Automated detection of mouse scratching through audio recording

Peter Elliott (Max G'Sell, Valerie Ventura)

Draft/Outline

1 Introduction

`This is not a true introduction but may be a helpful reference`

The paper I will be writing will be about methodology for automated detection of scratching by caged mice. I have worked on a method to take an audio recording of a mouse and return labeled time points when scratching occurs. The paper will need to explain why this work matters: why record scratches and why bother with automation? The paper will also need to explain how my method works and how well it works. Since the work is primarily of interest to non-statisticians, many of the techniques I've used will likely be unfamiliar to much of the audience.

The ultimate goal of the project is to aid efforts to research itch and pain. These sensations are often studied through behavioral observation. In the case of itch, mice may be injected with a chemical agent that induces itch then scratching behavior is observed over time. Quantity of scratching is used to assess how much itch the mouse felt (unfortunately, they can't tell us directly). This can be used to measure the effect of various interventions, such as genetic modifications to the mice or skin treatment. This observation helps us understand the mechanisms by which itch and pain work. Improved understanding and the ability to test interventions may eventually lead to improved treatment options for humans.

Currently, observations of mice are typically done manually. That is, someone has to sit, watch and count. As a result, the sample sizes collected are often very small. Furthermore, often only a small fraction of the mouse's response to induced itching is observed - the effect of the chemical agent continues long after observation stops. These problems would be fixed by automated observation. This would allow many mice to be observed simultaneously and would allow for extended observation times with negligible extra expense to the scientist's time.

The method the paper will propose uses audio data to automate the observation process. The method requires that a single mouse be placed in a sound-proofed cage with a microphone for the duration of the experiment. The resulting audio

data is analyzed using a two-stage process: first segmentation into candidate time intervals then classification. The segmentation process makes use of scratching's characteristic rhythmic pattern. Time intervals with regular peaks in energy are marked as candidates. The classification is done using a random forest trained using features we have designed. These features attempt to encode both time and frequency (i.e. audio pitch) information. We will discuss the design of these features.

The method was developed using data collected by the Ross Neuroscience Lab at the University of Pittsburgh, and we use that data to assess the accuracy of our method. The classifier is not perfect, so the statistical implications of mislabeled scratches will need to be discussed.

2 Data

At which university or lab?

This might be a matter of personal opinion, but it might help to make this be in active instead of passive form. I say that because I'm a little confused thinking about who did what. Who injected the mouse, who analyzed the audio/video, and who is "us"?

The raw data come from an audio recording of a lone mouse in a sound-proofed cage. Before recording, the mouse was injected at its neck with an itch-inducing agent. The cage was sound-proofed to minimize audio contamination from outside sources of noise. The recording used two microphones, placed in different corners of the cage. A camera was placed at the top of the cage. While our goal is to perform automated labeling using audio, the video was used to manually label the data.

Mouse scratching occurs in small sets of rapid swipes, which we refer to as *scratch bouts*. The goal of the procedure is to detect when these scratch bouts occur, rather than labeling individual scratches. An example of a scratch bout is shown in Figure 1. The pattern of scratching can be seen in both the waveform and spectrogram. Scratches within bouts tend to occur with a period of around 50 ms. A full scratch bout tends to last between 200 and 500 ms but can be as long as a full second.

3 Methodology

Maybe don't use the word "obviously"?

The proposed procedure for scratch bout detection has two main steps. First, the recording is segmented to choose time intervals that are candidates to be scratch bouts. This segmentation procedure helps to remove from consideration time periods that are obviously not scratch bouts. It also helps to ensure alignment between labeled intervals and scratch bouts. In the second step, a classifier is used to label the candidate time intervals based on extracted features. The classifier allows us to discrminate in cases where simple rules are insufficient. `This may seem like an obvious point, but previous methods don't use a classifier`

3.1 Segmentation

In the first stage of processing, we perform two tasks. We remove from consideration any time period without any scratch-like pattern. We also align time intervals to be

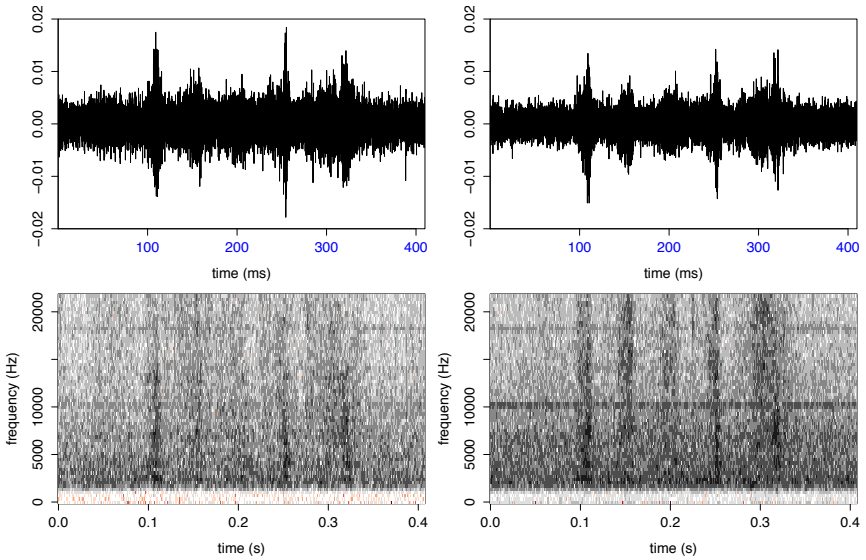

Figure 1: An example of a scratch bout. The top row shows the audio waveforms in the time interval for each of the two microphones. The bottom row shows spectrograms of the time interval corresponding to the waveforms. The time-frequency representation given by the spectrograms provides a useful basis for extracting relevant information.

classified with scratch-like observations. As this is only the first stage, the removal process is tuned so that the vast majority of scratch bouts pass, even at the cost of many passing false discoveries.

"Pass", as in pass the test? Maybe you could say "are accepted"?

Our strategy for segmentation relies on the observation that the series of scratches within a scratch bout creates a characteristic rhythm. The audio from a scratch bout should show peaks in signal power with a period around 50 ms. Exploratory analysis of the data also reveals that sound caused each swipe occupies a broad frequency band. This peak in power across a broad range of frequencies can be seen in the spectrograms in Figure 1. Additionally, while lower frequencies are often contaminated by other sources of sound, higher frequencies tend to uncontaminated. The segmentation process therefore locates time points with strong power in higher frequencies and checks that these peaks in power occur close to 50ms apart.

Segmentation goes as follows:

1. Create a spectrogram from the audio recording. We use time bins of 128 samples (approximately 3ms) with an overlap of 96 samples and a Gaussian windowing function (`reference may be needed here`).

2. For each time bin, sum the modulus of Fourier coefficients corresponding to frequencies over 10kHz. This measures the power in the frequency range above 10kHz.

Why are there two columns? I see that the top and bottom rows are spectrograms and integrated power. But I don't see why you need two of them.

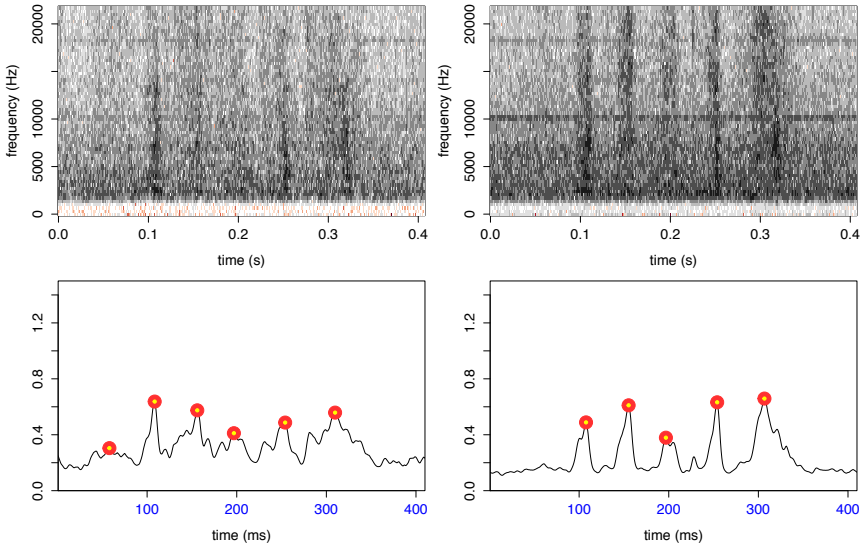

Figure 2: An example of a scratch bout. The top row shows spectrograms of the time interval for each of the two microphones. The bottom row shows the integrated power above 10 kHz, smoothed over time. Detected peaks are marked in red. This bout shows a consistent period of about 50 ms. There is some variation in other characteristics of the peaks.

3. Apply a triangular kernel smoother to the time series of power measurements. The kernel shape matches the expected shape in the time series of peaks caused by scratches. We use a bandwidth that corresponds to approximately 8ms.

4. Find time points in the smoothed time series that (a) have maximal power within 25ms, and (b) have power greater than the minimal time bin within 25ms by some chosen parameter h.

5. Group sequences of time points found in part 4 that have between-adjacent-point distances under 120ms. Groups that include at least three time points are marked as candidate time intervals.

An example of the transformation used for segmentation is shown in Figure 2. The top row displays the spectrograms obtained in step 1. The bottom row displays the smoothed time series of power measurements obtained in step 3, along with the peaks obtained in step 4. As the peaks are less than 120ms apart, they are grouped into a candidate time interval.

Most of the parameters used are a reflection of exploratory analysis. The 10kHz cutoff is where sound contamination starts to be rare. The bandwidth of the kernel smoother is chosen to mirror the typical duration of a swipe. The 25ms range in which a time point must be a maximum is a reflection of the minimum amount of

time there can be between swipes in a scratch bout. The 120ms distance is used to avoid splitting a scratch bout in the event that one of its swipes is not detected.

Sensitivity of the segmentation process is controlled by the choice of peak threshold h in step 4. Higher values of h require that time bins be more pronounced maxima to be considered as candidate swipes, and therefore fewer candidate time intervals will be chosen. An appropriate value for h can be chosen using grid search with a training set.

3.2 Classification

After segmentation, we use a classifier to label the selected candidate time intervals. Labeling the candidate time intervals requires the design of features that capture important characteristics of the data as well as the selection of a classification method capable of distinguishing between the class-conditional distributions. Many of the features we use try to find and describe rhythmic patterns in the data. We also make use of audio pitch (frequency) information via the Fourier transform.

Our choice of classifier is a Random Forest (Breiman, 2001). The benefits are discussed in Section 3.2.2.

3.2.1 Feature design

Recall that for the purpose of segmentation, the waveform of the audio was transformed in steps 1-3 to give a new time series representation containing useful information. There are other useful transformations that can also be made. Thee different transformations allow us to describe different characteristics of the data. We can extract common features from these transformations to get a large body of potentially informative features.

Transformations

We use two main types of transformations. The first involves convolution of power measurements across time with a template. The transformation used for segmentation (described in steps 1-3 in Section 3.1) is an example - in that case the triangular kernel acts as the template. The second type of transformation uses localized frequency information. We compute a spectrogram for the time interval and calculate a correlation or inner product with a periodogram template at each time bin. The transformations we use are as follows:

1. Convolve the power above 10kHz with a triangular kernel.
2. Convolve the power above 15kHz with a triangular kernel.
3. Convolve the power above 20kHz with a triangular kernel.

4. Calculate the correlation of the spectrum at each time point with an average of periodograms drawn from a sample of scratches.

5. Calculate the inner product of the spectrum at each time point with an average of periodograms drawn from a sample of scratches.

I'm using spectrum and periodogram sort of interchangeably which is probably bad

There are a couple other transformations that need to be described

Common features

For each transformation, we wish to extract a set of features. For scratch bouts, we expect to again see a rhythmic pattern of peaks in the transformed time series. The features we extract should describe the set of peaks found in the candidate time interval. Peaks in the transformed time series can be found following step 4 in the segmentation process. We then extract the following features:

1. The number of peaks

2. The average time between peaks

3. The standard deviation of times between peaks

4. Summary statistics (mean, median, minimum, maximum, and standard deviation) of the powers at each peak

5. Summary statistics of the full width at half maximum of each peak

6. Summary statistics of the transformed time series itself

In some cases, the feature to extract is not defined. For example, we cannot calculate the standard deviation of the times between peaks if the transformed time series only has two peaks. In these cases, the feature can be coded as -1. There might be a better place to mention this

Miscellaneous features

- time
- large-binned spectrograms and frequency distributions

3.2.2 Random Forests

Random Forest classifiers allow us to handle several difficulties. The feature extraction process gives us a high dimensional feature set, with many that are highly correlated. There are also potentially many complex interactions between features. The decision boundary for classification is likely nonlinear. These are all characteristics that are well-handled by Random Forests.

4 Results

- `accuracy measures (sensitivity, FDR, ROC?)`
- `analysis of where in the process errors occur (segmentation vs classification)`
- `comparison to previous methods`
- `stereo recording vs mono?`
- `comparison of classifier choices?  (RF, logistic lasso)`
- `should evaluate robustness to parameter choice`

5 Discussion