Pitfalls of Using Feature-Based Classification for Mouse Ultrasonic Vocalizations

Reyhaneh Abbasi[1,2,3], Peter Balazs[1], Dustin J. Penn[2], Sarah M. Zala[2]

[1]*Acoustic Research Institute, Austrian Academy of Science, 1040 Vienna, Austria*
E-Mails: reyhaneh.abbasi@oeaw.ac.a and peter.balazs@oeaw.ac.at
[2]*Konrad Lorenz Institute of Ethology, Department of Interdisciplinary Life Sciences, University of Veterinary Medicine, 1160 Vienna, Austria*
E-Mails: Dustin.Penn@vetmeduni.ac.at and Sarah.Zala@vetmeduni.ac.at
[3]*Vienna Doctoral School of Cognition, Behaviour and Neuroscience, University of Vienna, Vienna, Austria*

Abstract

House mice communicate through ultrasonic vocalizations (USVs), which are above the range of human hearing (>20 kHz), and are often classified into ca. 10 syllable types. Feature-based classifiers are mainly used to study USVs, and yet these are outperformed by image-based models in studies of other types of acoustic signals. In this study, we examine the pitfalls of feature-based classification of USVs comparing them with a state-of-the-art image-based classification method. Our results show that the limitation of feature-based methods can be attributed to two main reasons: 1) features are defined by the frequency track (FT) of USVs, and as a result, any error in determining the FT has a profound impact on the extracted features and thus the performance of the classifier. 2) The features are based on human-determined thresholds, which may not be met in case of any trivial variations in the calculated FTs. We provide the results of the numerical experiments.

Introduction

The ultrasonic vocalizations (USVs) of house mice and rats are becoming increasingly used as models to investigate to animal communication [1-3] and for studying the genetic basis of autism and speech disorders in humans [4, 5]. In these studies, USVs are usually categoried to investigate the potential information content of these signals. USVs are manually classified into as many as different 12 classes, depending on their spectro-temporal features [5-9]. These classes are based on mandy criteria, including frequency changes [9] (> 5 kHz increase "up", > 5 kHz decrease "d"), the number of components ("c2" with 2 and "c3" with 3 components), changes of frequency direction (≥ 2 changes "c") or shape (u-shape, "u", u-inverted shape, "ui"), frequency modulation (< 5kHz, "f"), length (5-10 ms, "s", < 5ms, "us") and presence of harmonic elements ("h"). Manual methods used to classify USVs are very time-consuming. Hence, researchers often point out for the need to develop better automatic techniques for USV classification.

Classification of USVs can be achieved through both unsupervised [10-14] and supervised [10, 15] classifiers. For example, Vogel, et al. [15] classified USVs from C57BL/6J mice into 9 classes, including "up", "d", "c2", "c3", "c", "ui", "f", "s", and "d". To provide the input, numerical features had first been extracted with a 25-high signal-to-noise-ratio instances from each class, and their feature-based classifier (i.e., Random Forest (RF) [16]) yielded a classification accuracy of 85%. Abbasi, et al. [14], used a image-based classifier, classified the detected segments from mouse vocalizaitons into the classes "c2", "c3", USVs without jumps, and false positives (FP). In previous study, a supervised Convolutional Neural Network (CNNs) [17] was trained using 1200 samples and fed by 2D Gammatone filtered spectrograms (GSs), adapted to the frequency range of mice. This classifier yielded a macro-F1 score of 90±2.7%. Recently, a study compared the accuracy of different machine learning methods for classifying 48669 mouse USVs into 10 classes [18]. Classes include "up", "d", "c2", "c3", "c", "ui", "f", "s", and "h" and composite (meaning two harmonically independent components). Commersial software (*Avisoft Bioacoustics,* Germany) was used to detect USVs and calculate their features. CNNs fed by spectroscopic images were found to outperform various classical feature-based machine learning algorithms (including supoort vector machines (SVM) [19]). It was concluded that image-based methods are superior over feature-based ones, though the authors did not address the underlying reasons for this difference in performance.

In the present study, we focused on the pitfalls of feature-based classification of USVs comparing them with a state-of-the-art image-based classification method. This does not neccesarily mean that feature-based classification is alway inferior, but that the features have to be well-designed. For image-based classification, we used CNN fed by 2D images of USVs, and for feature-based methods, we applied three machine learning algorithms, including a Multilayer Perceptron (MPL) [20], SVM, and RF.

Data

To develop the automatic USVs classification we used recordings of adult wild-derived house mice. Animal housing and weaning conditions have been previously described [21]. The dataset consisted of recordings from male (n=11) and female (n=3) subjects during 10 min in the presence of an unfamiliar female stimulus mouse (as described in Zala, et al. [21]).

Before data processing, USVs need to be deteted. For detection, we applied the Automatic Mouse Ultrasound Detector (A-MUD) [21] using its default parameters. After automatically detecting all segments, the dataset was manually labeled into 9 classes: "up", "d", "c2", "c3", "c", "u", "ui", "f", and "s" (Figure 1). The dataset consisted of 4049 USVs, and had a very uneven distribution of different classes, e.g, fewest (50 samples) were "c3" and most were "up" calls (1948 samples).

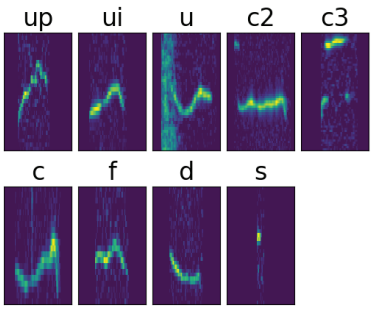

Figure 1: Exemplar spectrograms of mouse USVs illustrating nine distinct classes. USVs representing "up", inverted-u ("ui"), u-shape ("u"), one jump included ("c2"), two jumps included ("c3"), complex ("c"), flat ("f"), down ("d"), and short ("s") USV classes

Methods

The classification of USVs was performed using both feature-based and image-based classifiers. In both approaches, the first step consisted of features extraction, followed by the classification phase. To evaluate the classifiers, we used stratified 7-fold cross-validation. The dataset was split into subsets: training (70% of data), validation (15%), and test (15%) data.

Feature-based classification

We developed a feature based classifier by calculating the frequency track (FT)-derived features. FT is the estimated track of the instantaneous frequency over time, which is given in the A-MUD output [9]. Briefly, to calculate FTs, the first frequencies of the three highest peaks of the liftered spectrum are found for each time point. Then, using the Viterbi algorithm, the most probable sequence of elements is selected. An example of a FT obtained for a USV of "c3" class member is shown in Figure 2.

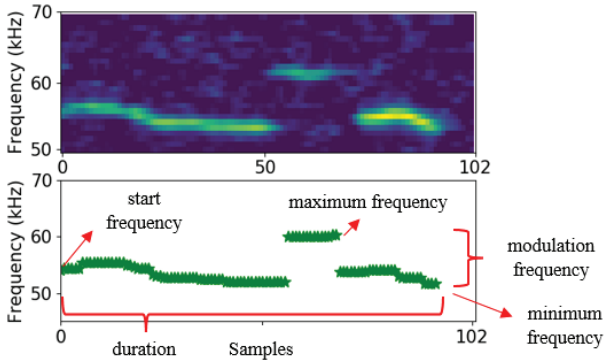

Figure 2: The USV spectrogram of a sample in the class "c3" (top). The extracted frequency track (FT) obtained from A-MUD is shown at the bottom

The 16 features include duration, slope, the number of jumps (calls with a change between frequencies of consecutive points > 10 kHz), number of changes in direction (a change in frequency > 5 kHz), minimum, maximum, standard deviation, and mean frequency, starting and ending point frequencies, the coefficients of the 3rd order linear equation fitted to the FT of USVs (excluding intercept), number of positive and negative convexities in FT (calculated based on the slope of the FT), the difference between lowest and ending point, between lowest and starting point, and between starting and ending point frequencies, are extracted from the FTs.

Feature selection was conducted using Recursive Feature Elimination (RFE) [22] based on the RF classifier. This method selects the features by recursively considering smaller and smaller sets of features, which are fed to the classifier. Feature selection is applied only on the training dataset. The selected features were then fed into three machine learning methods including SVM, RF, and MLP.

We tested the performance of MLP with different numbers of neurons and layers. The best performing configuration was the one that consisted of five fully connected layers, respectively with 20, 32, 64, 128, and 10 neurons. The network was trained using a decreasing learning rate (10^{-2} to $0.5*10^{-5}$), for 200 epochs. For the developement of RF, 200, 300, 400, and 500 trees were trained with the 16-dimensional features. The best results were obtained using 400 trees. The SVM was trained using the following parameteres: one versus rest configuration, Radial Basis Function kernel, C equal to 5000, and Gamma with the value of 0.0005. Because the training data is imbalanced, in all methods we have used weights which are inversely proportional to class frequencies in the train data.

Image-based classification

In image-based classification algorithms, data are given in the form of two-dimensional images to a classifier. In the present study, the images were produced using the short-time Fourier transforms of elements in the frequency interval of 20 kHz to 120 kHz. Then, following previous studies [12, 23] the Gammatone filter bank is used to reduce the size of the STFT array along the frequency axis from 251 × 401 to 64 × 401. Then, USVs were classified by feeding these 2D images into a CNN.

The architecture of our CNNs included 5 convolutional layers with ELU activation function, each of them is composed of 32, 64, 64, 128, and 128 filters, and 3 fully connected layers composed of 64, 64, and 10 neurons. The kernel size was 3*18 in the first layer, and 3*3 in the successive four layers. A max-pooling layer was applied after each of them. The pooling size was fixed to 2*2. The fully connected layers had 128 neurons with ELU activation function. The last fully connected layer had 10 neurons with a Softmax activation function. We trained our CNNs with a batch size of 32 with 200 epochs. We used the categorical cross-entropy as the loss function. To promote the generalizability of the model, we performed the augmentation of the training dataset using random shifts of width and height by 10%. To investigate the effect of imbalanced training data we have used weights in the loss function, inversely proportional to class frequencies. The number of epochs was set to 200 with batch size and learning rate equal to 32 and a decreasing learning rate (10^{-2} to $0.5*10^{-5}$), respectively.

Results

According to RFE, the following 10 features are used for the development of feature-based classifier. They include duration, modulation frequency, the difference between ending and starting points of FT, the difference between

lowest and starting point frequencies, the difference between lowest and ending point frequencies, the standard deviation of FT, maximum frequency, number of negative convexities, and the first and third coefficients of the linear regression fitted to FT.

We compared the performance of CNN to RF, SVM, and MLP (Table 1). The F1-score results clearly show that image-based USV classification works significantly better than feature-based methods. Similar to [15], we also found that among the feature-based algorithms, RF outperformed the other two feature-based methods. This result is probably because RF is based on bagging a large number of trees which improves its generalizability on the test dataset.

Table 1: The comparison of macro F1-score obtained from 4 classification methods. The results are reported based on test data from stratified 7 folds.

Metric \ Model	CNN	RF	SVM	MLP
Macro F1-score	**71±3**	67±3	62±1	57±4

Comparing the class-wise performance of models (Table 2), we found that RF outperformed CNN for the classes "up", "f", and "d", they had similar performance for the class "s", and CNN was the superior classifier in the other classes. The outperformance of RF for the classes "up", "d", and "f" is likely because using visual inspection, the calculation of FT is less error prone than other classes. In addition, their features extracted from FT, are normally more distinctive than other classes (as they depend on modulation frequency). Conversely, because of the complexity of classes "c", "c2", and "c3", the claculated FT is less accurate and less reliable to build on extracted features (Figure 3). Even if the computational error is small in the FT computational stage, it will not be easy to determine the appropriate features for these classes again due to the complexity. As a result, image-based methods that do not require the use of these features can model the output more accurately.

Table 2: The comparison of class-wise F1-score, obtained from 4 classification methods. The results are reported based on test data from 7 folds.

Model	Studied classes								
	c	c2	c3	d	f	s	u	ui	up
CNN	**57 ±8**	**66 ±5**	**77 ±5**	64 ±4	64 ±6	**63 ±10**	**73 ±6**	**72 ±4**	85 ±2
RF	43 ±8	65 ±7	64 ±4	**69 ±7**	**76 ±4**	**63 ±10**	70 ±7	68 ±4	**86 ±2**
SVM	43 ±4	54 ±6	48 ±14	66 ±6	69 ±3	58 ±7	66 ±6	70 ±2	82 ±0
MLP	32 ±8	52 ±4	40 ±7	63 ±6	71 ±3	51 ±5	61 ±6	59 ±10	77 ±4

In addition to the problem described above, we found that the two additional issues may also affect the performance of feature-based classifiers: (1) there is inevitable error during extracting FT because of the background noise and faint USVs, and as a result, derived features are miscalculated. (2) Features are defined based on human-determined thresholds. For example, jumps in frequency are found if the change in frequency is more than 10 kHz. Obviously, if there are trivial changes in the calculated FTs, this threshold may not be met and as a result, the target USV is misclassified as an element with no jump.

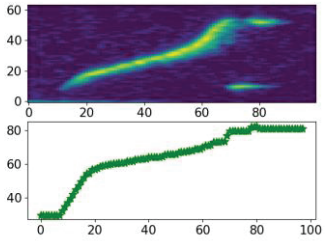

Figure 3: A sample of miscalculated frequency track for a member of the class "c3". It can be seen that the jump in the USV was not detected due to a computational error.

Conclusions

In this study, we addressed the pitfalls of classifying USVs using feature-based classifiers. We compared the performance of three feature-based classifiers against one image-based model. We found that the image-based classifier outperfomred feature-based classifiers, especially in classes that the calculation of FT and FT-derived features are error-prone. We also found that error may caused by background noise, faint USVs, and the sensitivity of the features to human-defined thresholds. We conclude that a feature-based method neccessiates well-designed features.

Acknowledgment

This work was supported by the Austrian Science Fund (AT) P28141-B25 to DJP and SMZ.

References

[1] G. Ehret, "Characteristics of vocalization in adult mice," in *Handbook of behavioral neuroscience*. vol. 25, ed: Elsevier, 2018, pp. 187-195.

[2] S. M. Brudzynski, *Handbook of Ultrasonic Vocalization: A Window Into the Emotional Brain* vol. 25: Academic Press, 2018.

[3] J. Heckman, B. McGuinness, T. Celikel, and B. Englitz, "Determinants of the mouse ultrasonic vocal structure and repertoire," *Neuroscience & Biobehavioral Reviews,* vol. 65, pp. 313-325, 2016.

[4] J. Fischer and K. Hammerschmidt, "Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication," *Genes, Brain and Behavior,* vol. 10, pp. 17-27, 2011.

[5] M. L. Scattoni, S. U. Gandhy, L. Ricceri, and J. N. Crawley, "Unusual repertoire of vocalizations in the BTBR T+ tf/J mouse model of autism," *PloS one,* vol. 3, p. e3067, 2008.

[6] K. Musolf, S. Meindl, A. L. Larsen, M. C. Kalcounis-Rueppell, and D. J. Penn, "Ultrasonic vocalizations of male mice differ among species and females show assortative preferences for male calls," *PLoS one,* vol. 10, 2015.

[7] M. A. Marconi, D. Nicolakis, R. Abbasi, D. J. Penn, and S. M. Zala, "Ultrasonic courtship vocalizations of male house mice contain distinct individual signatures," *Animal Behaviour,* 2020.

[8] D. Nicolakis, M. A. Marconi, S. M. Zala, and D. J. Penn, "Ultrasonic vocalizations in house mice depend upon genetic relatedness of mating partners and correlate with subsequent reproductive success," *Frontiers in zoology,* vol. 17, pp. 1-19, 2020.

[9] S. M. Zala, D. Nicolakis, M. A. Marconi , A. Noll, T. Ruf, P. Balazs *, et al.*, "Primed to vocalize: Wild-derived male house mice increase vocalization rate and diversity after a previous encounter with a female," *PLoS ONE,* vol. 15, p. e0242959, 2020.

[10] K. R. Coffey, R. G. Marx, and J. F. Neumaier, "DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," *Neuropsychopharmacology,* vol. 44, pp. 859-868, 2019.

[11] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference and Prediction.," 2 ed: Springer, 2009, pp. 485-585.

[12] M. Van Segbroeck, A. T. Knoll, P. Levitt, and S. Narayanan, "MUPET—mouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations," *Neuron,* vol. 94, pp. 465-485. e5, 2017.

[13] J. Chabout, J. Jones-Macopson, and E. D. Jarvis, "Eliciting and analyzing male mouse ultrasonic vocalization (USV) songs," *Journal of visualized experiments: JoVE,* 2017.

[14] X. Dou, S. Shirahata, and H. Sugimoto, "Functional clustering of mouse ultrasonic vocalization data," *PloS ONE,* vol. 13, p. e0196834, 2018.

[15] A. P. Vogel, A. Tsanas, and M. L. Scattoni, "Quantifying ultrasonic mouse vocalizations using acoustic analysis in a supervised statistical machine learning framework," *Scientific reports,* vol. 9, p. 8100, 2019.

[16] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* vol. 1: MIT press Cambridge, 2016.

[18] M. Premoli, D. Baggi, M. Bianchetti, A. Gnutti, M. Bondaschi, A. Mastinu*, et al.*, "Automatic classification of mice vocalizations using Machine Learning techniques and Convolutional Neural Networks," *PLOS ONE,* vol. 16, p. e0244636, 2021.

[19] I. Steinwart and A. Christmann, *Support vector machines*: Springer Science & Business Media, 2008.

[20] M. H. Hassoun, *Fundamentals of artificial neural networks*: MIT press, 1995.

[21] S. M. Zala, D. Reitschmidt, A. Noll, P. Balazs, and D. J. Penn, "Automatic mouse ultrasound detector (A-MUD): A new tool for processing rodent vocalizations," *PloS One,* vol. 12, p. e0181200, 2017.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning,* vol. 46, pp. 389-422, 2002.

[23] R. Abbasi, P. Balazs, A. Noll, D. Nicolakis, M. A. Marconi, S. M. Zala*, et al.*, "Applying convolutional neural networks to the analysis of mouse ultrasonic vocalizations," 2019.