Detecting Vocalizations of Individual Monkeys in Social Groups

Alireza Bayestehtashk[1], Izhak Shafran[1], Kristine Coleman[2] and Nicola Robertson[2]

***Abstract*— Vocalization is an important clue in recognizing monkeys' behaviors. Previous studies have shown that the frequencies, the types and the lengths of vocalizations reveal significant information about social interactions in a group of monkeys. In this work, we describe a corpus of monkey vocalizations, recorded from Oregon National Primate Research Center with the goal of developing automatic methods for recognizing social behaviors of individuals in groups. The constraints of the problem necessitated using tiny low-power recorders, mounted on their collars. The recordings from each monkeys' recorder nonetheless contains vocalizations from not only the monkey wearing the recorder but also its spatial neighbors. The devices recorded vocalizations for two consecutive days, 12 hours each day, from each monkey in the group. Like in sensor networks, low power recorders are unreliable and have sample loss over long durations. Furthermore, the recordings contain high-levels of background noise, including clanging of metal collars against cages and conversations of caretakers. These practical issues poses an interesting challenge in processing the recordings. In this paper, we investigate our automated approaches to process the data efficiently, detect the vocalizations and align the recordings from the same sessions.**

I. INTRODUCTION

Current approaches for observing the animal behaviors completely depend on human observation. A highly trained observer watches the animals in the group and records the occurrence or duration of the behaviors listed on an ethogram (a set of behaviors with their quantitative descriptions) [1]. There is a wide range of behaviors such as aggression, displace, fear grimace, lipsmack, scream, grunting etc. that can be used in studies of social behaviors [2]. Human observation has two major limitations: First, feasible ethograms are limited to a small subset of behaviors since the rate of analyzing the data and its accuracies drop when an observer annotates more behaviors. Second, it is impossible to annotate all behaviors of every animal in a group in a single pass. In practice, the observer is forced to go through the data multiple times and in each pass, annotate a specific behavior of all animal or a particular individuals' activities. In addition, the behaviors with auditory modality such as barking, cooing and grunting are difficult and time consuming for human observers to annotate.

Having an automated method for observing and modeling the social activities could lead to a better understanding of behaviors of social animals and open up new directions for researchers in behavioral ecology, anthropology, evolutionary psychology, conservation biology, and neuroscience.

[1]A.Bayestehtashk and I.Shafran are with the Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA bayesteh_ar at yahoo.com

[2]K.Coleman and N.Robertson are with the Oregon National Primate Research Center, Oregon Health & Science University, Portland, OR, USA

In this paper, we describe a corpus of vocalizations that was collected with the goal of developing automatic methods for recognizing behavior of individuals monkeys' in social groups. In Section II, we describe this corpus along with the challenges posed in processing them. Subsequently, we delve into different components of our analysis. In Section III-A, we address the issue of background noise in the recordings. From the cleaned signal, we extract segments with potential vocalizations, compute feature vectors from them and classify them. Then, in Section III-C, we examine the problem of aligning the vocalizations from different recordings so that we can identify which monkey vocalized when. Finally, we highlight the steps that we found most effective in processing these recordings.

Fig. 1. *A group of monkeys in Pen.*

II. THE CORPUS OF RHESUS MACAQUE VOCALIZATIONS

Our corpus consists of audio and video recordings of social behaviors of groups of rhesus macaques. The study and the data collection was approved by OHSU's Institutional Animal Care and Use Committee. Groups of 4-6 animals were formed, introduced into the pen, which is about 12 ft long, 7 ft deep and 7 ft tall as shown in Figure 1, and observed over a period of about 2 months. We recorded behavior as the group settled into their stable social hierarchy. After approximately two weeks, we perturbed the social hierarchy of the groups using standard procedures such as presence of an unfamiliar human (outside the cage), and introduction of toys and desirable food. The observations were performed to minimize the disruption of animal care and husbandry. This meant swapping the spent audio recorder, housed in their collars, with a fully charged one on a specified day of the week. The recordings were performed till about 7pm on the same day and between about 7am and about 7pm the

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

subsequent day, corresponding to the hours when the lights remained on. In all, 80 such sessions were recorded from 5 different groups.

Video recordings were captured by three cameras mounted on three different corners of the pen and one fisheye-lens camera mounted on the ceiling. All four cameras were fully synchronized in the frame level and their frame rate was controlled by an external trigger to be exactly 12 fps. The mounting locations of cameras were carefully chosen to support 3D reconstruction of the observation sessions and maximize the coverage of the visible space in the cage.

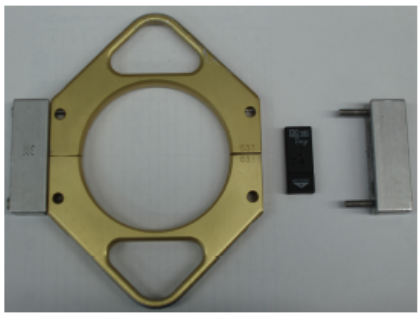

Fig. 2. *Tiny low-power audio recorder along with its housing that attaches to the monkey's collar*

Audio was recorded using tiny recorders, EDIC B21, which is about 40 x 15 x 10 mm in dimension, 8g in weight, and has a battery life of 2-3 days. These recorders were placed in a custom housing that was attached to a standard collar, as shown in Figure 2. Each recorder was programed to record 12 hours at 8 kHz sampling rate for each session. Unlike the video recordings, the audio recordings could not be synchronized via hardware or other means. Our calibration attempts using chirp signals show that the asynchrony is erratic and not easily predictable such as a constant offset or a linear drift. In all, we have about 3800 hours of audio recordings.

III. The Problem and Our Approach

There are several challenges in processing the above mentioned audio recordings and this paper addresses them.

1) High background noise: Monkeys move about such that their collars hit the walls and metal mesh of the pen. In addition, the recordings also contain the conversations of human caretakers.
2) Multiple speakers: Even though each monkey has a separate collar-mounted recorder, the recordings contain vocalizations from its neighbors. So, attributing which monkey spoke when is non-trivial.
3) Sample dropout: The recorders appear to lose sample randomly over the course of the long 12 hour recording sessions. This is similar to the problem that occurs in unreliable low-power sensor networks and complicates the problem of aligning the recordings which is necessary for identifying which monkey vocalized when.
4) Length of recordings: The sessions are about 12 hours long, which makes it infeasible to apply conventional solutions such as dynamic programming to align waveforms.

Given the amounts of data, the first step was clearly eliminating the segments with very low probability of vocalizations. From listening to several random examples and from preliminary experiments, simple methods based on energy or spectral entropy were confounded by large amounts of background noise. So, before we could eliminate segments without vocalization, we had to improve the signal to noise ratio. After enhancing the signal and removing unvocalized segments, we were able to achieve high accuracies in detecting vocalizations using a supervised classifier fairly easily. In contrast, without the signal enhancement, the supervised classifier was unusable. Having identified the vocalized segments, we aligned the recordings by just focusing on these segments containing high signal-to-noise ratio. This improved the quality of alignment compared to aligning with portions that included background noise without vocalizations. Below, we describe each of the steps in more detail.

A. Signal Enhancement and Candidate Segments

The pen housing used for collecting the corpus is part of a bigger laboratory, which was not designed for high quality audio recordings. The infrastructure including ventilation and lighting introduced a significant amount of background noise. The walls are acoustically reflective and not dampened in any way, causing significant reverberations.

The recordings contain two sources of additive noise – a significant amount of background noise that was largely constant in nature, on top of which there were bursts of metallic clangs from different distances. Knowing that the first component is a good candidate for signal enhancement techniques, we applied noise spectral subtraction.

Noise spectral subtraction is a simple and computationally efficient method for reducing the background noise and enhancing the audio. It is a nonparametric method and has two major steps. The first and the more important step is to estimate the background noise. The more sophisticated techniques locate a segment in the recording which contains only noise. Simpler approaches typically assume the initial few milliseconds are noise and estimate the background from it. We were interested in quickly characterizing the potential benefit of this simple technique, so we resorted to the implementation in Audacity [5], where the user needs to manually choose an appropriate segment containing noise, from which a noise profile is created. The noise profile simply consists of a set of statistics like maximum for each frequency bin in Discrete Fourier transform(DFT) computed across all the frame of noise segment. The second step uses the noise profile to attenuate the power spectrum of the parts of signal that are similar to the noise and leave the rest unchanged. Finally frequency-smoothing and time-smoothing are applied to produce a natural sound and prevent rapid changes in the gain of the output signal.

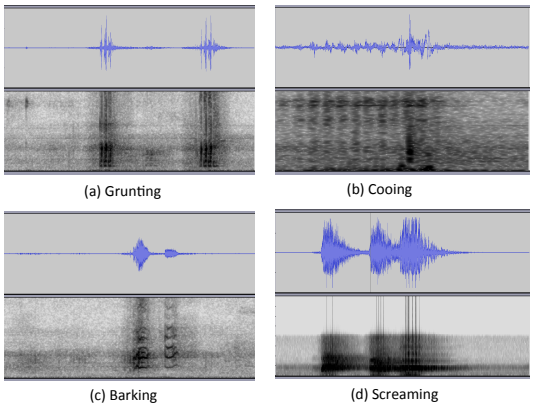

Fig. 3. *Different types of monkey vocalizations*

This simple approach was remarkably effective. After signal enhancement with spectral subtraction, it was relatively easy to filter out unwanted segments which contained only silence or background noise and no vocalizations. This was useful in reducing the size of the data significantly. Energy-based segmentation is simple and computationally efficient method for removing such segments. All segments below -35 db were removed. This reduced the corpus by a factor of 10 and made it feasible to process the data using the next few steps.

B. Detecting Segments Containing Vocalizations

The candidate segments extracted from the previous step contains three types of audio – the vocalizations from the monkeys, the bursty noises such as metal clangs, and the human conversations. Human interference is unavoidable in the standard animal laboratory setting since animal husbandry requires mandatory routine checks, multiple times a day, by the staff to feed and monitor them.

The difficulty in isolating the human conversations is that monkey vocalizations vary largely depending on the type (e.g., grunting, cooing, barking and screaming), as illustrated in Figure 3. We manually checked a 12-hour recording from one monkey several times and carefully annotated all segments as belonging to the monkeys and humans. From this, we created a balanced data set of about 1200 segments.

For each segment, we extracted a fixed dimension feature vector using OpenSmile [3], a standard feature extraction tool that extracts a rich set of features for each segment. Briefly, the toolkit extracts features in two steps. First step is extraction of 25 msec long frames using a Hanning window at a rate of 100 frames/sec and computation of frame-level features such as RMS, MFCCs, ZCR, voicing probability, F0 and their deltas. The second step is aggregating frame-level features into the segmental feature vector by applying statistical functions such as mean, median, variance, minimum and maximum across all frame-level features of a segment. We extract about 400 features our labeled segments. We utilized standard supervised classifiers and compared three methods – support vector machines, K- nearest neighbors and Gaussian mixture models.

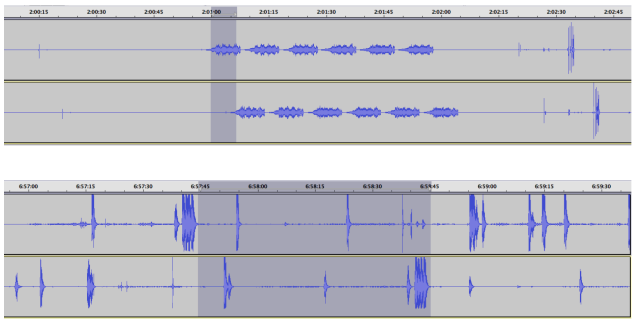

Fig. 4. *The lag between two recordings (Top) about 6 seconds in the beginning of a session (Bottom) about 60 seconds in the middle of a session.*

C. Diarization and Solving the Alignment Problem

The larger goal of this study of quantifying monkeys' social behaviors requires identifying which specific individual monkey vocalized. The recordings from each monkey, as mentioned before, contains vocalization from nearby monkeys too. A simple energy threshold is insufficient since the monkeys vocalize with wide dynamic range including soft coos. The identification of the monkey that is vocalizing at a specific time will require aligning all the recordings from a group and finding the one with the maximum energy at the given time.

The alignments, as mentioned before, are not merely a constant offset. Figure 4 illustrates an example from our corpus where the offset is different in the beginning and in the middle of the recording.

A straightforward dynamic time warping (DTW) for aligning audio from two recorders would be infeasible for 12 hours of audio due to the computational complexity. The DTW backtrace matrix is $O(N_1 N_2)$ for sequences of lengths N_i[4]. At 8KHz sampling, that would require about 1.2 petabytes!

The noise in the recordings poses another problem. Preliminary experiments with multiple sequence alignments produced inconsistent results. Specifically, when 3 streams A, B and C were tested, the sum of delays between A-B and A-C did not match that from A-C.

Our approach is to align detected vocalized segments, instead of aligning samples. After discarding the non-vocalized segments and those related to human conversations, the scale of the alignment problem decreases dramatically as the corpus reduced by a factor of 80. The average number of segments for 12 hours recordings is about 1.5K, which becomes feasible for solving with DTW. Additionally, based on observations of random samples, we also constrained the maximum allowable time delay between the streams to be about 5 minutes. This reduces the DTW search space to a banded diagonal matrix, speeding up the computation considerably.

The noise in the vocalized segments and the clipping of the waveforms, however, introduces considerable jumps in

the DTW solution. We applied a modification to the warping cost to introduce a penalty to avoid rapid changes. This smoothness constraint improves the alignment considerably in practice.

Let $S_1 = \{S_1(1), \ldots, S_1(N))\}$ and $S_2 = \{S_2(1), \ldots, S_2(M))\}$ be two sequences of vocalized segments and each segments $S_k(i)$ starts at time $t_{k,i}$. The warping path is computed as follows:

$$F(i,j) = dist(S_1(i), S_2(j)) + \min \begin{cases} F(i-1, j-1) + \phi(|t_{1,i-1} - t_{2,j-1}|, |t_{1,i} - t_{2,j}|) \\ F(i-1, j) + \phi(|t_{1,i-1} - t_{2,j}|, |t_{1,i} - t_{2,j}|) \\ F(i, j-1) + \phi(|t_{1,i} - t_{2,j-1}|, |t_{1,i} - t_{2,j}|) \end{cases}$$

Where the distance between two segments is measured based on Normalized Cross-Correlation(NCC):

$$dist(S_1(i), S_2(j)) = 1 - max\{NCC(S_1(i), S_2(j))\}$$

And the penalizing term is quadratic function [6]:

$$\phi(x, y) = \frac{|x - y|^2}{2 * \sigma}$$

IV. Experimental Results

A. Classification Task

In this section, we compare the performance of several binary classifiers in detecting monkey vocalizations using features described earlier in Section III-B. The classifiers we compared includes K-nearest neighbor classifier, GMM-based classifier and support vector machines (SVM). For the SVM, we investigated two different types of kernel, namely, the radial and the polynomial basis functions. Using the manually labeled 12-hour session, the parameters of the classifiers were tuned using a grid search on a 5-Fold cross validation over train set and evaluated over the held-out set. Results, reported in Table I, show that the SVMs with polynomial kernel work best for this task.

TABLE I

The performance (accuracy) of different classifiers in detecting segments with vocalization from the monkeys.

Method	Ave. Accuracy	Std. Accuracy
K-NN	78.2	5.6
GMM	83.4	4.2
SVM-poly	85.1	3.8
SVM-rbf	88.9	3.3

B. Alignment Task

1) Verification Against Manual Alignments: We converted two audio recordings from one session into two sequences of vocalization segments and manually aligned them. The Figure 5(a) shows the comparison of manually aligned data and the alinement obtained using DTW. The results show the DTW can mostly track the actual path but it still suffers from spurious picks. Our experiments show a median filter removes the picks and results in a promising alignment curve as shown in Figure 5(b).

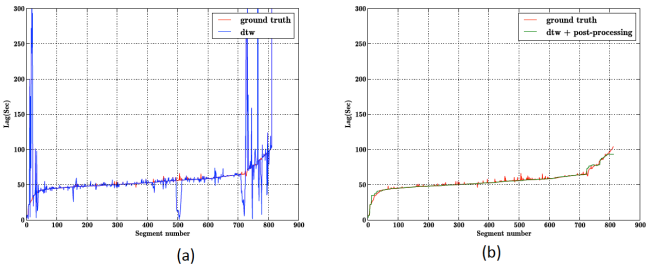

Fig. 5. *Alignment curves where x-axis and y-axis denote time in terms of segment index and the lags in secs respectively. Red and green lines mark ground truth and estimated lags respectively. Subfigures show: (a) our method as described in Section III-C, and (b) post-processed with a median filter.*

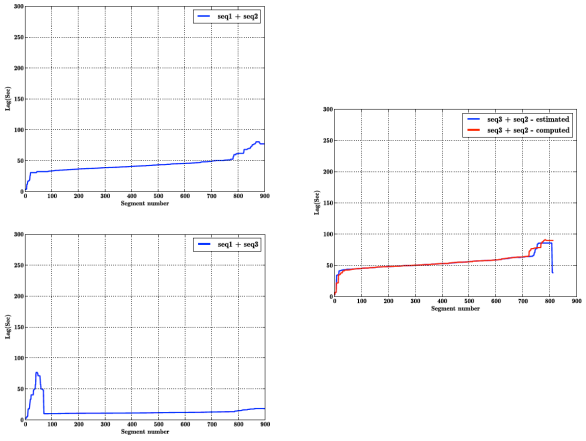

Fig. 6. *Pairwise agreement between two alignment curves. The left subfigures show alignments for two different pairs (A-C and B-C) while the right one shows two graphs one computed directly from waveforms (A-B) and the other inferred indirectly from (A-C) and (B-C).*

2) Consistency Verification: Manual verification of alignment is enormously labor intensive and not scalable. Alternatively, we can check consistency without needing any manual alignments. Alignments between two waveforms (A-B) can be computed directly or inferred from alignments of both waveform with respect to a third waveform (A-C and B-C). Figure 6 illustrates an example of pairwise agreement. In our experiment, the percentage of estimated alignments with coarse disagreement (deviation of more than 10%) is about 3.4%, which demonstrates that our method is effective.

V. Conclusions

In this paper, we have addressed the problem of processing noisy recordings obtained from low power unreliable recorders mounted on collars of monkeys, in the context of observing their social behavior in groups. The challenges in processing these data include high background noise, vocalizations from multiple monkeys in each recordings, randomly distributed missing samples, and long recordings of up to 12 hours. We developed a pipeline with three stages to address these challenges. First, a signal enhancement stage was used to remove the background noise using spectral subtraction, which was found to be very effective.

From the resulting waveform, we were able to trim out segments without any candidate vocalizations, thus reducing the data by factor of 10. Second, we devised a supervised classifier to detect segments with vocalization. This classifier was remarkably effective with an accuracy of about 89%. Finally, we reformulated the problem of aligning 12 hour long waveforms into aligning about 1.5K segments, reducing the problem by a factor of 80. We demonstrate that our alignment algorithm works as well as labor-intensive manual alignments and has good pairwise consistency.

REFERENCES

[1] Altmann, J."Observational study of behavior: sampling methods", Behaviour, 49(3):227267, 1974. Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10, 2010.

[2] Maestripieri, D., and Wallen, K., "Affiliative and submissive communication in rhesus macaques", Primates, 38(2):127138, 1997.

[3] Eyben, F., Wollmer, M. and Schuller, B., "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10, 2010.

[4] Sakoe, H. and Chiba, S.,"Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43 49, ISSN: 0096-3518, 1978. Elsevier, 1991.

[5] Mazzoni, D., and Dannenberg, R. . "Audacity [software]", Pittsburg, PA: Carnegie Mellon University, 2000.

[6] Asgari, M., Shafran, I. and Bayestehtashk, A., "Robust Detection of Voiced Segments in Samples of Everyday Conversations Using Unsupervised HMMs", Proc. IEEE Spoken Language Technology (SLT), 2012.