Submitted: 01/22/21

Research Article

[1]Faculty of Science, McGill University, Montreal, QC, Canada

Keywords

Pharmacology, ultrasonic vocalization, convolutional neural network, rat, call classification

Email Correspondence

samir.gouin@mail.mcgill.ca

Samir Gouin[1]

Automated ultrasonic vocalization analysis: Training and testing VocalMat on a rat-based dataset

Abstract

Background: Ultrasonic vocalizations (USVs) offer another way to study the behaviour of rodents in addition to commonly used visual methods. USV subtypes have been associated with behaviour such as the concurrence of 22-kHz calls and signs of distress (defensive behaviour). (1,2) However, the categories used to analyze USVs are a source of contention, most notably with 50-kHz calls, and may even be arbitrary altogether. (3) To facilitate subtyping calls, VocalMat has been developed for USV identification and classification, and it has shown an accuracy of greater than 98% for mice USV detection and 86% for mice USV classification. (4) In this project, we have constructed a rat-based dataset of USVs and then used it to train the VocalMat program to assess automated USV classification.

Methods: Avisoft-SASLab Pro was used to manually classify USVs from 216 audio files. The sorted USVs were then used to train VocalMat's classification program.

Results: Our results show overall accuracies greater than 90% with the highest in the trill and flat categories (97.2% and 91.0%). We experimented with the number of USV categories and found high accuracies when grouping spectrographically similar calls, which are flat calls with up and down ramp calls (96.9%) and trill calls with trill jump and flat-trill calls (98.7%).

Limitations: There are large variations in the number of calls per category in our dataset. More data is needed to fill these gaps and provide more training samples for infrequent calls.

Conclusions: By creating a database of rat USVs and then using it to train VocalMat, we have shown the potential of its adaption to a rat vocal repertoire. Going forward, we hope to test more variations of USV categories on machine learning programs to establish a robust approach to classifying USVs.

Introduction

Ultrasonic vocalizations (USVs) offer perspective into the social and emotional states of rodents. (5) Rodent USVs are composed of syllables collectively ranging from 20-110 kHz, a range higher than human perception. USVs have distinctive acoustic features including frequency, duration, and pitch. The analysis of these acoustic features and spectrographic shapes allows USVs to be categorized into subtypes. USVs have been associated with behaviour, such as the concurrence of 22-kHz calls and signs of distress, (2) and are thought to play an important role in affective signaling, social communication, and thermoregulation. (5) USV analysis can be used to study the effects of drugs on rodents by monitoring changes in calling during experiments, making it an important tool in pharmacology research. (6-8)

Laboratory rats make three broad types of calls: pup calls, 22-kHz calls, and highly heterogenous 50-kHz calls. The 22-kHz and 50-kHz labels are approximations, as in reality the frequencies of these calls vary from 18-32 kHz and 35-72 kHz respectively. (9) There appear to be spectral graphic similarities between mice and rat USVs. However, mice lack 22-kHz and trill call equivalents. (5) Moreover, there are spectral differences between rat and mice USVs, such as shorter short calls and greater modulation of frequency in upward ramp and downward ramp calls in mice. (4, 10) The categories used to analyze USVs are a source of contention, most notably with 50-kHz calls, and may even be arbitrary altogether. (3) Regardless, USV analysis has value as a non-invasive method to monitor changes in rodent behaviour.

Categorization schemes for rat 50-kHz USVs have evolved and increased in complexity. Initially, calls were categorized in only two groups, frequency-modulated and flat calls. (11, 12) They were subsequently categorized in three/four groups, (13, 14) and currently, in 14 groups (Fig. 1). (1)
At present, most groups studying rat USVs use software called Avisoft-SASLab Pro developed by Avisoft Bioacoustics. Avisoft software is commonly used to record USVs, analyze spectrograms, and categorize calls. Post-recording, the sound files can be preprocessed in Avisoft-SASLab Pro through contrast adjustment and noise reduction. Subsequently, USVs can be identified and manually classified by subtype. Avisoft-SASLab Pro also includes an automatic call selection option. However, this approach is unreliable and requires frequent manual configuration. When analyzing

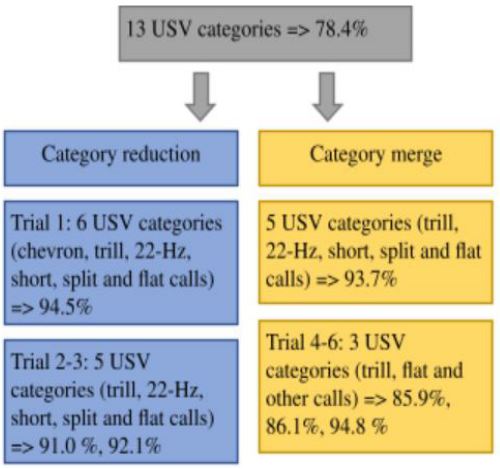

Figure 1. A graphical representation of the two approaches tested, category reduction (removing USC categories) and category merge (combining USV categories), and the overall accuracies of each trial.

hundreds of calls, manual processing is both time-consuming and prone to error. In addition, some USVs may be difficult to readily classify, which highlights the need for more robust classification procedures.

In the field of USV analysis, convolutional neural networks (CNN), a machine learning approach that leverages computer vision, have been increasingly used to identify and classify USVs. Programs such as DeepSqueak and VocalMat have shown high accuracies of USV identification and classification in mice. (15, 4) Furthermore, these programs have implemented adaptive methods to reduce noise and heighten the clarity of USV spectrograms. Notably, machine learning approaches can be trained on different datasets. This increases a program's flexibility for use in different experiments.

Although these new programs are promising, they have been developed largely or exclusively to recognize mouse USVs. VocalMat has not been established for use with rat USVs. This motivated us to develop a rat dataset, train VocalMat with it, and evaluate its performance on rat USV classification. By extending a program developed for use in mice to classify rat calls, we aim to advance automated USV analysis.

Review of Automated USV Analysis Approaches

Many programs have been used for automating rodent USV analysis, including MUPET, (16) A-MUD, (17) WAAVES, (18) XBAT, (19) DeepSqueak, (15) and VocalMat. (4) Since many of these programs do not employ machine learning, they lack the necessary flexibility to deal with different datasets that each contain a variety of USV categories. As a result, these programs, in addition to the other non-machine learning approaches, would need to be tailored to specific testing conditions. Most recently, DeepSqueak and VocalMat have been developed to use machine learning in the identification and classification of mouse calls.

DeepSqueak is a MATLAB package that couples an object detection network with a region proposal network[1] (R-CNN). DeepSqueak detects USVs by analyzing USV length, frequency range, and a classification confidence parameter. The classification confidence parameter is based on the probability that the region of interest contains a call or background noise. The calls can be discriminated from background noise (denoised) automatically to reduce the number of false positives. Thresholds for tonality, or the amount of energy focused at a single frequency, are applied to further reduce silence and noise, and subsequently, optimize the clarity and detection of USVs. The authors showed that high levels USV detection were maintained despite the addition of white noise or natural noise.

Supervised or unsupervised training methods can be used in DeepSqueak to classify USVs based on contour extraction of their spectrographic shapes. When employing supervised[2] training, Coffey et al. used five categories (split, inverted U, short rise, wave and step) to sort ~56,000 mouse USVs and train the classification CNN. For unsupervised training, the contours were automatically clustered by shape, duration and frequency. Individual USVs were subsequently sorted by their degree of similarity to the clusters. When optimized, 20 clusters were found to be most effective[3]. Therefore, DeepSqueak recognized 20 potential categories of mouse USVs with unsupervised training. (15)

VocalMat, developed by the Dietrich lab (Yale School of Medicine), is a MATLAB package that employs image-processing and differential geometry[4] approaches to analyze USVs. (4) In contrast to DeepSqueak, VocalMat identifies calls based on intensity thresholds calculated for each recorded segment. Sound recordings are converted to gray-scale spectrograms in which brighter pixels represent high-intensity values, and contrast enhancement is used to emphasize the USVs against background noise. Noise is further reduced via a local median filter based on the ratio of the median intensity of pixels in the detected USV candidates versus the background. Several morphological operations and the removal of small noise artifacts (≤ 60 pixels) enhance the clarity of USVs. The remaining USV candidates are classified into subtypes. In developing VocalMat, the Dietrich lab used transfer learning from AlexNet[5], pretrained on the ImageNet dataset, to train the CNN on prelabeled data (supervised learning method). They trained their program on images of individual calls extracted from spectrograms rather than trained on acoustic datafiles. They used 12,954 mice USVs across 12 categories: chevron, reverse chevron, down-frequency modulation, up-frequency modulation, flat, short, complex, step up, step down, two steps, multiple steps and noise.

When testing VocalMat and DeepSqueak, which were trained on a mouse datasets, on rat calls, there was poor performance (<50% accuracy). To test the efficacy of a CNN approach to USV classification, we modified VocalMat and trained it on a rat USV dataset. To date, no other labs have developed automated USV analysis programs specific to rats. VocalMat was selected over DeepSqueak due to its higher detection rate of mouse calls: 91.7% compared to 78.0%. (4) VocalMat classified USVs in 12 categories of mouse calls at an overall accuracy of 86.0%. (4) Additionally, VocalMat allows for efficient adaption of the program to different datasets as it uses transfer learning rather than training a network from scratch[6].

Methods

USV Categories

The USVs were classified according to the following categories (1):

USV	Description
Complex	Contain two or more directional changes in frequency of at least 3 kHz each
Upward Ramp	Monotonically increasing in frequency, with a mean slope not less than 0.2 kHz/ms
Downward Ramp	Monotonically decreasing in frequency, with a mean negative slope not less than 0.2 kHz/ms
Flat	Near-constant frequency greater than 30 kHz with a mean slope between -0.2 and 0.2 kHz/ms
Short	Duration less than 12 ms
Split	Middle component "jumps" to a lower frequency and contains a harmonic
Step-Up	Instantaneous frequency change to a higher frequency
Step-Down	Instantaneous frequency change to a lower frequency
Multi-Step	Two or more instantaneous frequency changes
Trill	Rapid frequency oscillations with a period of approximately 15 ms (either sinusoidal or appearing as repeated "inverted- U's")
Flat-Trill Combination	A trill that is flanked on one or both sides by a monotonic portion that is no less than 10 ms
Trill with Jumps	A trill that contains one or more higher-frequency components
Chevron	A monotonic increase followed by a monotonic frequency decrease, each of at least 5 kHz
Composite	Calls (other than flat/trill combinations) that comprise two or more categories
Unclear	Unclassifiable
Miscellaneous	Calls that are clear but do not fit into any of the above call categories
22- kHz	Near-constant frequency calls between 20 and 25 kHz

[1] A method used to predict the regions of objects and reduce the time of detection
[2] The use of a prelabelled dataset
[3] Based on the elbow method - the inflection point at which the introduction of new clusters produced diminishing improvement on error reduction
[4] Used to fine-tune CNNs with image features such as curvature
[5] A large CNN used for object recognition
[6] A computationally intensive process that could last days-weeks with negligible performance differences

Animals

Eight adult male Long-Evans rats (290-344 g, weighed at the start of USV recordings) were housed in a reverse cycle room (2 per cage). Food and water were accessible ad libitum outside of testing sessions, and the home cages were maintained by the animal facility of McGill University. All procedures were approved by the McGill Animal Care Committee in accordance with the guidelines of the Canadian Council on Animal Care.

Recording Sessions

The rats were recorded in a plexiglass chamber (29.5 cm high × 57.6 cm wide × 53.5 cm deep, ENV-007CT, Med Associates, St. Albans, VT). The chambers contained bedding that was changed between recording sessions and were encased in sound-reduction acoustic foam (Primacoustic, Port Coquitlam, British Columbia). USVs were recorded with a CM16 (Avisoft Bioacoustics, Berlin Germany) microphone angled towards the center of the box with the Avisoft Bioacoustics RECORDER software (version 4.2.29). Recording sessions were conducted on alternating days. During each recording session (20 min), four rats, each with their own chamber and microphone, were recorded as they naturally vocalized. Recording was started before the rats were placed in the chambers. A total of 216 audio (WAV) files were obtained.

Dataset

Spectrograms were generated from the audio files using a fast Fourier transform (FTT) length of 512 points and an overlap of 75% (FlatTop window, 100% frame size). The USVs had been previously manually identified and classified using Avisoft-SASLab Pro by another lab member (Adithi Sundarakrishnan) who provided the audio files used in this project. The next challenge was to export the classified USVs in separate image files with a standardized time and frequency axis. To this end, we adjusted the export parameters of the batch progress function with Raimund Specht (Berlin-based Avisoft Bioacoustics developer) and implemented new features that allowed the export of USVs into image (PNG) files organized by their classification. To standardize the time axis, we added margins to the USVs by inserting random background sections of the spectrograms. Following this procedure for 216 audio files, we obtained a total of 19,769 USVs[7].

VocalMat

The next step was to train VocalMat's USV classification program[8] on the rat dataset of images. Each training trial was independent of the others. We implemented a custom reader to standardize image size (227 x 227) and convert grayscale images to RGB for use with VocalMat.

```
function data = customreader(filename)
onState = warning('off', 'backtrace');
c = onCleanup(@() warning(onState));
data=imread(filename);
data = imresize(data, [227 227]);
data=data(:,:,min(1:3,end));
end
```

The classification CNN was trained using the Dietrich Lab recommended settings: a batch size of M=128 images and a maximum epoch number of 100. 33 epochs was the highest quantity used in our training. A stochastic gradient descent with momentum (0.9) at a learning rate of $\alpha = 10\text{-}4$ and weight decay $\lambda = 10\text{-}4$ was used to optimize training parameters. As this process is stochastic, the results from training a dataset differed between trials. Thus, we trained similar categories multiple times to compare results. In addition, we varied the number of categories to avoid overfitting[9] on our relatively small dataset. Specifically, we reduced the number of categories to lower the complexity, or number of parameters used for classification. The classification was trained on 90% of the inputted images and the accuracy, defined as the rate at which the automated classification matches the manual one, was assessed on the validation set (10%).

Results

We experimented with removing and merging categories of USVs used to train VocalMat (Fig. 1). The latter approach was used when USV subtypes shared stereotypical features such as trill jump and flat-trill USVs. Notably, removing categories would limit the pipeline's usefulness and merging categories would require manual categorization to distinguish the USVs within the merged category. When balancing the complexity of the model, or the number of categories with the degree of accuracy of specific USV subtypes, we prioritized high accuracy over including many categories. This would be of use when quantifying specific USV subtypes over recording sessions, rather than the prevalence of all USV subtypes over recording sessions.

Training with Original Categories

We trained the program to recognize all the categories of USVs described by Wright et al. (2010) with two exceptions: complex and composite calls were excluded as there were insufficient images available for training purposes. We obtained an overall accuracy of 78.4%. An accuracy higher than 85% was obtained for 22-kHz calls and for 50-kHz call types, i.e. trill (91.2%), short (93.3%), split (86.7%) and flat (87.1%). Based on these results, we reduced the USV categories from the initial 13 to 6 to omit the categories with the lowest accuracies, with the exception of chevron calls (Fig. 2a: Trial 1). We kept the chevron category because it had high accuracies when VocalMat was trained on the mouse dataset. The accuracy improved to 94.5%. All categories had an accuracy over 90.0% except for the classification of chevron calls (75.8%). Subsequently, we discarded the least accurate category, chevron calls, with its samples, and trained the program two times, now with only five categories: trill, 22-kHz, short, split and flat (Figure 2a: Trials 2 & 3). Trills, splits and flats showed the most consistent results. The overall accuracies were 91.0% and 92.1%, slightly below the previously obtained 94.5%.

Training with Merged Categories

To reduce potential overfitting, we next experimented with combining spectrographically similar USVs into categories. We merged the trill, flat-trill and trill jump calls (n=7724 calls), the split calls with the multi-step calls (n=899 calls), and the up ramp, down ramp and flat calls (n=5808 calls). The combined trill group resulted in an accuracy of 93.7%, lower than the mean of trills alone (~ 97%). The split calls combined with the multi-step calls had a lower accuracy of 53.3%. Notably, the multi-step

Figure 2a The performance of classification training when grouping the images in chevron, trill, 22-kHz, short, split and flat categories across three trials (blue, grey and orange). Figure 2b. The performance of classification training when grouping the images in trill, flat and other categories across three trials (yellow, blue and green).

[7] Please contact the author for access to the dataset
[8] Available at https://github.com/ahof1704/VocalMat.git
[9] This occurs when the model is too closely fitted to the training data and cannot be generalized and successfully used on different data
[10] A lack of capturing characteristic features, often the result of biased or limited data

calls have a much smaller training set and are thus susceptible to underfitting[10]. The flat calls combined with up and down ramp calls showed an accuracy of 93.8%, comparable to a mean accuracy of 92.0% of flats alone. Every region of a spectrogram that VocalMat identifies as salient (i.e. pinpointed by high-intensity pixels) must be classified in a category. Consequently, the authors included a "noise" category. Similarly, we implemented a three-way classification consisting of an "other" group, in addition to the trill and flat groups (the two most prevalent USVs). This gave us a better indication of how VocalMat would perform when analyzing highly variable noise artifacts and audio files, which are replete with individual calls that bear similarity to multiple categories and cannot be readily classified. We trained the program three times with different combinations of the flat and trill calls. In the first instance, we assessed trill (n=6330 calls) and flat calls (n=5103 calls) alone with the "other" category (n=7625 calls) and obtained a total accuracy of 85.9% (Fig. 2b: Trial 4). We then grouped the trill category with the trill jump and flat-trill categories (n=7724 calls) and retrained VocalMat. The accuracy of the combined trill category decreased from 92.2% to 90.0%, but the overall accuracy showed a negligible change of 85.9% to 86.1% (Fig. 2b: Trial 5). We observed the strongest performance when grouping the trill category with the trill jump and flat-trill categories (n=7724 calls), and the flat USVs with the up ramp and down ramp categories (n=5808 calls). Compared to the other two trials, the performance was higher in all groups and the overall result was 94.8% accuracy (Fig. 2b: Trial 6).

Discussion

Our results show high overall accuracy with a 6-category classification scheme of chevron, trill 22-Hz, short split and flat calls (94.5%, Trial 1) and with 3 merged categories (94.8%, Trial 6) as reductions from the initial 13 categories. These results are the first foray into adjusting VocalMat for rat, rather than mouse, USV classification. These 2 categorization approaches could be used depending on the degree of complexity required; the former is best suited for experiments evaluating multiple USV subtypes and the latter is best for focused analysis of specific subtypes.

As predicted, omitting certain call subtypes and merging other call subtypes to form larger categories had an effect on training accuracy. Counterintuitively, removing the most inaccurate USV category of chevron calls with its accompanying samples did not increase overall accuracy (Fig. 2a). This may be the result of decreased variance in the dataset. Merging categories also led to a decrease in accuracy except flat calls with up and down ramp calls (96.9%, Fig. 2b) and trill calls with trill jump and flat-trill calls (98.7%, Fig. 2b). Merging multiple USV subtypes that may not look alike within a category was expected to increase intra-category variability and decrease the accuracy of categorization. This approach is sufficient for experiments focused on trill and flat calls but will require adaptation for other common call subtypes. Further manual processing will be required to classify USVs within merged categories, such as discriminating between trill, trill jump and flat-trill calls. As VocalMat relies on a stochastic training process, more trials are required to compare the accuracies of different USV group combinations and determine the highest potential accuracies of each combination of categories, in addition to testing calls that are unable to be categorized.

While not a direct correlation, the USV categories with the most images, trill and flat calls, retained relatively consistent accuracies across the training trials and notably, the highest average accuracies (97.2% and 91.0%, Fig. 2a). These calls have relatively simple spectrographic shapes in comparison to the other call categories. As a result, the dataset for these categories may have less variability, leading to more consistent categorization. Our dataset is a reflection of the actual USV prevalence by category; notably, trills and flat calls each represent on average about 30% of all 50-kHz calls. In addition, there are large variations in the number of calls per category. More data is needed to fill these gaps and provide additional training samples for infrequent calls. Specifically, a large dataset comprising calls from male and female rats at different ages will be critical to capture variations within USV categories, (20, 21) as individual rats differ in terms of which USV subtype they most commonly emit. (1) Through following the procedure outlined in this paper, it is possible to construct and grow databases for use with machine learning programs.

In summary, we have adapted VocalMat for use in adult rats and have shown that it can classify 50-kHz calls with a high degree of accuracy. These results show promise and the established procedure will help build upon these results. By experimenting with other datasets, we hope to further test VocalMat and improve USV call classification.

Acknowledgements

I thank my supervisor, Dr. Paul Clarke, for his insightful feedback and continual guidance and Adithi Sundarakrishnan for her generous support. This work was partially funded by an NSERC undergraduate student research award. The author declares no conflict of interest.

References

1. Wright JM, Gourdon JC, Clarke PBS. Identification of multiple call categories within the rich repertoire of adult rat 50-kHz ultrasonic vocalizations: effects of amphetamine and social context. Psychopharmacology. 2010;211(1):1-13.

2. Litvin Y, Blanchard DC, Blanchard RJ. Rat 22kHz ultrasonic vocalizations as alarm cries. Behavioural Brain Research. 2007;182(2):166-72.

3. Goffinet J, Mooney R, Pearson J. Inferring low-dimensional latent descriptions of animal vocalizations. bioRxiv. 2019:811661.

4. Fonseca AHO, Santana GM, Bampi S, Dietrich MO. Analysis of Ultrasonic Vocalizations from Mice Using Computer Vision and Machine Learning. bioRxiv. 2020:2020.05.20.105023.

5. Clarke PB WJ. Rodent ultrasonic vocalizations. In: Stolerman I, editor. I.P. Encyclopedia of Psychopharmacology: Springer-Verlag Berlin Heidelberg; 2015. p. 1918.

6. Simola N, Granon S. Ultrasonic vocalizations as a tool in studying emotional states in rodent models of social behavior and brain disease. Neuropharmacology. 2019;159:107420.

7. Simola N. Rat Ultrasonic Vocalizations and Behavioral Neuropharmacology: From the Screening of Drugs to the Study of Disease. Curr Neuropharmacol. 2015;13(2):164-79.

8. Brudzynski SM. Pharmacology of Ultrasonic Vocalizations in adult Rats: Significance, Call Classification and Neural Substrate. Current neuropharmacology. 2015;13(2):180-92.

9. Boulanger-Bertolus J, Rincón-Cortés M, Sullivan RM, Mouly A-M. Understanding pup affective state through ethologically significant ultrasonic vocalization frequency. Scientific reports. 2017;7(1):13483-.

10. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes, Brain and Behavior. 2011;10(1):44-56.

11. Burgdorf J, Wood PL, Kroes RA, Moskal JR, Panksepp J. Neurobiology of 50-kHz ultrasonic vocalizations in rats: Electrode mapping, lesion, and pharmacology studies. Behavioural Brain Research. 2007;182(2):274-83.

12. Ahrens AM, Ma ST, Maier EY, Duvauchelle CL, Schallert T. Repeated intravenous amphetamine exposure: rapid and persistent sensitization of 50-kHz ultrasonic trill calls in rats. Behavioural brain research. 2009;197(1):205-9.

13. Kaltwasser M-T. Acoustic signaling in the black rat (Rattus rattus). Journal of Comparative Psychology. 1990;104(3):227-32.

14. Vivian JA, Miczek KA. Morphine attenuates ultrasonic vocalization during agonistic encounters in adult male rats. Psychopharmacology (Berl). 1993;111(3):367-75.

15. Coffey KR, Marx RG, Neumaier JF. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology. 2019;44(5):859-68.

16. Van Segbroeck M, Knoll AT, Levitt P, Narayanan S. MUPET—Mouse Ultrasonic Profile ExTraction: A Signal Processing Tool for Rapid and Unsupervised Analysis of Ultrasonic Vocalizations. Neuron. 2017;94(3):465-85.e5.

17. Zala S, Nicolakis D, Noll A, Balazs P, Penn D. Automatic mouse ultrasound detector (A-MUD): A new tool for processing rodent vocalizations. PLOS ONE. 2017;12:e0181200.

18. Reno JM, Marker B, Cormack LK, Schallert T, Duvauchelle CL. Automating ultrasonic vocalization analyses: The WAAVES program. Journal of Neuroscience Methods. 2013;219(1):155-61.

19. Barker DJ, Herrera C, West MO. Automated detection of 50-kHz ultrasonic vocalizations using template matching in XBAT. Journal of Neuroscience Methods. 2014;236:68-75.

20. Mittal N, Thakore N, Bell RL, Maddox WT, Schallert T, Duvauchelle CL. Sex-specific ultrasonic vocalization patterns and alcohol consumption in high alcohol-drinking (HAD-1) rats. Physiology & Behavior. 2019;203:81-90.

21. Wöhr M, Schwarting RKW. Affective communication in rodents: ultrasonic vocalizations as a tool for research on emotion and motivation. Cell and Tissue Research. 2013;354(1):81-97