represents the number of phrases. Dur represents the duration of a Twitter call. IPI represents average interval between peak amplitudes in every phrase. Tphr represents the average sweeping time of a phrase. Fmin represents the minimum frequency. Fmax represents the maximal frequency. Fdom represents the dominant frequency which has maximal intensity. Fbw represents the frequency bandwidth. tKnee represents the fraction of time to a Knee point over the phrase time. fKnee represents the fraction of frequency difference between a Knee point and starting frequency over the frequency bandwidth. Knee point was defined as the point where the frequency slope of a phrase changed dramatically. Feature statistics with the mean value and standard deviation were listed in Table 3.2.

While medians between two generations were shown significantly different for all the Twitter features except for frequency bandwidth of the last phrase (Fbw_e), these features generally had medium effect size, suggesting small changes in Twitter features across different generations, as shown in Figure 3.6.

For Phee calls, there were 10 features being compared between the current monkey group (n = 10595) and the past (n = 2246). Dur represents the duration of a Phee call. CF represents center frequency. Fmin represents the minimum frequency. Fmax represents the maximal frequency. Fdom represents the dominant frequency. Fstart represents the initial frequency. Fend represents the final frequency. Fbw represents the frequency bandwidth. Tfmin represents the fraction of time to Fmin over duration. Tfmax represents the fraction of time to Fmax over duration. Feature statistics were listed in Table 3.3.
While medians between two generations were shown significantly different for all the Phee features except for dominant frequency (\(F_{dom}\)) and frequency bandwidth (\(F_{bw}\)), these features all had small effect size, suggesting Phee features remained almost the same across different generations, as shown in Figure 3.7.

For Trill calls, there were 12 features being compared between the current monkey group (\(n = 547\)) and the past (\(n = 1740\)). \(FM\ rate\) represents the mean frequency modulation rate. \(FM\ depth_{\text{max}}\) represents the maximal frequency modulation depth. \(FM\ depth\) means the mean frequency modulation depth. The other features had the same definition as Phee call features. Trill feature statistics were listed in Table 3.4.

Almost all Trill features were shown to have no significant difference between two generations with Mann-Whitney U-test except for maximal frequency (\(F_{max}\)). However, \(F_{max}\) also had a really small feature size (Hedges’ \(g = 0.0735\)), suggesting the Trill features remained almost the same across different generations, as shown in Figure 3.8.

For Trillphee calls, there were 14 features compared between the current monkey group (\(n = 844\)) and the past (\(n = 1528\)). \(FM\ rate\) represents the mean frequency modulation rate of the trilling part. \(FM\ depth_{\text{max}}\) represents the maximal frequency modulation depth of the trilling part. \(FM\ depth\) means the mean frequency modulation depth of the trilling part. \(t_{\text{Trans}}\) means the fraction of time to transition point when trilling stops over duration. Feature statistics with the mean value and standard deviation were listed in Table 3.5.

While medians between two generations were shown significantly different for all the Trillphee features except for time to maximal frequency (\(T_{f_{\text{max}}}\)) and frequency bandwidth (\(F_{bw}\)), these features generally had medium effect size, suggesting small changes in Phee features across different generations, as shown in Figure 3.9.
3.2.5 Feature convergence and divergence between two groups

After examining the vocal variations between two generations, we started to examine the vocal plasticity between two groups while a new social environment was formed.

Call number of each call type of both Colony and Utah groups in different periods were shown in Table 3.6. \textit{1W.b} represents the one-week period before two monkey groups were placed together in a same room. \textit{Merge} represents the period within first month after merging day. \textit{1M.a} represents the period after a month from merging day. \textit{4M.a} represents the period after four months from merging day. Feature distributions of both Colony and Utah groups in these four periods were then investigated.

To avoid the bias given from individual animals, a permutation was done for calls in both Colony and Utah groups. Calls originally classified as Colony group were randomly split into two sub-groups based on different target channels. Same procedure was done for calls in Utah group. As a result, there were four statistical tests (Mann-Whitney U-test) to measure the difference between any Colony-Utah subgroup pair. For example, \textit{Fdom\_b} of Twitter calls in both groups showed significant difference during \textit{1W.b} period, as shown in Figure 3.10m. This was supported by the fact that all pairs showed significant difference during \textit{1W.b} period, as shown in Table 3.7. In contrast, \textit{Fmin\_b} of Twitter calls in both groups showed no significant difference during \textit{1M.a} period, as shown in Figure 3.10g. Yet two pairs showed significant difference during \textit{1M.a} period, as shown in Table 3.7.

With the statistical test results of four sub-group pairs within four periods, it would be possible to understand whether feature convergence/divergence really happened between two groups, or that was just the epic phenomena of individual fluctuations.
Twitter feature distributions of both Colony and Utah groups in different periods were shown in Figure 3.10. Together with Table 3.7, we found that among 24 features, $F_{dom\_e}$, $F_{bw\_m}$, $tKnee\_m$, $fKnee\_m$, and $fKnee\_e$ of two groups converged during different periods after merging day. $Nphr$, $Dur$, $F_{max\_e}$, $F_{dom\_b}$, and $F_{dom\_m}$ of two groups converged after merging day, but diverged after four months. The other features of two groups had no significant difference and remained unchanged across all the periods. All the statistical tests were done with Mann-Whitney U-test.

Phee feature distributions of both Colony and Utah groups in different periods were shown in Figure 3.11. Together with Table 3.8, we found that all acoustic features of two groups converged during periods after merging day. Dynamics of Phee features were faster than the average convergence time of Twitter features.

Trill feature distributions of both Colony and Utah groups in different periods were shown in Figure 3.12. Together with Table 3.9, we found that among 12 features, $Dur$, $CF$, $F_{max}$, and $F_{end}$ of two groups converged after one month from merging day. Distance of $F_{dom}$ and $F_{start}$ between two groups was getting close after one month from merging day although they were significantly different across all the periods.

Trillphee feature distributions of both Colony and Utah groups in different periods were shown in Figure 3.13. Together with Table 3.10, we found that most acoustic features of two groups had no significant difference across all the periods. Few acoustic features such as $F_{max}$ and $CF$ converged after merging day, while $T_{fmax}$ diverged after merging day.
3.2.6 Temporal sequence of call events between groups and within groups

While call features showed convergence/divergence, it could result from either vocal exchanges between groups or exposure in a new social environment with different dialects. Therefore, we examined the temporal relationship between calls made from two groups.

Inter-call-interval was defined as the period between offset time of a call event and onset time of the next call event. We then defined a temporal interaction event as any interval between two successive calls less than 9 seconds, based on marmoset antiphonal calling studies which showed 9-second was the boundary of antiphonal response interval different from spontaneous call interval (Miller et al. 2009). Although they used antiphonal Phee calls in a computer playback experiment paradigm that might not be applied to other call types, we used 9 seconds as a controlled parameter in our thesis to examine the change of temporal sequence correlation across different periods. Probability of events following Poisson process occurring within 9 seconds was also calculated to model spontaneous and independent call events. N is the total call counts, T is the recording time, Δt is the call interval. Equations for both probabilities were shown as follows.

\[
P_{\text{temporal interaction}} = \frac{n(\Delta t \leq 9)}{N}
\]

\[
P_{\text{Poisson process}} = 1 - e^{-\frac{N}{T}9}
\]

Inter-call-interval distribution within 9 seconds in four periods was examined. Within each period, inter-call-interval distribution on each recording channel and the distribution across all the channels were shown in 50-bin histograms in Figure 3.14. The peak occurred at around 0.72 second for call events across channels, and the peak occurred at around 0.9 second for call events on individual channel.
The probability of temporal interaction across all the channels in each session was calculated and shown in Figure 3.15a. Temporal interaction increased from 0.59 to 0.67 during Merge period, and then decreased back to 0.60 a month later. The probability of temporal interaction between two groups was calculated and shown in Figure 3.15b. The probability of temporal interaction within colony channels in each session was calculated and shown in Figure 3.15c. The probability of temporal interaction within Utah channels in each session was calculated and shown in Figure 3.15d.

Within all the temporal interaction pairs, it included cross-group interaction, within-Colony-group interaction, and within-Utah-group interaction. They were defined as below, where C represents call event from Colony group and U represents calls from Utah group.

\[ P_{\text{cross}} = \frac{\{C - U\} \cup \{U - C\}}{\{C - U\} \cup \{U - C\} \cup \{C - C\} \cup \{U - U\}} \]

\[ P_{\text{WithinColony}} = \frac{\{C - C\}}{\{C - U\} \cup \{U - C\} \cup \{C - C\} \cup \{U - U\}} \]

\[ P_{\text{WithinUtah}} = \frac{\{U - U\}}{\{C - U\} \cup \{U - C\} \cup \{C - C\} \cup \{U - U\}} \]

Proportion of cross-group interaction among all the temporal interaction was shown in Figure 3.16a, which did not have a significant change. Proportion of within-Colony-group interaction was shown in Figure 3.16b. It increased from 0.18 to 0.21, and then decreased to 0.15 after 4 months. Proportion of within-Utah-group interaction was shown in Figure 3.16c. It decreased from 0.38 to 0.34, and then increased to 0.45 after 4 months. Among all these interactions, within-Colony-group interaction was less than the other two interaction types. To examine the raw temporal sequence relationship without assumptions on interaction window (9 sec), we further applied cross-correlation on call sequences.
For each session, channels were split into two channels: one recorded Utah group calls, and the other recorded Colony group calls. Cross-correlation of call sequences between these two channels were calculated. Within a period, the correlation was averaged across sessions, as shown in Figure 3.17a. Peak occurred at around 2.1 sec during Merge period, and the cross-correlation peak decreased after 4 months later. Within Utah channels, cross-correlation of call sequences between channels recorded Utah group calls were also calculated. Averaged cross-correlation within a period was shown in Figure 3.17b. Cross-correlation of call sequences between channels recorded Colony group calls were shown in Figure 3.17c. As a whole, cross-correlation of call sequences between two groups were stronger than within group cross-correlation.

To further examine whether peak at 2 sec is an epic phenomena, we further examined auto-correlation on each recording channel and time shift predictor. Auto-correlation of call sequences onset time within Colony channels was calculated and averaged, as shown in Figure 3.18a. Auto-correlation of call sequences onset time within Utah channels was shown in Figure 3.18b. Small peaks at around 2 seconds can be observed. We then shifted a call sequence at 60 seconds and 50 seconds later and calculated the cross-group cross-correlation, as shown in Figure 3.18c. We found the peak just directly shifted 60 seconds and 50 seconds later compared to Figure 3.17a. Similar effect could also be observed in within-Colony-group cross-correlation with time shifts, as shown in Figure 3.18d. This reflected cluster structure within call sequences on each channel and across channel, as suggested from Figure 3.14.
3.3 Discussion

3.3.1 Social interaction revealed by changes in vocalization structures

Communication structures consisted of the percentage of each call type produced, call acoustic features, and the temporal sequence between calls.

The difference in call type percentage between two generations probably came from the social structure change across two generations. Trill and Trillphee calls might not play an important role compared to Twitter and Phee calls in the current colony. Such difference could come from the sampling bias in our recording set up that Trill and Trillphee calls had less intensity to be picked up by our segmentation algorithm, or they were overlapped with other calls that could not be identified in our classification algorithm.

The difference between vocalizations made by two generations could also be found in the acoustic feature distributions. Most of the call features were shown to be similar for a small effect size between two generations, suggesting a limitation of vocal control in common marmosets. Other call features, however, showed a large difference. In Twitter calls, average phrase sweeping time ($T_{phr}$) was generally smaller now than in the past, which might be correlated to a larger animal population in the current colony. Number of phrases ($N_{phr}$) was more now, and frequencies were generally lower. In Trillphee calls, duration ($Dur$) was shorter and frequencies were generally lower now. While there are some other factors such as the individuals, gender, weights, recording time of a day that could contribute to the difference in these features as well, the similarities of most of the features suggested the major factor for changes in duration and frequency should come from the adaptation to a larger population size. This also suggested that Phee and Trill call features were more robust to be inherited across generations than other call types.
General increase in call counts of each call type during the Merge period suggested animals could detect changes in social structures. While more submissive calls and contact calls were made during the formation of new social structures, vocal exchanges were also more within the first month after merging day based on the temporal sequence analysis. Percentage difference across different periods after formation of new social structures also indicated the communication context difference in different periods. More experiments needed to be done in the future to verify the behavioral relevance of each call type. Stronger call power was observed during periods with more call counts, which suggested a vocal compensation in a noisy environment to convey information.

Acoustic features such as frequencies in Phee calls and duration in Twitter calls were also found to converge between two groups after placing them together to form new social structures. Few other features such as time to maximal frequency (Tfmax) in Trillphee calls showed divergence between two groups after merging day. However, some features such as dominant frequency (Fdom), minimum frequency (Fmin) in Twitter calls, and minimum frequency (Fmin) in Phee calls showed both convergence and divergence across different periods, which suggested social interaction was balanced between caller identification and group cohesion instead of dominated by individual one process when new social members were introduced. This might come from natural variations across time.

Feature convergence was mostly dominated by changes in features of Utah monkeys while features of Colony monkeys remained unchanged. This suggested foreigners to an existing colony had higher vocal plasticity and sensitivity to fit in a new social environment with larger population size. We would suggest isolate two monkey groups into two separate rooms again, and test whether call features of each group diverged again.
Besides separating animals again to verify the convergence/divergence process of acoustic features, these feature variations could come from changes of neighboring animals. In future experiments, we would suggest fix neighboring animals, and change neighboring animals several months later as a comparison.

The limitations of this study included the ambiguity of animal individual identity. The source localization algorithm could recognize group identity but not individual identity due to the reverberation issue. This would bias the sampling in this thesis. Although recording locations were rotated and a permutation test was also applied to balance the sampling, we could not exclude the possibility that most changes in call counts, call features, temporal sequence interaction were dominated by few individual monkeys.

We would suggest using a parabolic microphone or a collar microphone in the future colony recording experiments to enhance the speaker identification accuracy. However, collar microphones might make monkeys vocalize less, and the cost to give every monkey a collar microphone would be pricy. Parabolic microphones would still face a sampling challenge to capture the dialogue dynamics within a social environment. We suggest put acoustic foams on the wall to reduce reverberation and increase few more microphones, so that we could track vocalizations of every monkey in the room.

Another limitation of this study is that there was only one week for recording animals in isolated rooms. Furthermore, vocalizations of Utah group recorded before merging day could be different from their ordinary calls because of getting into a new environment. We suggest recording start a month after Utah monkeys were transported from Utah.
3.3.2 Implications in behavior relevance and motor control mechanisms

Difference in vocal plasticity of different acoustic features might also reveal different behavioral relevance of each call feature. Unchanged acoustic features might be accounted for a rather robust mechanism to maintain the caller identification and call type recognition. This could also come from the limitation of vocal production system to vary these features. On the other hand, acoustic features that were converged/diverged might function as a group ascription marker in communication. This could also come from a higher flexibility in these vocal production system for these features. To test whether these acoustic features had behavioral relevance in caller identification and context recognition, a behavioral experiment with a tool to vary different acoustic features systematically (e.g. a virtual marmoset vocalization synthesis tool in Session 4.3) would need to be done.
<table>
<thead>
<tr>
<th>Call Type</th>
<th>Colony</th>
<th>Utah</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>3109</td>
<td>4854</td>
<td>7963</td>
</tr>
<tr>
<td>Phee</td>
<td>4620</td>
<td>6080</td>
<td>10700</td>
</tr>
<tr>
<td>Trill</td>
<td>1158</td>
<td>1518</td>
<td>2676</td>
</tr>
<tr>
<td>Trillphee</td>
<td>370</td>
<td>728</td>
<td>1098</td>
</tr>
<tr>
<td>Compound</td>
<td>1000</td>
<td>1420</td>
<td>2420</td>
</tr>
<tr>
<td>Other</td>
<td>955</td>
<td>2208</td>
<td>3163</td>
</tr>
<tr>
<td>Total</td>
<td>11212</td>
<td>16808</td>
<td>28020</td>
</tr>
</tbody>
</table>

Table 3.1 Overall call counts detected
<table>
<thead>
<tr>
<th>Twitter Feature</th>
<th>Overall (N = 1672)</th>
<th>Colony (N = 572)</th>
<th>Utah (N = 1100)</th>
<th>Past (N = 963)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nphr</td>
<td>9.44 ± 2.90</td>
<td>8.77 ± 2.51</td>
<td>9.79 ± 3.02</td>
<td>7.72 ± 2.55</td>
</tr>
<tr>
<td>Dur (sec)</td>
<td>1.34 ± 0.45</td>
<td>1.26 ± 0.42</td>
<td>1.39 ± 0.46</td>
<td>1.01 ± 0.36</td>
</tr>
<tr>
<td>IPI (msec)</td>
<td>143.83 ± 19.40</td>
<td>145.31 ± 20.18</td>
<td>143.06 ± 18.94</td>
<td>139.80 ± 0.02</td>
</tr>
<tr>
<td>Tphr_b (msec)</td>
<td>35.59 ± 22.24</td>
<td>36.26 ± 21.33</td>
<td>35.24 ± 22.70</td>
<td>46.90 ± 20.68</td>
</tr>
<tr>
<td>Tphr_m (msec)</td>
<td>41.47 ± 11.70</td>
<td>40.09 ± 11.76</td>
<td>42.18 ± 11.61</td>
<td>39.26 ± 9.24</td>
</tr>
<tr>
<td>Tphr_e (msec)</td>
<td>40.31 ± 12.88</td>
<td>40.12 ± 13.21</td>
<td>40.41 ± 12.71</td>
<td>42.37 ± 10.22</td>
</tr>
<tr>
<td>Fmin_b (kHz)</td>
<td>7.14 ± 1.71</td>
<td>7.38 ± 1.69</td>
<td>7.01 ± 1.71</td>
<td>8.39 ± 1.24</td>
</tr>
<tr>
<td>Fmin_m (kHz)</td>
<td>5.63 ± 0.69</td>
<td>5.87 ± 0.60</td>
<td>5.51 ± 0.70</td>
<td>6.10 ± 0.62</td>
</tr>
<tr>
<td>Fmin_e (kHz)</td>
<td>5.59 ± 1.13</td>
<td>5.77 ± 1.11</td>
<td>5.50 ± 1.13</td>
<td>6.02 ± 0.60</td>
</tr>
<tr>
<td>Fmax_b (kHz)</td>
<td>10.54 ± 2.44</td>
<td>10.87 ± 2.58</td>
<td>10.36 ± 2.34</td>
<td>12.87 ± 1.92</td>
</tr>
<tr>
<td>Fmax_m (kHz)</td>
<td>11.40 ± 1.46</td>
<td>11.47 ± 1.38</td>
<td>11.36 ± 1.50</td>
<td>12.23 ± 1.59</td>
</tr>
<tr>
<td>Fmax_e (kHz)</td>
<td>9.39 ± 1.78</td>
<td>9.58 ± 1.74</td>
<td>9.30 ± 1.79</td>
<td>9.67 ± 1.44</td>
</tr>
<tr>
<td>Fdom_b (kHz)</td>
<td>8.32 ± 1.96</td>
<td>8.61 ± 2.08</td>
<td>8.17 ± 1.88</td>
<td>9.72 ± 1.47</td>
</tr>
<tr>
<td>Fdom_m (kHz)</td>
<td>7.17 ± 0.60</td>
<td>7.07 ± 0.63</td>
<td>7.23 ± 0.57</td>
<td>7.47 ± 0.68</td>
</tr>
<tr>
<td>Fdom_e (kHz)</td>
<td>6.65 ± 1.39</td>
<td>6.65 ± 1.30</td>
<td>6.65 ± 1.44</td>
<td>6.79 ± 0.60</td>
</tr>
<tr>
<td>Fbw_b (kHz)</td>
<td>3.40 ± 2.00</td>
<td>3.49 ± 1.99</td>
<td>3.35 ± 2.00</td>
<td>4.48 ± 1.73</td>
</tr>
<tr>
<td>Fbw_m (kHz)</td>
<td>5.85 ± 1.65</td>
<td>5.68 ± 1.56</td>
<td>5.93 ± 1.68</td>
<td>6.19 ± 1.74</td>
</tr>
<tr>
<td>Fbw_e (kHz)</td>
<td>3.80 ± 1.86</td>
<td>3.81 ± 1.87</td>
<td>3.80 ± 1.86</td>
<td>3.64 ± 1.49</td>
</tr>
<tr>
<td>tKnee_b</td>
<td>0.64 ± 0.28</td>
<td>0.63 ± 0.28</td>
<td>0.65 ± 0.29</td>
<td>0.74 ± 0.22</td>
</tr>
<tr>
<td>tKnee_m</td>
<td>0.61 ± 0.25</td>
<td>0.59 ± 0.25</td>
<td>0.62 ± 0.25</td>
<td>0.70 ± 0.22</td>
</tr>
<tr>
<td>tKnee_e</td>
<td>0.73 ± 0.24</td>
<td>0.71 ± 0.25</td>
<td>0.74 ± 0.23</td>
<td>0.80 ± 0.19</td>
</tr>
<tr>
<td>fKnee_b</td>
<td>0.27 ± 0.24</td>
<td>0.27 ± 0.24</td>
<td>0.27 ± 0.23</td>
<td>0.30 ± 0.19</td>
</tr>
<tr>
<td>fKnee_m</td>
<td>0.30 ± 0.14</td>
<td>0.27 ± 0.14</td>
<td>0.32 ± 0.14</td>
<td>0.33 ± 0.13</td>
</tr>
<tr>
<td>fKnee_e</td>
<td>0.32 ± 0.23</td>
<td>0.29 ± 0.22</td>
<td>0.34 ± 0.23</td>
<td>0.37 ± 0.18</td>
</tr>
</tbody>
</table>

Table 3.2 Measured feature distribution of Twitter
Table 3.3 Measured feature distribution of Phee

<table>
<thead>
<tr>
<th>Phee Feature</th>
<th>Overall (N = 10595)</th>
<th>Colony (N = 4575)</th>
<th>Utah (N = 6020)</th>
<th>Past (N = 2246)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dur (sec)</td>
<td>1.21 ± 0.35</td>
<td>1.20 ± 0.33</td>
<td>1.21 ± 0.36</td>
<td>1.15 ± 0.50</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>7.48 ± 0.47</td>
<td>7.42 ± 0.45</td>
<td>7.53 ± 0.47</td>
<td>7.54 ± 0.53</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>6.83 ± 0.55</td>
<td>6.77 ± 0.53</td>
<td>6.87 ± 0.56</td>
<td>6.89 ± 0.43</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>8.14 ± 0.58</td>
<td>8.06 ± 0.57</td>
<td>8.19 ± 0.59</td>
<td>8.19 ± 0.75</td>
</tr>
<tr>
<td>Tfmin</td>
<td>0.07 ± 0.22</td>
<td>0.05 ± 0.18</td>
<td>0.08 ± 0.24</td>
<td>0.07 ± 0.22</td>
</tr>
<tr>
<td>Tfmax</td>
<td>0.82 ± 0.20</td>
<td>0.81 ± 0.19</td>
<td>0.82 ± 0.20</td>
<td>0.77 ± 0.24</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>7.11 ± 0.50</td>
<td>7.05 ± 0.49</td>
<td>7.16 ± 0.50</td>
<td>7.16 ± 0.48</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>6.95 ± 0.53</td>
<td>6.89 ± 0.50</td>
<td>7.00 ± 0.54</td>
<td>7.01 ± 0.50</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>7.88 ± 0.63</td>
<td>7.84 ± 0.60</td>
<td>7.91 ± 0.65</td>
<td>7.81 ± 0.71</td>
</tr>
<tr>
<td>Fbw (kHz)</td>
<td>1.30 ± 0.65</td>
<td>1.28 ± 0.64</td>
<td>1.32 ± 0.65</td>
<td>1.30 ± 0.60</td>
</tr>
<tr>
<td>Trill Feature</td>
<td>Overall (N = 547)</td>
<td>Colony (N = 310)</td>
<td>Utah (N = 237)</td>
<td>Past (N = 1740)</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Dur (sec)</td>
<td>0.47 ± 0.19</td>
<td>0.46 ± 0.21</td>
<td>0.48 ± 0.16</td>
<td>0.45 ± 0.16</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>6.88 ± 0.71</td>
<td>6.67 ± 0.68</td>
<td>7.16 ± 0.66</td>
<td>6.84 ± 0.83</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>5.99 ± 0.79</td>
<td>5.76 ± 0.73</td>
<td>6.29 ± 0.76</td>
<td>5.97 ± 0.86</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>7.78 ± 0.81</td>
<td>7.59 ± 0.78</td>
<td>8.04 ± 0.78</td>
<td>7.70 ± 0.91</td>
</tr>
<tr>
<td>T fmin</td>
<td>0.27 ± 0.38</td>
<td>0.27 ± 0.38</td>
<td>0.27 ± 0.39</td>
<td>0.29 ± 0.39</td>
</tr>
<tr>
<td>T fmax</td>
<td>0.71 ± 0.26</td>
<td>0.71 ± 0.25</td>
<td>0.70 ± 0.28</td>
<td>0.69 ± 0.27</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>6.66 ± 0.83</td>
<td>6.42 ± 0.74</td>
<td>6.98 ± 0.84</td>
<td>6.64 ± 0.82</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>6.34 ± 0.86</td>
<td>6.08 ± 0.79</td>
<td>6.69 ± 0.82</td>
<td>6.33 ± 0.91</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>7.03 ± 1.01</td>
<td>6.81 ± 0.97</td>
<td>7.31 ± 1.00</td>
<td>6.96 ± 1.03</td>
</tr>
<tr>
<td>FM rate (Hz)</td>
<td>29.69 ± 5.89</td>
<td>29.87 ± 4.93</td>
<td>29.45 ± 6.95</td>
<td>29.96 ± 5.38</td>
</tr>
<tr>
<td>FM depth (kHz)</td>
<td>0.87 ± 0.42</td>
<td>0.94 ± 0.42</td>
<td>0.79 ± 0.41</td>
<td>0.92 ± 0.36</td>
</tr>
<tr>
<td>FM depth (kHz)</td>
<td>0.49 ± 0.19</td>
<td>0.52 ± 0.20</td>
<td>0.44 ± 0.16</td>
<td>0.50 ± 0.18</td>
</tr>
</tbody>
</table>

Table 3.4 Measured feature distribution of Trill
<table>
<thead>
<tr>
<th>Trillphee Feature</th>
<th>Overall (N = 844)</th>
<th>Colony (N = 300)</th>
<th>Utah (N = 544)</th>
<th>Past (N = 1528)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dur (sec)</td>
<td>1.09 ± 0.36</td>
<td>1.08 ± 0.35</td>
<td>1.09 ± 0.37</td>
<td>0.95 ± 0.31</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>7.65 ± 0.55</td>
<td>7.55 ± 0.57</td>
<td>7.71 ± 0.53</td>
<td>7.39 ± 0.47</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>7.01 ± 0.72</td>
<td>6.94 ± 0.77</td>
<td>7.05 ± 0.68</td>
<td>6.78 ± 0.55</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>8.30 ± 0.61</td>
<td>8.17 ± 0.59</td>
<td>8.37 ± 0.61</td>
<td>8.00 ± 0.53</td>
</tr>
<tr>
<td>Tfmin</td>
<td>0.19 ± 0.32</td>
<td>0.15 ± 0.27</td>
<td>0.21 ± 0.34</td>
<td>0.11 ± 0.29</td>
</tr>
<tr>
<td>Tfmax</td>
<td>0.66 ± 0.34</td>
<td>0.64 ± 0.33</td>
<td>0.67 ± 0.35</td>
<td>0.68 ± 0.30</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>7.43 ± 0.72</td>
<td>7.33 ± 0.74</td>
<td>7.49 ± 0.70</td>
<td>7.17 ± 0.53</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>6.34 ± 0.86</td>
<td>6.08 ± 0.79</td>
<td>6.69 ± 0.82</td>
<td>6.33 ± 0.91</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>7.03 ± 1.01</td>
<td>6.81 ± 0.97</td>
<td>7.31 ± 1.00</td>
<td>6.96 ± 1.03</td>
</tr>
<tr>
<td>Fbw (kHz)</td>
<td>1.29 ± 0.75</td>
<td>1.23 ± 0.78</td>
<td>1.32 ± 0.73</td>
<td>1.22 ± 0.55</td>
</tr>
<tr>
<td>FM rate (Hz)</td>
<td>32.17 ± 11.52</td>
<td>31.71 ± 11.06</td>
<td>32.43 ± 11.76</td>
<td>25.78 ± 5.94</td>
</tr>
<tr>
<td>FM depth$^{max}$ (kHz)</td>
<td>0.40 ± 0.40</td>
<td>0.35 ± 0.32</td>
<td>0.43 ± 0.44</td>
<td>0.42 ± 0.23</td>
</tr>
<tr>
<td>FM depth (kHz)</td>
<td>0.09 ± 0.07</td>
<td>0.09 ± 0.08</td>
<td>0.09 ± 0.07</td>
<td>0.19 ± 0.09</td>
</tr>
<tr>
<td>tTrans</td>
<td>0.29 ± 0.24</td>
<td>0.29 ± 0.24</td>
<td>0.29 ± 0.24</td>
<td>0.40 ± 0.22</td>
</tr>
</tbody>
</table>

Table 3.5 Measured feature distribution of Trillphee
<table>
<thead>
<tr>
<th>Call Number N</th>
<th>1W.b</th>
<th>Merge</th>
<th>1M.a</th>
<th>4M.a</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Twitter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colony</td>
<td>102</td>
<td>190</td>
<td>161</td>
<td>102</td>
</tr>
<tr>
<td>Utah</td>
<td>226</td>
<td>423</td>
<td>232</td>
<td>169</td>
</tr>
<tr>
<td><strong>Phee</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colony</td>
<td>1301</td>
<td>1525</td>
<td>985</td>
<td>607</td>
</tr>
<tr>
<td>Utah</td>
<td>1990</td>
<td>1725</td>
<td>1092</td>
<td>1038</td>
</tr>
<tr>
<td><strong>Trill</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colony</td>
<td>75</td>
<td>88</td>
<td>113</td>
<td>24</td>
</tr>
<tr>
<td>Utah</td>
<td>13</td>
<td>70</td>
<td>81</td>
<td>48</td>
</tr>
<tr>
<td><strong>Trillphee</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colony</td>
<td>88</td>
<td>108</td>
<td>54</td>
<td>42</td>
</tr>
<tr>
<td>Utah</td>
<td>154</td>
<td>219</td>
<td>87</td>
<td>69</td>
</tr>
</tbody>
</table>

*1W.b*: within one week before merging; **Merge**: within first month after merging; **1M.a**: from a month after merging; **4M.a**: from four months after merging

Table 3.6 Measured call type number across time
<table>
<thead>
<tr>
<th>Twitter Feature</th>
<th>1W.b</th>
<th>Merge</th>
<th>1W.a</th>
<th>4M.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nphr</td>
<td>n.s. ***</td>
<td>n.s. **</td>
<td>n.s. **</td>
<td>* **</td>
</tr>
<tr>
<td>Dur (sec)</td>
<td>** n.s.</td>
<td>* *</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>IPI (msec)</td>
<td>n.s. n.s.</td>
<td>** n.s.</td>
<td>n.s.</td>
<td>* *</td>
</tr>
<tr>
<td>Tphr_b (msec)</td>
<td>* n.s.</td>
<td>n.s. **</td>
<td>n.s.</td>
<td>n.s. ***</td>
</tr>
<tr>
<td>Tphr_m (msec)</td>
<td>n.s. ***</td>
<td>** *</td>
<td>n.s.</td>
<td>* *</td>
</tr>
<tr>
<td>Tphr_e (msec)</td>
<td>n.s. n.s.</td>
<td>*** *</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmin_b (kHz)</td>
<td>n.s. **</td>
<td>n.s. **</td>
<td>n.s.</td>
<td>***</td>
</tr>
<tr>
<td>Fmin_m (kHz)</td>
<td>*** ***</td>
<td>n.s. n.s.</td>
<td>* n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmin_e (kHz)</td>
<td>** ***</td>
<td>n.s. n.s.</td>
<td>n.s. n.s.</td>
<td>* n.s.</td>
</tr>
<tr>
<td>Fmax_b (kHz)</td>
<td>n.s. n.s.</td>
<td>n.s. *</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmax_m (kHz)</td>
<td>n.s. n.s.</td>
<td>n.s. *</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmax_e (kHz)</td>
<td>n.s. *</td>
<td>*** **</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fdom_b (kHz)</td>
<td>*** ***</td>
<td>** ***</td>
<td>n.s.</td>
<td>*** ***</td>
</tr>
<tr>
<td>Fdom_m (kHz)</td>
<td>** ***</td>
<td>n.s. n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fdom_e (kHz)</td>
<td>** **</td>
<td>** n.s.</td>
<td>n.s.</td>
<td>n.s. ***</td>
</tr>
<tr>
<td>Fbw_b (kHz)</td>
<td>n.s. **</td>
<td>n.s. **</td>
<td>n.s.</td>
<td>n.s. ***</td>
</tr>
<tr>
<td>Fbw_m (kHz)</td>
<td>*** n.s.</td>
<td>*** ***</td>
<td>** *</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fbw_e (kHz)</td>
<td>n.s. n.s.</td>
<td>n.s. *</td>
<td>n.s.</td>
<td>n.s. ***</td>
</tr>
<tr>
<td>tKnee_b</td>
<td>*** ***</td>
<td>n.s. *</td>
<td>n.s.</td>
<td>* ***</td>
</tr>
<tr>
<td>tKnee_m</td>
<td>n.s. ***</td>
<td>n.s. n.s.</td>
<td>n.s.</td>
<td>* ***</td>
</tr>
<tr>
<td>tKnee_e</td>
<td>n.s. n.s.</td>
<td>n.s. n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>fKnee_b</td>
<td>*** n.s.</td>
<td>n.s. **</td>
<td>n.s.</td>
<td>n.s. **</td>
</tr>
<tr>
<td>fKnee_m</td>
<td>** n.s.</td>
<td>n.s. **</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>fKnee_e</td>
<td>n.s. *</td>
<td>*** n.s.</td>
<td>*** n.s.</td>
<td>* n.s.</td>
</tr>
</tbody>
</table>

n.s.: no significance difference using Mann-Whitney U test; *: p < 0.05; **: p < 0.01; ***: p < 0.001;

Table 3.7 Twitter feature distribution difference between sub-groups
<table>
<thead>
<tr>
<th>Phee Feature</th>
<th>1W.b</th>
<th>Merge</th>
<th>1W.a</th>
<th>4M.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dur (sec)</td>
<td>***</td>
<td>n.s.</td>
<td>***</td>
<td>n.s.</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>*</td>
<td>n.s.</td>
<td>***</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>Tfmin</td>
<td>***</td>
<td>**</td>
<td>***</td>
<td>n.s.</td>
</tr>
<tr>
<td>Tfmax</td>
<td>***</td>
<td>**</td>
<td>n.s.</td>
<td>***</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>*</td>
<td>n.s.</td>
<td>***</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fbw (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

n.s.: no significance difference using Mann-Whitney U test;
*: p < 0.05; **: p < 0.01; ***: p < 0.001;

Table 3.8 Phee feature distribution difference between sub-groups
### Table 3.9 Trill feature distribution difference between sub-groups

<table>
<thead>
<tr>
<th>Trill Feature</th>
<th>1W.b</th>
<th>Merge</th>
<th>1W.a</th>
<th>4M.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dur (sec)</td>
<td>**</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>***</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Tfmin</td>
<td>**</td>
<td>**</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Tfmax</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>**</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM rate (kHz)</td>
<td>***</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM depth$_{max}$ (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>*</td>
<td>*</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM depth (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>**</td>
<td>***</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>*</td>
<td>*</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

*n.s.: no significance difference using Mann-Whitney U test;  
*: p < 0.05; **: p < 0.01; ***: p < 0.001;
<table>
<thead>
<tr>
<th>Trillphee Feature</th>
<th>1W.b</th>
<th>Merge</th>
<th>1W.a</th>
<th>4M.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dur (sec)</td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>CF (kHz)</td>
<td>***</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmin (kHz)</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fmax (kHz)</td>
<td>***</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Tfmin</td>
<td>***</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Tfmax</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fdom (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fstart (kHz)</td>
<td>n.s.</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fend (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Fbw (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM rate (kHz)</td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM depth$^{\text{max}}$ (kHz)</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>FM depth (kHz)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>tTrans</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

n.s.: no significance difference using Mann-Whitney U test; 
*: p < 0.05; **: p < 0.01; ***: p < 0.001;

Table 3.10 Trillphee feature distribution difference between sub-groups
Figure 3.1 Overall classified call type percentage

a. Call count percentage of six classified call types. Noise group was excluded here.
b. Call count percentage comparison between colony group monkeys and Utah group monkeys.
Figure 3.2  Overall call count and signal power dynamic across time

a. Call counts of both colony and Utah groups were normalized to make each day has 4-hour recordings. Call counts were also normalized to 4-channel for each group per day. Red dash line is the merging date.

b. Dynamics of normalized call counts per hour across six periods. Each period has four days. Each day has four 1-hour sessions. Call counts in colony group increased significantly during Merge period. (Wilcoxon signed-rank test p < 0.05) Call counts in Utah group also increased. (Wilcoxon signed-rank test p < 0.01)

c. Call signal power (mean-square value) dynamics in different days.

d. Call signal power in colony group increased significantly during Merge period. (Wilcoxon ranksum test p < 0.001) Signal power in Utah group decreased. (Wilcoxon ranksum test p < 0.001)
Figure 3.3  Call count dynamic across time of different call types
Figure 3.4  Call percentage dynamic across time of different call types
Figure 3.5  Signal power dynamic across time of different call types
Figure 3.6  Overall measured feature distribution of Twitter

(a. - x.) 24 feature distributions of Twitter calls are compared between calls collected in this study (black: both colony and Utah groups, n = 1672) and calls collected 18 years ago (red: ancestors of colony group, n = 963). All the calls were applied with same feature measurement algorithms.

Each twitter call was separated into three parts. First represents the first phrase of Twitter. Middle represents the average of all the middle phrases. Last represents the last phrase. Feature statistics were listed in Table 3.2.

Two distributions of same feature were compared with three statistical methods. Mann-Whitney U-test compares two medians. The effect size was shown with Hedges’ g and the area under receiver operating characteristic curve (auROC). (Hedges’ g = 0.2: small effect size; g = 0.5: medium effect size; g = 0.8: large effect size; auROC = 0.5: totally overlapped; auROC = 1: totally separated.)
Twitter Number of Phrases

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6643 \]
\[ \text{Hedges g} = 0.6204 \]

Twitter Duration

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.7164 \]
\[ \text{Hedges g} = 0.7969 \]

Twitter Inter-Phrase-Interval

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.5479 \]
\[ \text{Hedges g} = 0.2189 \]

Twitter First Phrase Sweeping Time

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.001 \]
\[ \text{auroc} = 0.6631 \]
\[ \text{Hedges g} = 0.5211 \]

Twitter Middle Phrase Sweeping Time

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.001 \]
\[ \text{auroc} = 0.5431 \]
\[ \text{Hedges g} = 0.2028 \]

Twitter Last Phrase Sweeping Time

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.001 \]
\[ \text{auroc} = 0.5409 \]
\[ \text{Hedges g} = 0.1719 \]
Twitter First Minimum Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.7304$
- $\text{Hedges g} = 0.8019$

Twitter Middle Minimum Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.6853$
- $\text{Hedges g} = 0.7168$

Twitter Last Minimum Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.6179$
- $\text{Hedges g} = 0.4449$

Twitter First Maximal Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.7692$
- $\text{Hedges g} = 1.0310$

Twitter Middle Maximal Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.6397$
- $\text{Hedges g} = 0.5492$

Twitter Last Maximal Frequency

- Past (n = 963)
- Present (n = 1672)

- $p < 0.0001$
- $\text{auroc} = 0.5635$
- $\text{Hedges g} = 0.1650$
Twitter First Dominant Frequency

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.7332 \]
\[ \text{Hedges g} = 0.7804 \]

Twitter Middle Dominant Frequency

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6269 \]
\[ \text{Hedges g} = 0.4685 \]

Twitter Last Dominant Frequency

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.05 \]
\[ \text{auroc} = 0.5290 \]
\[ \text{Hedges g} = 0.1175 \]

Twitter First Frequency Bandwidth

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6720 \]
\[ \text{Hedges g} = 0.5686 \]

Twitter Middle Frequency Bandwidth

- Past (n = 963)
- Present (n = 1672)

\[ p < 0.001 \]
\[ \text{auroc} = 0.5435 \]
\[ \text{Hedges g} = 0.2036 \]

Twitter Last Frequency Bandwidth

- Past (n = 963)
- Present (n = 1672)

\[ p = 0.1338 \]
\[ \text{auroc} = 0.5175 \]
\[ \text{Hedges g} = 0.0910 \]
Twitter First Time to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.5901
Hedges g = 0.3557

Twitter Middle Time to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.6162
Hedges g = 0.3895

Twitter Last Time to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.5922
Hedges g = 0.3154

Twitter First Freq to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.5699
Hedges g = 0.1447

Twitter Middle Freq to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.5724
Hedges g = 0.2144

Twitter Last Freq to Knee Fraction

- Past (n = 963)
- Present (n = 1672)

p < 0.0001
auroc = 0.5799
Hedges g = 0.2431
Figure 3.7  Overall measured feature distribution of Phee

(a. - j.) 10 feature distributions of Phee calls are compared between calls collected in this study (black: both colony and Utah groups, n = 10595) and calls collected 18 years ago (red: ancestors of colony group, n = 2246). All the calls were applied with same feature measurement algorithms.

Dominant frequency is frequency with maximal energy. Time to maximal frequency is divided by call duration. Same as time to minimum frequency. Feature statistics were listed in Table 3.3.

Mann-Whitney U-test examines the distance between two medians. The effect size was shown with Hedges’ g and the area under the receiver operating curve (auROC). (Hedges’ g = 0.2: small effect size; g = 0.5: medium effect size; g = 0.8: large effect size; auROC = 0.5: totally overlapped; auROC = 1: totally separated.)
**Phee Duration**

- Duration (sec)
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.0001
- auroc = 0.5436
- Hedges g = 0.1647

**Phee Center Frequency**

- CF (kHz)
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.0001
- auroc = 0.5286
- Hedges g = 0.1235

**Phee Minimum Frequency**

- Fmin (kHz)
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.0001
- auroc = 0.5337
- Hedges g = 0.1192

**Phee Maximal Frequency**

- Fmax (kHz)
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.01
- auroc = 0.5174
- Hedges g = 0.0889

**Phee Time to Minimum Frequency Fraction**

- tFmin
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.0001
- auroc = 0.5436
- Hedges g = 0.0046

**Phee Time to Maximal Frequency Fraction**

- tFmax
- Percentage (%)
- Past (n = 2246)
- Present (n = 10595)
- p < 0.0001
- auroc = 0.5473
- Hedges g = 0.2283
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Past (n = 2246)</th>
<th>Present (n = 10595)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phee Dominant Frequency</td>
<td>p = 0.0869</td>
<td>auroc = 0.5115</td>
</tr>
<tr>
<td>Phee Start Frequency</td>
<td>p &lt; 0.0001</td>
<td>auroc = 0.5283</td>
</tr>
<tr>
<td>Phee End Frequency</td>
<td>p &lt; 0.0001</td>
<td>auroc = 0.5467</td>
</tr>
<tr>
<td>Phee Bandwidth</td>
<td>p = 0.9082</td>
<td>auroc = 0.5008</td>
</tr>
</tbody>
</table>
Figure 3.8  Overall measured feature distribution of Trill

(a - l.) 12 feature distributions of Trill calls are compared between calls collected in this study (black: both colony and Utah groups, n = 547) and calls collected 18 years ago (red: ancestors of colony group, n = 1740). All the calls were applied with same feature measurement algorithms.

Dominant frequency is frequency with maximal energy. Time to maximal frequency is divided by call duration. Same as time to minimum frequency. FM rate is obtained from the average interval between peaks in the fast frequency modulation. Feature statistics were listed in Table 3.4.

Mann-Whitney U-test examines the distance between two medians. The effect size was shown with Hedges’ g and the area under the receiver operating curve (auROC). (Hedges’ g = 0.2: small effect size; g = 0.5: medium effect size; g = 0.8: large effect size; auROC = 0.5: totally overlapped; auROC = 1: totally separated.)
Trill Duration

Trill Center Frequency

Trill Minimum Frequency

Trill Maximal Frequency

Trill Time to Minimum Frequency Fraction

Trill Time to Maximal Frequency Fraction
Trill Dominant Frequency

- Past (n = 1740)
- Present (n = 547)

$p = 0.5096$
auroc = 0.5093
Hedges g = 0.0237

Trill Start Frequency

- Past (n = 1740)
- Present (n = 547)

$p = 0.6773$
auroc = 0.5059
Hedges g = 0.0101

Trill End Frequency

- Past (n = 1740)
- Present (n = 547)

$p = 0.0880$
auroc = 0.5241
Hedges g = 0.0634

Trill Mean FM Rate

- Past (n = 1740)
- Present (n = 547)

$p = 0.0991$
auroc = 0.5233
Hedges g = 0.0489

Trill Max FM Depth

- Past (n = 1740)
- Present (n = 547)

$p < 0.001$
auroc = 0.5548
Hedges g = 0.1159
Figure 3.9  Overall measured feature distribution of Trillphee

(a. - n.) 14 feature distributions of Trillphee calls are compared between calls collected in this study (black: both colony and Utah groups, n = 844) and calls collected 18 years ago (red: ancestors of colony group, n = 1528). All the calls were applied with same feature measurement algorithms.

FM rate and FM depth were measured only for trilling part. Time of transition point from fast frequency modulation to slow frequency modulation was shown as fraction. Feature statistics were listed in Table 3.5.

Mann-Whitney U-test examines the distance between two medians. The effect size was shown with Hedges’ g and the area under the receiver operating curve (auROC). (Hedges’ g = 0.2: small effect size; g = 0.5: medium effect size; g = 0.8: large effect size; auROC = 0.5: totally overlapped; auROC = 1: totally separated.)
Trillphee Duration

- Past (n = 1528)
- Present (n = 844)

$p < 0.0001$
- $\text{auroc} = 0.6120$
- $\text{Hedges g} = 0.4218$

Trillphee Center Frequency

- Past (n = 1528)
- Present (n = 844)

$p < 0.0001$
- $\text{auroc} = 0.6558$
- $\text{Hedges g} = 0.5234$

Trillphee Minimum Frequency

- Past (n = 1528)
- Present (n = 844)

$p < 0.0001$
- $\text{auroc} = 0.6222$
- $\text{Hedges g} = 0.3705$

Trillphee Maximal Frequency

- Past (n = 1528)
- Present (n = 844)

$p < 0.0001$
- $\text{auroc} = 0.6480$
- $\text{Hedges g} = 0.5240$

Trillphee Time to Minimum Frequency Fraction

- Past (n = 1528)
- Present (n = 844)

$p < 0.0001$
- $\text{auroc} = 0.5926$
- $\text{Hedges g} = 0.2729$

Trillphee Time to Maximal Frequency Fraction

- Past (n = 1528)
- Present (n = 844)

$p = 0.5647$
- $\text{auroc} = 0.5071$
- $\text{Hedges g} = 0.0851$
Trillphee Dominant Frequency

\[ \text{Percentage} \%
\]

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6247 \]
\[ \text{Hedges g} = 0.4380 \]
Past (n = 1528)
Present (n = 844)

Trillphee Start Frequency

\[ \text{Percentage} \%
\]

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6477 \]
\[ \text{Hedges g} = 0.5366 \]
Past (n = 1528)
Present (n = 844)

Trillphee End Frequency

\[ \text{Percentage} \%
\]

\[ p < 0.0001 \]
\[ \text{auroc} = 0.6643 \]
\[ \text{Hedges g} = 0.5198 \]
Past (n = 1528)
Present (n = 844)

Trillphee Bandwidth

\[ \text{Percentage} \%
\]

\[ p = 0.6316 \]
\[ \text{auroc} = 0.5059 \]
\[ \text{Hedges g} = 0.1045 \]
Past (n = 1528)
Present (n = 844)
Trillphee Mean FM Rate

- Past (n = 1528)
- Present (n = 844)

p < 0.0001
auroc = 0.6820
Hedges g = 0.7647

Trillphee Max FM Depth

- Past (n = 1528)
- Present (n = 844)

p < 0.0001
auroc = 0.6185
Hedges g = 0.0590

Trillphee Mean FM Depth

- Past (n = 1528)
- Present (n = 844)

p < 0.0001
auroc = 0.8504
Hedges g = 1.2008

Trillphee Transition Time Fraction

- Past (n = 1528)
- Present (n = 844)

p < 0.0001
auroc = 0.6560
Hedges g = 0.4639
Figure 3.10  Dynamics of Twitter feature distributions between two groups

(a. - x.) 24 feature distributions of Twitter calls are compared between colony (blue) and Utah group (green) in different periods. Population number in each group in each period is listed in Table 3.6. Shaded area represents standard error.

Feature distributions between two groups were tested with Mann-Whitney U-test, shown in red mark. Feature distributions between different periods were also tested with Mann-Whitney U-test. (n.s.: p > 0.05, *: p < 0.05, **: p < 0.01, ***: p < 0.001) The effect size between two groups in four periods is listed on the top of the figure.
Twitter First Maximal Frequency Dynamics

Hedges g: 0.2879  0.1870  0.0001  0.6607
auROC:     0.5972  0.5402  0.5015  0.6880

Twitter Last Maximal Frequency Dynamics

Hedges g: 0.1994  0.2143  0.0032  0.2456
auROC:     0.6144  0.5687  0.5075  0.6053

Twitter Middle Maximal Frequency Dynamics

Hedges g: 0.1604  0.2294  0.0709  0.0719
auROC:     0.5431  0.5672  0.5252  0.5240

Twitter First Minimum Frequency Dynamics

Hedges g: 0.1278  0.2474  0.1409  0.5334
auROC:     0.5451  0.5722  0.5514  0.6541

Twitter Last Minimum Frequency Dynamics

Hedges g: 0.0635  0.3884  0.0926  0.4629
auROC:     0.5420  0.6297  0.5385  0.6087

Twitter Middle Minimum Frequency Dynamics

Hedges g: 0.8173  0.4680  0.3385  0.7014
auROC:     0.7159  0.6516  0.6253  0.7098
Twitter First Frequency Bandwidth Dynamics

<table>
<thead>
<tr>
<th>Hedges g</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2148</td>
<td>0.5613</td>
</tr>
<tr>
<td>0.0117</td>
<td>0.5051</td>
</tr>
<tr>
<td>0.1329</td>
<td>0.5390</td>
</tr>
<tr>
<td>0.3911</td>
<td>0.6317</td>
</tr>
</tbody>
</table>

Twitter Last Frequency Bandwidth Dynamics

<table>
<thead>
<tr>
<th>Hedges g</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1515</td>
<td>0.5749</td>
</tr>
<tr>
<td>0.0391</td>
<td>0.5091</td>
</tr>
<tr>
<td>0.0562</td>
<td>0.5227</td>
</tr>
<tr>
<td>0.0133</td>
<td>0.5079</td>
</tr>
</tbody>
</table>

Twitter Middle Frequency Bandwidth Dynamics

<table>
<thead>
<tr>
<th>Hedges g</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4292</td>
<td>0.6388</td>
</tr>
<tr>
<td>0.0201</td>
<td>0.5087</td>
</tr>
<tr>
<td>0.1692</td>
<td>0.5538</td>
</tr>
<tr>
<td>0.2657</td>
<td>0.5712</td>
</tr>
</tbody>
</table>

Twitter First Dominant Frequency Dynamics

<table>
<thead>
<tr>
<th>Hedges g</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3787</td>
<td>0.6444</td>
</tr>
<tr>
<td>0.1212</td>
<td>0.5245</td>
</tr>
<tr>
<td>0.0094</td>
<td>0.5244</td>
</tr>
<tr>
<td>0.1646</td>
<td>0.5416</td>
</tr>
</tbody>
</table>

Twitter Last Dominant Frequency Dynamics

<table>
<thead>
<tr>
<th>Hedges g</th>
<th>auROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3472</td>
<td>0.6725</td>
</tr>
<tr>
<td>0.0588</td>
<td>0.5441</td>
</tr>
<tr>
<td>0.4263</td>
<td>0.6425</td>
</tr>
<tr>
<td>0.3948</td>
<td>0.6823</td>
</tr>
</tbody>
</table>
Twitter First Freq to Knee Fraction Dynamics

Twitter Last Freq to Knee Fraction Dynamics

Twitter Middle Freq to Knee Fraction Dynamics

Twitter First Time to Knee Fraction Dynamics

Twitter Last Time to Knee Fraction Dynamics

Twitter Middle Time to Knee Fraction Dynamics
Figure 3.11  Dynamics of Phee feature distributions between two groups

(a. - j.) 10 feature distributions of Phee calls are compared between colony (blue) and Utah group (green) in different periods. Population number in each group in each period is listed in Table 3.6. Shaded area represents standard error. Statistical tests were same as described in Figure 3.10.
**Phee Bandwidth Dynamics**

- Hedges g: 0.2149, 0.0400, 0.1848, 0.1885
- auROC: 0.5651, 0.5055, 0.5492, 0.5488

**Phee Dominant Frequency Dynamics**

- Hedges g: 0.5425, 0.0328, 0.1813, 0.0084
- auROC: 0.6469, 0.5083, 0.5487, 0.5019

**Phee End Frequency Dynamics**

- Hedges g: 0.3081, 0.0417, 0.0193, 0.1249
- auROC: 0.6014, 0.5059, 0.5067, 0.5363

**Phee Start Frequency Dynamics**

- Hedges g: 0.4796, 0.0792, 0.1827, 0.0080
- auROC: 0.6335, 0.5196, 0.5471, 0.5052
Figure 3.12  Dynamics of Trill feature distributions between two groups

(a. - l.) 12 feature distributions of Trill calls are compared between colony (blue) and Utah group (green) in different periods. Population number in each group in each period is listed in Table 3.6. Shaded area represents standard error. Statistical tests were same as described in Figure 3.10.
Trill Dominant Frequency Dynamics

Hedges g: 2.2114  0.7540  0.3891  0.7888  
auROC:     0.9328  0.7099  0.5991  0.7665

Trill Start Frequency Dynamics

Hedges g: 2.4126  0.7821  0.3890  0.7888  
auROC:     0.9472  0.7190  0.6042  0.7183

Trill End Frequency Dynamics

Hedges g: 0.9452  0.5330  0.3681  0.2636  
auROC:     0.7179  0.6562  0.6020  0.5799

Trill Mean FM Depth Dynamics

Hedges g: 0.7542  0.4589  0.4617  0.1519  
auROC:     0.6882  0.6227  0.6056  0.5091

Trill Max FM Depth Dynamics

Hedges g: 0.0863  0.2611  0.4495  0.0910  
auROC:     0.5836  0.5838  0.6292  0.5156

Trill Mean FM Rate Dynamics

Hedges g: 0.2244  0.0629  0.1213  0.0253  
auROC:     0.6646  0.6239  0.5055  0.6293
Figure 3.13  Dynamics of Trillphee feature distributions between two groups

(a - n.) 14 feature distributions of Trillphee calls are compared between colony (blue) and Utah group (green) in different periods. Population number in each group in each period is listed in Table 3.6. Shaded area represents standard error. Statistical tests were same as described in Figure 3.10.
Bandwidth (kHz)

Trillphee Bandwidth Dynamics

Hedges g: 0.0869 0.1239 0.1521 0.5792
auROC: 0.5162 0.5541 0.5476 0.6577

Colony Utah

Fdom (kHz)

Trillphee Dominant Frequency Dynamics

Hedges g: 0.7368 0.1885 0.1404 0.4707
auROC: 0.6653 0.5671 0.5758 0.6104

Colony Utah

Fend (kHz)

Trillphee End Frequency Dynamics

Hedges g: 0.3306 0.0792 0.0993 0.3488
auROC: 0.6037 0.5318 0.5314 0.5968

Colony Utah

Fstart (kHz)

Trillphee Start Frequency Dynamics

Hedges g: 0.6946 0.2543 0.2532 0.4277
auROC: 0.6665 0.5671 0.5758 0.6304

Colony Utah
Figure 3.14  Inter-call-interval distribution across different periods

(a,b,c,d) Inter-call-interval distribution within 9 seconds in four different periods are shown. Within each period, inter-call-interval distribution on each channel and across all the channels were shown in different colors.

a.  Period when two marmoset groups were housed in different rooms.
b.  Period within first month after a new social environment was formed
c.  Period starting a month after a new social environment was formed
d.  Period starting 4 months after a new social environment was formed
Temporal interaction was defined as inter-call-interval within 9 seconds. Probability of events following Poisson process occurring within 9 seconds in one session was calculated. First period had 40 sessions, second period had 38 sessions, third period had 36, and the last period had 32 sessions. Mann-Whitney U-test was applied.

a. Probability of temporal interaction across 4 periods.
b. Probability of cross-group interaction over temporal interaction.
c. Probability of within-Colony-group over temporal interaction.
d. Probability of within-Utah-group over temporal interaction.
Figure 3.16  Target proportion in temporal interaction

Temporal interaction pair was defined as inter-call-interval within 9 seconds. Within all the temporal interaction pairs, the proportion of colony-colony, Utah-Utah, (colony-Utah& Utah-colony) was shown.

a. Proportion of cross-group interaction over temporal interaction.
b. Proportion of within-Colony-group over temporal interaction.
c. Proportion of within-Utah-group over temporal interaction.
Figure 3.17  Cross-correlation of temporal sequence

Onset time of calls in one channel to onset time of calls in another channel (left) and Offset time of calls in one channel to onset time of calls in another channel (right)

a. Cross-group cross-correlation.
b. Within-Colony-group cross-correlation.
c. Within-Utah-group cross-correlation.
Figure 3.18  Auto-correlation and time shift predictor of temporal sequence

a. Auto-correlation of Colony channel average across different periods.
b. Auto-correlation of Utah channel average across different periods.
c. Cross-group cross-correlation (onset time to onset time) with time shifts.
d. Within-Colony-group cross-correlation with time shifts.
CHAPTER 4:

Selectivity and Tolerance of Neuronal Responses for Conspecific Vocalizations

4.1 Introduction

Communication is one of the fundamental components of both human and nonhuman animal behavior that is crucial for social interactions. It is believed that revealing neural basis of conspecific vocalizations in the auditory cortex of non-human primates can give us insights on how speech perception is formed in human brain.

Several studies suggested a hierarchy in primate cortex for encoding vocalizations. First of all, it is found that there are anatomical connections between auditory belt region and prefrontal cortex in nonhuman primates by injecting anterograde and retrograde tracers (Romonski et al. 1999). From these traces, projections formed two anatomical pathways: one from anterolateral belt (AL) to the frontal pole, the rostral principal sulcus, the inferior convexity, and caudal dorsal lateral prefrontal cortex; the other from caudolateral belt (CL) to the caudal principal sulcus, the inferior convexity, and ventral prefrontal areas.

Secondly, it is found that cortical neurons conveyed more information about abstract auditory entities than spectral-temporal patterns (Chelchik and Nelkin 2012). Besides, it is found that CL neurons have greater spatial sensitivity while AL neurons are more sensitive to call types (Tien et al. 2001). VLPFC neurons are also found to be more sensitive to call repertoire than AL neurons (Romanski et al. 2005).
Furthermore, cooling deactivation and double dissociation were also used to search whether there is a vocalization processing stream in auditory cortex. Bilateral deactivation of the anterior auditory field resulted in deficits in a call-type discrimination task, whereas bilateral deactivation of the posterior auditory field caused deficits in sound localization, but not vice versa (Lomber et al. 2008). Human auditory cortex was also found to process speech identity in anterior regions and spatial location in posterior regions, with double dissociation (Ahveninen et al. 2006). Last but not least, imaging studies also showed neural response to call types is higher in rostral lateral cortical regions while neural response to caller identities is higher in caudal temporal cortical regions (Petkov et al. 2005).

While these findings suggested dual streams in brain to process auditory information, whether vocalization is processed through a hierarchical pathway is still debated because of the following reasons. First of all, there are more connections within these areas instead of directly projecting to the prefrontal cortex (Kaas and Hackett 2000). It is found that lateral belt neurons also project to the parabelt regions, and then both caudal parabelt and rostral parabelt neurons project to the temporal-parietal areas. These interconnections may integrate auditory information and then give projections to the prefrontal cortex, so it is still unclear whether there are separate processing pathways in the auditory system.

Besides, these studies usually used few call examples (Winter and Funkenstein 1973, Mergoliash 1986, Manley and Muller-Preuss 1978, Wang et al. 1995, Tien et al. 2001, Cohen et al. 2007, Recanzone 2008). Single units are often found to respond to more than one call type or multiple acoustic features in these studies. It is possible that these neurons would respond to other untested call types, or do not respond to a same call type produced from different individuals because of acoustic variations.
Therefore, it will be necessary to manipulate vocalization features to determine which features are responsible for driving neuronal responses. Besides, conspecific vocalizations are defined by their behavioral relevance, unique acoustic spectral-temporal features, and their intrinsic statistical variations for each call type and caller. An examination of how statistical variations in vocalizations are encoded across cortical regions will be needed.

In this chapter we present a systematic stimuli design to explore how vocalizations are encoded in auditory cortex. We also propose a possible model to test the hierarchy of neurons in response to vocalizations. Several example neurons using these stimuli design were shown as preliminary data. More neurophysiology data will be needed in the future to test our neuronal hierarchy model for vocalization processing.

4.2 Theoretical hypothesis

To test our hypothesis that there is a hierarchical structure for neural representation of vocalizations in marmoset auditory cortex, we made the following assumptions.

First of all, boundaries of marmoset conspecific vocal categories are defined based on their acoustic feature distributions. It may not be true since their perceptual categorization does not necessarily meet the categorizations in acoustic space. However, it is reasonable to use statistical difference in acoustics as a first step to classify their vocal repertoire before more behavioral evidence was shown.

Secondly, the statistical distributions of their vocal repertoire was based on previous colony recordings, as described in Chapter 3. Although distributions might be biased by our recorded individuals, the small variations found across different groups and generations suggested a stable statistical distribution for each call category.
Third, the acoustic features we selected to represent vocalizations were based on their weights to discriminate call categories, individuals and gender in acoustic space found in previous study (Agamaite 1997). These features are not necessarily behavioral relevant but would be good indicators to start with.

Based on the above assumptions, our hypothesis consists of the following statements. First of all, there is a hierarchy in auditory cortical regions to process vocalizations and eventually acoustic information is transformed into call perception. Second, neurons in higher hierarchy along the processing pathway are more invariant to spatial location. Third, neurons in higher hierarchy along the processing pathway are more invariant to sound level and background noise. Fourth, statistical structures of vocalizations are represented in a distributed way that is correlated to spectral and temporal characteristics of neurons in A1. For example, the distribution of preferred frequency modulation rates of A1 neurons should match with the distribution of frequency modulation rates in Trill and Trillphee calls. Last, rostral lateral neurons had higher selectivity to vocal repertoire and higher nonlinearity of spectral and temporal integration.

Our neural hierarchical model for vocalization coding is defined as below.

1. *The representation to vocalizations applies maximal excitation coding strategy.*

Here we assumed the brain uses maximal excitation as their coding strategy to vocal repertoire, and the excitatory response projected onto a feature can be fit into a bell-shaped curve, as shown in Figure 4.1a. For neurons showed response-invariant to one feature but selective to the other feature, examples are shown in Figure 4.1b. We did not consider the inhibitory response and the saddle shaped response to vocalizations, which are two possible coding strategies as shown in Figure 4.1c and Figure 4.1d.
2. *The deterministic feature space for tuning becomes larger at a higher hierarchy.*

Vocalization can be described with n-dimensional features, and the feature space of a call type \( S \) can be defined as

\[
S = \{f_1, f_2, ..., f_n\}
\]

The neural response \( R \) (driven firing rate) to the vocalization stimuli characterized by features gave the deterministic feature space \( A \), a p-dimensional subspace of \( S \).

\[
R = R_{\text{original}} - R_{\text{spontaneous}}
\]

\[
R_A = (A(A^T A)^{-1} A^T) \cdot R
\]

\[
R_{S\setminus A} = 0
\]

\[
A = \{f_{a_1}, f_{a_2}, ..., f_{a_p}\} \text{ where } a_i \in \{1, 2, ... n\} \text{ for } i < n \text{ s. t. } A \subseteq S
\]

3. *The selectivity to features within natural call distribution is larger at a higher hierarchy.*

The neural selectivity to a feature corresponding to the natural call distribution is shown in **Figure 4.2a**. The evoked neural response to vocalization stimuli is characterized within a high-dimensional space. After projecting onto a feature dimension \( f_i \), we can fit the driven firing rate to parameters of feature \( f_i \) into a distribution with probability density function \( \varphi \).

The overall area under the distribution curve will be one, and the selectivity index for that feature is defined as the area between the distribution curve and the natural call distribution range of feature \( f_i \). Equations are shown as follows.

\[
SI_{f_i} = \int_{\mu - 2\sigma}^{\mu + 2\sigma} \varphi(f_i) \, df_i
\]

where \( \mu \) is the mean of statistical distribution of vocalizations on feature \( f_i \), and \( \sigma \) is the standard deviation of statistical distribution of vocalizations on feature \( f_i \). Selectivity would be bounded between 0 and 1.
4. *The tolerance to features within natural call distribution is larger at a higher hierarchy.*

The neural tolerance to a feature corresponding to the natural call distribution is shown in Figure 4.2b. The evoked neural response to vocalization stimuli is characterized within a high-dimensional space. After projecting onto a feature dimension $f_i$, we can fit the driven firing rate to parameters of feature $f_i$ into a distribution with probability density function $\phi$. The overall area under the distribution curve will be one, and the selectivity index for that feature is defined as the area between the distribution curve and the natural call distribution range of feature $f_i$. Equations are shown as follows.

\[
T_{f_i} = 1 - \int_{\mu-2\sigma}^{\mu+2\sigma} (\phi(f_i) - E[\psi(f_i)])^2 \, df_i
\]

\[
E[\psi(f_i)] = \int_{\mu-2\sigma}^{\mu+2\sigma} f_i \cdot \phi(f_i) \, df_i
\]

where $\mu$ is the mean of statistical distribution of vocalizations on feature $f_i$, and $\sigma$ is the standard deviation of statistical distribution of vocalizations on feature $f_i$. Tolerance would be bounded between 0 and 1.

5. *The neural discriminability for different call types is larger at a higher hierarchy.*

The neural discriminability for different call types becomes larger at a higher hierarchy, as shown in Figure 4.3. In auditory nerve fiber (ANF), the neural response to different call types are overlapped with their spectral distributions. In primary auditory cortex (A1), the distributions of neural response to different call types are expected to be more separated because of the increased nonlinearity. As the hierarchy goes higher, the discriminability of different call types based on neural response also increased. Eventually, two separated neural distributions formed call categorization perception.
6. The nonlinearity of neural representations is larger at a higher hierarchy.

The nonlinearity of a neuron encoding vocalizations also becomes larger at a higher hierarchy. This means that the neural response to partial features cannot directly predict the neural response to the deterministic feature combination. The neural response to the combinations of features is larger than the sum of neural response to individual feature.

$$R_A \gg \sum_i R_{P_i}$$

$$\sum_i P_i = A \text{ and } P_i \subseteq S$$

7. The hierarchy index can then be obtained from selectivity and tolerance.

The hierarchy of a neuron encoding vocalizations can be defined as

$$H = \frac{|A|}{|S|} \cdot \prod_{j \in A} S_{f_j} \cdot T_{f_j}$$

where SI is selectivity index and TI is tolerance index calculated from the deterministic feature space A. Hierarchy would be bounded between 0 and 1. Neuron at highest hierarchy is regarded as a grandmother cell for encoding a call category. Neuron at lowest hierarchy responds only to one feature, and thus cannot discriminate a call category from any other stimuli with the same feature.

Overall, a neuron selective to a call type has to meet the following requirements. First of all, the deterministic feature space A equals the feature space of that call type S. Second, the parameters this neuron tuned for each feature should be within the statistical acoustic space for that call type. Third, it does not respond to other call types and partial features.
To explore the possible vocalization processing pathway in auditory cortical regions, different degrees of the hierarchy calculated from the model could help us construct the pathway. The following are possible outcomes from this model fitting result. First, neuron is only selective one call category, and not responsive to other stimuli that shared similar acoustic features. Second, neuron is selective to subspace of a call category feature space. Third, neuron is selective to particular combinations of features. Fourth, neuron is selective to only one feature of vocalizations. Last, neuron is not selective to vocalization features at all. In our thesis, we focus on exploring neurons that followed the first fourth outcomes.

4.3 Stimuli design

Although signal processing methods like filter bank decompositions and independent component analysis (ICA) were used in some neural coding studies (Nagarajan et al. 2002, Averbeck and Romanski 2004, Nagarajan et al. 2002, Theunissen and Doupe 1998), these stimuli described in abstract mathematical dimensions might not directly correspond to behaviorally relevant features compared to parametric stimuli. Parametric synthetic stimuli have been used to elucidate neural mechanisms of communications in other animal species such as bats (Suga et al. 1979, Suga 1988). Therefore, we used parametric synthetic stimuli based on the equations for quantifying marmoset vocalizations (DiMattina and Wang 2006) in our thesis. The implementation codes were uploaded to the lab server.

We further selected features which played important roles in discriminating different call categories and caller identities in acoustic space as variable parameters, and the rest of the features which had less influence on call type discrimination as fixed constants. The fixed constant is set to be the average of the natural call distribution on that feature.
The variable features for narrowband calls included center frequency, slow frequency modulation bandwidth, fast sinusoidal frequency modulation rate, sinusoidal frequency modulation depth, and harmonic attenuation. The variable features for broadband calls included center frequency, phrase number, mean frequency bandwidth of a phrase, mean sweeping time of a phrase, mean time fraction of a knee point, and mean frequency fraction of a knee point. Other general variable features included sound level, spatial locations, and background noise level. Parameter range for variable features started from three standard deviations below average to three standard deviations above average, and sometimes the parameter range would be even larger when standard deviation was small.

We did not choose amplitude modulation and stimulus duration as variable features. This does not mean that these two features are not behaviorally relevant. Instead, amplitude modulation varies a lot across the same call category, which might make it less relevant to call categorization, but more important to caller identification. Duration could be important in dialogue semantics and call type discrimination, but the temporal firing pattern is not considered in our present model.

Note that the features we selected to test neural coding to vocalizations may not have direct behavioral relevance; however, they were found to have discriminability power in call categorization in acoustic space, which would be a good assumption to start with.

While synthetic marmoset vocalizations could lose some information because of the simplified acoustic structures, we selected examples of real vocalizations from our colony recordings, including simple call types (Twitter, Trill, Phee, Trillphee), and commonly observed compound call types, as a comparison of neural response to vocalizations.
4.4 Recording setup

4.4.1 Electrophysiology procedure

Electrophysiology recordings were conducted in a double-walled sound-attenuated chamber (Industrial Acoustics Company) with acoustic absorption foam (Sonex) attached on the internal walls. Marmoset subject was trained to sit in a primate chair with head fixed. Single neuronal responses in auditory cortex were recorded from right hemisphere of an awake adult marmoset (M6x) following the chronic recording preparation (Lu et al. 2001). Small craniotomies (1.1 mm diameter) were made in the skull over superior temporal gyrus, where tungsten electrodes (A-M Systems, 3-5 MΩ impedance) were allowed to penetrate through perpendicularly via a hydraulic microdrive (Trent-Wells). Each recording session lasted about 5 hours, and animals were awake but not performing any behavioral tasks. Single neuronal were sorted via a template-based online sorting software (MSD, Alpha Omega Engineering). All experimental and surgical procedures were approved by the Johns Hopkins University Animal Care and Use Committee.

4.4.2 Acoustic stimuli

All acoustic stimuli were generated digitally at a sampling rate of 97.7 kHz in Matlab, converted to analog signals (RX6, Tucker-Davis Technologies), attenuated to a desired sound pressure level (PA5, Tucker-Davis Technologies), power- amplified (Crown Audio), multiplexed and relayed (PM2R, Tucker-Davis Technologies), and delivered through an array of 32 speakers (FT28D, Fostex) covering a complete sphere. These speakers were located 1m away from the animal in a free field setup (Remington 2012).
Stimuli intensity was calibrated to 95 dB SPL at 0 dB attenuation for a 4 kHz tone. While possible, neurons were characterized for frequency, intensity, and spatial tuning following our lab protocol. All the simple acoustic stimuli such as pure tone, band-pass noise, and frequency modulated tone were played with a 500ms inter-stimulus-interval and 5 mse onset/offset ramp. While not specified, these stimuli were played with 50 dB sound attenuated level. For real and synthesized vocalization stimuli, they were played with a 1000ms inter-stimulus-interval. While not specified, these stimuli were played with 30 dB sound attenuated level. Every stimulus was presented at least 5 repetitions in a random order, and delivered from preferred spatial location of the recorded neuron.

4.5 Results

The tonotopic map of M6x right hemisphere was shown in Figure 4.4. Our recorded craniotomies were located in A1 and CM/CL. Because of the limited number of single units found responsive to vocalizations in our recordings, we only presented some examples in this thesis as preliminary data for exploring neural representations of vocalizations. Note that the limited neurons responsive to vocalizations could result from the unexperienced recording techniques and the incomplete testing protocol for the first three craniotomies. More data from other auditory cortical areas such as R, RT, and lateral belt will be needed in the future follow-up experiments as well as data from other hemispheres to test whether our hypothesis is true or not.
4.5.1 Neuronal response to call categories

Ten call categories were chosen from our colony recording library. Four simple main call types (Phee, Trill, Trillphee, Twitter) and six commonly observed compound call types (dTwitter, Tsikbark, Tsikstring, Tsikstring, Trillwitpeep, Peepstring). These vocal tokens were selected because they had acoustic features that were close to the call distribution average (statistics could be found in Chapter 3). Time-reversed vocal tokens were also delivered as a comparison of neuronal response to natural vocalizations.

Three example neurons were shown in Figure 4.5. For single units not responsive to puretones and narrowband noise, but found to be responsive to call categories, they usually also responded to broadband noise. As Figure 4.5a shows, neuron (M6xU124) responded selectively to narrowband Trill and broadband Tsikbark, Tsikstring, Peepstring calls over other call categories. Neural response to these preferred call categories was also larger than response to puretone and narrowband noise. However, their firing rates were less than the response to a broadband noise with 30 dB attenuated sound level. Besides, neuron showed onset response instead of sustained response. These suggested that vocalization stimuli might not be this neuron’s preferred stimuli.

Another neuron (M6xU135) example was shown in Figure 4.5b. This unit responded less selectively to call categories compared to former neuron example (M6xU124). Besides Trill, Tsikbark, Tsikstrg and Peepstrg calls, it also responded to Twitter and Trtwpp calls. Again, this neuron also showed onset response, and phase-locking coding strategy for some call categories. The other neuron (M6xU117) showing sustained response to the same call categories was shown in Figure 4.5c. Evoked firing rates were similar between response to broadband noise and response to preferred call categories.
4.5.2 Spatial effect on neuronal selectivity to vocalizations

In previous studies, spatial selectivity and acoustic stimuli selectivity were usually studied separately. However, this might bias our understanding of how neurons encode vocalizations in different auditory cortical areas such as A1. While neurons were found not responsive to some call categories, this could come from the inhibition from the delivered spatial location. Figure 4.6a shows a neuron example (M6xU126) that neuron was not responsive to vocalizations while they were delivered through front speaker. However, the neuron showed sustained response to the same vocalization stimuli but delivered through the best speaker, which was concluded from the spatial tuning to a broadband noise.

We further explored whether spatial effect on neural selectivity to vocalizations could be predicted by spatial selectivity obtained with broadband noise. As Figure 4.6b shows, an example neuron (M6xU128) showing contralateral preference in response to broadband noise had similar spatial preference in response to different call categories, regardless of their bandwidths. Based on the firing patterns in presence of different acoustic stimuli, this neuron was selective to spatial locations, but not vocalizations. Similar effect could also be found in neurons showing ipsilateral preference, as shown in Figure 4.6c.

4.5.3 Noise masking effect on neuronal selectivity to vocalizations

While most of the studies in previous literature directly used recorded vocal tokens, neural characteristics were usually ignored. However, neural selectivity to calls could be affected by the neural sensitivity to overall spectrum. Therefore, background noise in vocal tokens could contribute to the neural response to calls with different degrees.
As shown in Figure 4.7, example neuron (M6xU180) found responsive to real Trill token did not respond to synthesized Trill calls, regardless of sound levels. However, adding background noise in the stimuli would evoke neural response. The cut-off threshold for SNR was about 36 dB. This neuron was not selective to vocalizations, and preferred stimuli with broad spectrum in general.

4.5.4 Neuronal response to call features

An example neuron (M6xU214) was shown in Figure 4.8. Its characteristic receptive field was shown in Figure 4.8a. Its best frequency was 18.379 kHz, but showed a rather broad response bandwidth. Its preferred sinusoidal FM stimuli with low modulation rates. Its responses to different features in synthesized Trill calls were shown in Figure 4.8b to Figure 4.8j. Neuron was found to be invariant to changes in individual feature parameters. Regardless of tested features, neuron showed strong onset and offset response. We also found that feature combination evoked more response than sum of response to individual features. For features that were not directly associated with call intrinsic properties, this neuron preferred loud sounds delivered through contralateral speakers.

Neural representation to narrowband vocalization does not limit to be rate coding, but could also be temporal coding. An example neuron (M6xU4) showing firing pattern shifted from sustained response to onset response while increasing slow FM modulation depths of Phee calls. While slow FM modulation depth was 0, the stimulus was a puretone with same center frequency. Another example neuron (M6xU10) showing phase-locking response to the transition point (tTrans) in Trillphee calls. Raster plots were shown in Figure 4.9.
Neural tuning to call features did not necessarily follow our proposed model that their tuning curves could be described as bell-shaped curve, where selectivity and tolerance were obtained with consideration of call feature distribution. As shown in Figure 4.10, neuron could utilize a U-shape coding strategy to the call feature. In this case, natural distribution of FM/AM rate in Trill calls was between 20 and 40 Hz, corresponding to the parameters which evoked the least neural response.

4.6 Discussion

4.6.1 Optimal stimuli approach for future direction

In our thesis, we varied one feature at a time while keeping other features unchanged. While it is a useful way to examine neural tuning and determine the next stimuli to deliver based on the neural response, it may not reveal the whole picture of neural representations of a high-dimensional stimulus such as vocalization. In our hypothesis and model, neurons at higher hierarchy should respond to feature combinations more than the sum of response to individual features. This will require presenting stimuli varying several features at the same time. Furthermore, it was observed that neurons selectively responding to few call categories over other call categories could be predicted from the response to spectrum, and their firing rates to vocalizations were less than response to broadband noise or puretones in A1. Although more data will be required to conclude neural coding strategy for calls in A1, it suggests that our vocalization stimuli might not be the optimal stimuli for neurons. In the future experiments, we would suggest delivering numerous vocalization stimuli from the high-dimensional acoustic space together with an gradient search algorithm, and project neuronal response to each dimension to test our model.
Figure 4.1  Possible neural coding strategies for vocalization

Possible neuronal coding strategies were shown with neural firing rates in response to stimuli projected onto a 2D feature plane. Red cylinder bounds the natural vocalization distribution with 95% calls (within 2 Std).

a.  Bell-shaped
b.  Ridge-shaped
c.  Inhibitory
d.  Saddle-shaped
Figure 4.2  Neural selectivity and tolerance to a vocalization feature model

Assume neurons utilized maximal bell-shaped coding strategy for vocalizations. Neural response projected to one feature. Two red lines bounded the natural vocalization distribution within 2 standard deviations (std).

a. Neuron with higher selectivity (black) and neuron with lower selectivity (blue) to a vocalization feature.

b. Neuron with higher tolerance (blue) and neuron with lower tolerance (black) to a vocalization feature.
Figure 4.3  Neural discriminability for call categories model

Neuronal response projected onto 2D feature plane. Each ellipse represents receptive field of a neuron.

a. Receptive field of each neuron is large. Neural discriminability for call categories based on feature distribution is small.

b. Receptive field of each neuron is smaller, and more selective to features. Neural discriminability for call categories based on feature distribution is larger.
Figure 4.4  Topographic map in right hemisphere

Each colored dot represents a neuron. Color of the dot represents the best frequency of that neuron in response to puretone. Neurons not responsive to puretones were not shown here. Recorded neurons were mostly located in A1 and CM/CL.
Figure 4.5  Examples of neural response to call categories

Examples of three neurons responsive to different call categories and their characteristic properties responding to puretone and noise are shown. Each stimulus has five repetitions delivered in random order during experiments. Raster plots are shown in left column, and rate plots in right column.

10 call categories were chosen from our colony recordings. Odd stim number represents the natural real vocal token for that call category, while even stim number represents the time-reversed vocal token for that call category. Call categories are listed below. First four stimuli represent simple calls, and the rest represents compound calls.

Stim #1: Phee; Stim #3: Trill; Stim #5: Trilphee; Stim #7: Twitter; Stim #9: dTwitter; Stim #11: Tsikbark; Stim #13: Tsikstrg; Stim #15: Tsikstrg; Stim #17: Trtwpp; Stim #19: Peepstrg

a. M6xU124. Neuron responsive to few call categories and broadband noise, but not to puretones and narrowband noise.

b. M6xU135. Neuron responsive to more call categories, broadband noise, and narrowband noise (onset response), but not to puretones.

c. M6xU135. Neuron responsive to several call categories and broadband noise (sustained response), but not to puretones and narrowband noise.
Figure 4.6  Examples of spatial effect on neural response to vocalizations

Examples of three neurons responsive to vocalizations delivered from different spatial speakers are shown. Each stimulus has five repetitions delivered in random order during experiments.

Spatial receptive field is obtained without doing any interpolations to avoid the bias from low repetition rates. Three values on the colorbar represents average firing rate, firing rate above one standard deviation from average, and below one standard deviation from average, respectively. There are 32 speakers in total, and the front view of the monkey M6x is shown with the shaded gray area. Negative azimuths represent contralateral side, while positive azimuths represent ipsilateral.

10 call categories were chosen from our colony recordings. Detailed descriptions for these call categories are shown in Figure 4.5 notes.

a. M6xU126. Neural response to call categories delivered from different spatial speakers are compared.

b. M6xU128. Neural spatial receptive field obtained from different call categories are compared (contralateral response).

c. M6xU127. Neural spatial receptive field obtained from different call categories are compared (ipsilaterial response).
b

Broadband Noise (M6xU128)

Tsikbark (M6xU128)

Trill (M6xU128)

Azimuth (degree)
Elevation (degree)
Neural selectivity to vocalizations could come from the sensitivity to the background noise in some cases. An example neuron (M6xU180) is shown.

a. Neural response to real Trill call delivered from different speakers.

b. Neural response to synthesized Trill call delivered from speaker 11 (contralateral) at different sound attenuated levels.

c. Neural response to synthesized Trill call with different background SNR. 0 dB represents call totally masked by broadband noise. 100 dB represents call without background noise. Cut-off threshold is at around 36 dB from interpolation.
Figure 4.8 Examples of neural response to features in synthesized calls

An example neuron (M6xU214) is shown with its response to simple acoustic stimuli and features in synthesized vocalizations.

(a.) Basic characteristics of the example neuron. Best frequency is 18.379 kHz. Stimuli were delivered with center frequency 18.379 kHz.

(b. - j.) Neuron tuned to different features of synthesized Trill call. For features that were not varied in each figure, they were fixed at the value of an average feature of Trill calls based on Table 3.4. For features that were varied, definitions are described in Figure 3.8 notes. Neuron is invariant to changes in individual feature parameters, and raster plots shared similar temporal patterns.
e. Synthesized Trill (M6xU214)

f. Synthesized Trill (M6xU214)

g. Synthesized Trill (M6xU214)
Figure 4.9  Examples of temporal coding for synthesized vocalizations

Neural response to a simple continuous call category could use temporal coding instead of rate coding. Two example neurons are shown. Left column is neural response to puretones at 50 dB from preferred spatial location. Right column is neural response to synthesized calls with varied acoustic feature parameters.

d. M6xU4. Neuron tuned to synthesized Phee call with different slow frequency modulation depth (FMMod).

e. M6xU10. Neuron tuned to synthesized Trillphee call with different time fractions of trill-to-phee transition point (tTrans)..
Neural tuning to co-varied FM/AM rate of Trill showed U-shaped coding instead of bell-shaped coding strategy. The natural distribution of Trill FM/AM rate within two standard deviations is between 20 and 40 Hz (reference from Table 3.4), which happens to evoke least neural response. An example neuron (M6xU4) is shown.

Figure 4.10  Examples of U-shaped coding for synthesized vocalizations
References


Beecher MD. and Stoddard PK. (1990). The role of bird song and calls in individual recognition: Contrasting field and laboratory perspectives.


Chia-Jung Chang

Email: chiajung.chang@jhu.edu
Web: www.cjchang.weebly.com

Laboratory of Auditory Neurophysiology
Department of Biomedical Engineering
Johns Hopkins University, Baltimore MD

EDUCATION

9/2012-8/2014  M.S.E. in Biomedical Engineering, Johns Hopkins University, Baltimore, MD
  • Thesis Advisor: Dr. Xiaoqin Wang

9/2008-6/2012  B.S. in Electrical Engineering, National Taiwan University, Taipei, Taiwan
  • Overall GPA: 3.98/4.00
  • Presidential Award, Dean’s List

RESEARCH EXPERIENCE

9/2012-8/2014  Auditory Neurophysiology Laboratory, Johns Hopkins University
  - Advisor: Dr. Xiaoqin Wang
  • Designed a virtual vocalization panel to systematically study the neural selectivity and tolerance of conspecific vocalizations
  • Obtained cortical neural responses to complex tone stimuli with single-unit recording in awake marmoset to explore the hierarchical object processing pathway
  • Implemented automatic speech recognition to categorize vocalizations and analyze acoustic features to investigate the vocal plasticity interaction between two groups of marmosets originally from different colonies

  - Advisors: Dr. Krishna Shenoy and Dr. Teresa Meng
  • Analyzed neural variability of motor preparation across trials and within a single trial
  • Implemented a FPGA-based brain machine interface for a wireless neural recording system for rhesus monkeys

2/2011-9/2011  Nano Engineering and Micro Systems Laboratory, National Tsing Hua University
  - Advisor: Dr. Chao-Min Cheng
  • Studied the dynamics of individual polymers using micro-curvilinear flow
  • Analyzed spatially localized nanopatterns on glass surface by modifying the filament orientation with myosin-II
PUBLICATIONS & PRESENTATIONS

- Journal Article

- Presentations


HONORS & AWARDS

2012  Google Anita Borg Memorial Scholarship
      - *Google, Inc.*
      • Recognizes female engineers with excellent academic performance and leaderships

2012  Irving T. Ho Fellowship
      - *Irving T. Ho Memorial Foundation*
      • Awards 3 undergraduates with excellent performance in electrical engineering in Taiwan

2012  Presidential Award
      - *National Taiwan University, Taiwan*
      • Recognizes top 5 % students

2011  Stanford Summer Undergraduate Research Fellowship
      - *Stanford University*
      • Awards 18 selected visiting undergraduates from Taiwan and China

2009  Gold Medal, 3rd Asian Science Camp Poster Session
      - *The High Energy Accelerator Research Organization (KEK), The International Center for Elementary Particle Physics (ICEPP), and Heisei Foundation for Basic Science, Japan*
      • Our team proposed three innovative ways to cure cancer in poster session