BRINGING IN THE OUTLIERS: A SPARSE SUBSPACE CLUSTERING APPROACH TO LEARN A DICTIONARY OF MOUSE ULTRASONIC VOCALIZATIONS

Jiaxi Wang[⋆] *Karel Mundnich*[⋆] *Allison T. Knoll* [†] *Pat Levitt*[†] *Shrikanth Narayanan*[⋆]

[⋆] Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, USA

[†] Department of Pediatrics and Program in Developmental Neuroscience and Neurogenetics, The Saban Research Institute, Children's Hospital Los Angeles, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

ABSTRACT

Mice vocalize in the ultrasonic range during social interactions. These vocalizations are used in neuroscience and clinical studies to tap into complex behaviors and states. The analysis of these ultrasonic vocalizations (USVs) has been traditionally a manual process, which is prone to errors and human bias, and is not scalable to large scale analysis. We propose a new method to automatically create a dictionary of USVs based on a two-step spectral clustering approach, where we split the set of USVs into inlier and outlier data sets. This approach is motivated by the known degrading performance of sparse subspace clustering with outliers. We apply spectral clustering to the inlier data set and later find the clusters for the outliers. We propose quantitative and qualitative performance measures to evaluate our method in this setting, where there is no ground truth. Our approach outperforms two baselines based on k-means and spectral clustering in all of the proposed performance measures, showing greater distances between clusters and more variability between clusters.

Index Terms— sparse subspace clustering, subspace similarity, clustering, mouse ultrasonic vocalizations.

1. INTRODUCTION

Mice vocalize in the ultrasonic range between 30 and 150kHz [1]. These vocalizations are primarily observed during social interactions and their properties vary across different social contexts [2, 3, 4]. Given the social nature of this behavior, mouse ultrasonic vocalizations (USVs) have become of special interest in biomedical research to obtain information about complex social communication behaviors, especially in genetic models of neurodevelopmental disorders [5] by using vocal communication as a proxy [6, 7] to study different behavioral patterns and states.

Generating a dictionary (inventory) of USVs is a critical step in the behavioral analysis of social communication and interaction. The study of vocalizations has traditionally been performed in a manual process, where human annotators annotate and cluster USVs into a small number of groups (4–10) according to subjective rules, including their shape in the frequency domain and their duration [8]. However, this is a time-consuming process that is prone to annotation errors and annotator biases, both of which potentially affect subsequent analyses, and which may not necessarily reflect the structure of the data. Therefore, the problem of finding a dictionary of vocalizations in animals using automated tools has started to attract interest in recent years [9, 10, 11, 12].

Jiaxi Wang's work was performed during an internship at USC SAIL. He is now with the Department of Electrical Engineering, Tsinghua University.

To analyze the sounds emitted by mice, it is common to analyze the spectrograms of the ultrasonic audio signals. An example of these vocalizations in the frequency domain is shown in Fig. 1, where different shapes can be identified in time in the spectrogram of the ultrasonic audio signal. From Fig. 1, we can see that the clustering of USVs into groups is a non-trivial problem, due to the lack of ground truth (the true number of different vocalization types is unknown), no intuitive distance or similarity between the USVs to aid the comparison between them, and the difficulty involved in creating rules that generalize over different sets of USVs (for example, generated by different strains of mice).

1.1. Related work

Different approaches have been proposed in the literature to analyze USVs in an automated fashion. A common first step is to detect the USVs in an audio signal. Several automated tools have been proposed [10, 11, 12, 13, 14], all of which extract these vocalizations from the spectrograms of the audio signal through different methods. For example, MUPET [11] detects the vocalizations based on the frame energy after a noise-reduction step, in a similar fashion to voice activity detection (VAD) in quiet environments. DeepSqueak [12] takes a computer vision approach, where a Faster-RCNN [15] is trained to detect the vocalizations in spectrograms as in object detection tasks.

The clustering approach to learn a dictionary of USVs was proposed in MUPET [11], and later employed in DeepSqueak [12]. Both methods use k-means at their core: [11] uses features obtained from the audio signal through a gammatone filterbank to perform clustering, while [12] uses intuitive features such as shape, frequency, and duration in time to cluster the USVs. In MUPET, the number of clusters is user-defined, and different clusters have overlapping features (leading to clusters that are noisy). A different approach is taken by MSA [9], where a small number of clusters or categories is used (often only 4), leading to high variability in the shapes of USVs within clusters: upward shapes that are both long and short, as well as complex shapes with different feature values.

1.2. Contributions

In this work, we propose a novel approach to cluster USVs based on the hypothesis that the information in the frequency domain lies in low-dimensional subspaces. We are motivated by the notion of distances between subspaces proposed in [16], and use these distances to perform sparse subspace clustering (SSC) over sets of segmented USVs. Since sparse subspace clustering is sensitive to outliers, we also propose a two-step approach to sparse subspace clustering, where we first cluster an inlier set of USVs and then add

outliers back into our clusters. Finally, we propose a new way to evaluate the quality of the clusters for this particular task.

2. BACKGROUND: SUBSPACE CLUSTERING

Subspace clustering aims to find subspaces in which the data samples lie by posing the clustering problem as a regression problem. In particular, sparse subspace clustering (SSC) [17, 18] is a popular version of subspace clustering methods due to its good performance in different application areas, as well as theoretical guarantees [16, 19]. For an in-depth overview of the methods, we refer the readers to [20].

The basic idea behind SSC is that data points that lie in a linear subspace can be linearly represented by other data points located in the same subspace. The coefficients that represent these subspaces, which can be found through the LASSO [21], reflect the similarity between different data points and can be used to build an affinity matrix. Then, spectral clustering [22, 23] is applied to this affinity matrix to get different clusters. According to [16, 19], the success of SSC depends mainly on three properties of the data: (1) low affinity between subspaces, (2) enough samples for a certain subspace, and (3) the data not having too many outliers (where an outlier is defined as a data point that does not lie in any of the subspaces).

Formally, sparse subspace clustering can be posed as an ℓ_1 minimization problem over the subspace coefficients:

$$\min_{\boldsymbol{y}} \|\boldsymbol{y}\|_1 \quad s.t. \quad \boldsymbol{s} = \boldsymbol{S}^* \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{y}$ contains the coefficients of the subspace and $\boldsymbol{S}^*$ contains all (vectorized) USVs in its column, except for USV $\boldsymbol{s}$.

An equivalent formulation uses LASSO instead:

$$\min_{\boldsymbol{y}} \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{S}^* \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{y}\|_1. \tag{2}$$

We can rewrite Eq. 2 in matrix form to include all USVs $\boldsymbol{s}_i, i \in \{1, \ldots, N\}$:

$$\min_{\boldsymbol{y}} \frac{1}{2}\|\boldsymbol{S} - \boldsymbol{S}\boldsymbol{Y}\|_2^2 + \lambda \|\boldsymbol{Y}\|_1 \quad s.t. \quad \text{diag}(\boldsymbol{Y}) = 0, \tag{3}$$

where $\text{diag}(\boldsymbol{Y})$ is a vector with the diagonal elements of $\boldsymbol{Y}$. Each column of $\boldsymbol{S}$ contains a vocalization and each column of $\boldsymbol{Y}$ contains the coefficients for the different USVs.

From $\boldsymbol{Y}$, the similarity matrix $\boldsymbol{A}$ is computed by:

$$\boldsymbol{A} = |\boldsymbol{Y}| + |\boldsymbol{Y}|^T. \tag{4}$$

The elements in $\boldsymbol{A}$ are the subspace similarities. To find the clusters, we apply spectral clustering (with random walk Laplacian) [23, 24] to $\boldsymbol{A}$.

3. PROPOSED METHOD

Our proposed method is a hybrid approach between sparse subspace clustering and more traditional clustering techniques. We are motivated by the fact that sparse subspace clustering does not perform well in the case of many outliers, which are expected to exist in spectrally diverse and non-stereotyped mouse USV data sets. Therefore, we propose a simple approach for outlier detection, perform sparse subspace clustering on the inlier data set, and then add the outliers to our clusters. The steps for the method are as follows: (1) detection of USVs using MUPET, (2) pre-processing of the segmented USVs, (3) dividing our data set into inliers and outliers, (4) performing subspace clustering on the inlier data set, and (5) assigning a cluster to each USV in the outlier set. These steps are summarized in Fig. 2.

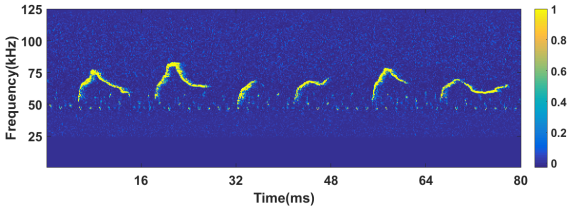


Fig. 1: A spectrogram depicting a sequence of mouse ultrasonic vocalizations (USVs). Brighter colors represent higher energy in a given frequency band.

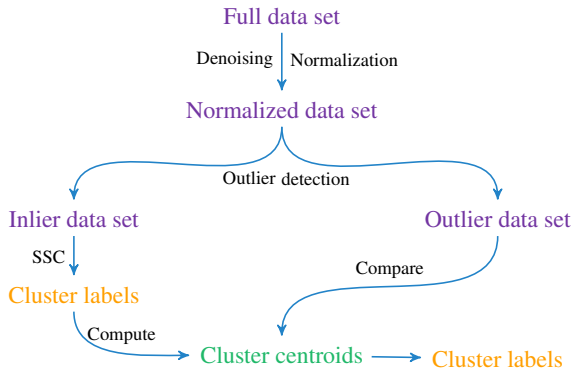


Fig. 2: Summary of the proposed method.

3.1. Preprocessing

3.1.1. Segmentation, denoising and normalization

To detect and segment USVs from audio files, we use MUPET. After segmentation, the vocalizations may have different time durations or be located in distinct frequency bands. Therefore, we normalize the vocalizations to a same size $F \times T$ (note that we depart from MUPET in this step). Therefore, we represent each USV by employing the regions in the frequency domain whose energy is greater than the average energy, by clipping the areas with energy below the average energy. We note that the segmented USVs contain energy in different frequency bands, and are of different time-lengths. Therefore, we use MATLAB's `imresize` function (which implements a bi-cubic interpolation) to normalize the segmented USVs into a same size, after which the USVs with similar shapes are frequency-aligned and have the same length in time. After the normalization, we vectorize the USV matrices.

3.1.2. Outlier detection

Since the performance of sparse subspace clustering highly degrades in the case of a large amount of outliers [16], we first remove the outliers and perform sparse subspace clustering on the resulting inlier data set.

Intuitively, an outlier is an USV that differs from any of all other USVs. Let $\boldsymbol{s}_i$ denote the vector of the i^{th} vocalization, where $i \in \{1, \ldots, N\}$. We define a vocalization $\boldsymbol{s}$ as an outlier if:

$$\max_{\boldsymbol{s} \neq \boldsymbol{s}'} \cos(\boldsymbol{s}, \boldsymbol{s}') < \tau, \tag{5}$$

where τ is a threshold to be chosen and $\cos(\cdot, \cdot)$ is the *cosine similarity* (CS) between vectors:

$$\cos(\boldsymbol{s}, \boldsymbol{s}') = \frac{\boldsymbol{s} \cdot \boldsymbol{s}'}{\|\boldsymbol{s}\| \|\boldsymbol{s}'\|}. \tag{6}$$

3.2. Step one: clustering inliers

We perform sparse subspace clustering as described in Sec. 2 in the inlier data set. In this case, $\boldsymbol{S}$ is the matrix of (vectorized) USVs and $\boldsymbol{Y}$ is the matrix of subspace coefficients.

3.3. Step two: clustering outliers

After assigning clusters to the inlier data set, we assign clusters to each one of the USVs in the outlier data set. There are mainly two ways of assignment: (1) we can cluster the outliers into existing categories, and (2) we can create new clusters for outliers. Here, we simply cluster the outliers into existing categories, and leave the problem of assigning novel clusters for future work. To perform cluster assignments, we first define the concept of *centroids* in this setting as the mean of the USVs inside a cluster. Assume $\boldsymbol{c}_k$ is the centroid of cluster k we get from sparse subspace clustering, $k = 1, 2, ..., K$ where K is the number of clusters. We assign an outlier $\boldsymbol{s}_{out}$ into the most similar cluster k using:

$$k = \underset{j}{\operatorname{argmax}} \cos(\boldsymbol{s}_{out}, \boldsymbol{c}_j). \tag{7}$$

4. EXPERIMENTS

We now describe our experiments. All the code is available online[1].

4.1. Data

The data set[2] we use contains 40 records sampled at 250kHz emitted by both DBA/2J (DBA) and C57Bl/6J (C57) mouse strains. After vocalization detection using MUPET [11], we used approximately 9000 vocalizations per strain. For details on how the data was collected, we refer readers to [25].

4.2. Parameters

We set $F = T = 64$, such that each vector $\boldsymbol{s}$ is of length 64×64. We choose the outlier threshold $\tau = 0.8$ for DBA and $\tau = 0.7$ for C57 (Eq. 5). According to [16], the selection of λ depends on the dimension of the subspaces, which remains unknown in our case. In practice, we find the dimension of the subspaces empirically by trial and error, and settle for $\lambda = 0.3$ for both data sets, which contains moderately sparse represented coefficients. We also test the different clustering methods with $K = 20, 40$, and 60 clusters.

Given that small perturbations in the adjacency matrix $\boldsymbol{A}$ may influence the eigenvectors of the Laplacian matrix computed during spectral clustering (Eq. 4), negatively affecting the results of the clustering performance, we set the elements in $\boldsymbol{Y}$ smaller than 0.001 to 0. We regard these values of $\boldsymbol{Y}$ as noise.

4.3. Baselines and Methods

We compare our proposed approach with two different baselines. The first one is the k-means approach used by MUPET. The second baseline uses spectral clustering by computing a similarity matrix based on the cosine similarity (CS), where the similarity is defined as:

$$\boldsymbol{A}_{ij} = \cos(\boldsymbol{s}_i, \boldsymbol{s}_j) \tag{8}$$

between USVs $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. We also use two approaches to solve the SSC problem: Orthogonal Matching Pursuit (OMP-SSC) [26] to solve the matrix form of Eq. 1 and the LASSO formulation in Eq. 3 (LASSO-SSC).

Table 1: Results for the clustering methods. CS + SC: spectral clustering using cosine similarity, LASSO-SSC: LASSO-based subspace clustering, OMP-SSC: OMP-based subspace clustering.

Strain	K	Method	Inliers only / With outliers	
			$\bar{d}_{\cos}(\boldsymbol{C})$	$\sigma(d_{\cos}(\boldsymbol{C}))$
DBA	20	k-means	0.374 / 0.372	0.215 / 0.209
		CS + SC	0.428 / 0.418	0.209 / 0.200
		LASSO-SSC	**0.572 / 0.463**	**0.168 / 0.164**
		OMP-SSC	0.296 / 0.315	0.242 / 0.229
	40	k-means	0.383 / 0.380	0.223 / 0.216
		CS + SC	0.448 / 0.439	0.215 / 0.206
		LASSO-SSC	**0.529 / 0.484**	**0.178 / 0.174**
		OMP-SSC	0.212 / 0.225	0.271 / 0.257
	60	k-means	0.383 / 0.381	0.225 / 0.219
		CS + SC	0.448 / 0.439	0.217 / 0.207
		LASSO-SSC	**0.529 / 0.484**	**0.185 / 0.178**
		OMP-SSC	0.192 / 0.203	0.271 / 0.259
C57	20	k-means	0.345 / 0.338	0.189 / 0.181
		CS + SC	0.369 / 0.357	0.174 / 0.166
		LASSO-SSC	**0.438 / 0.391**	**0.157 / 0.135**
		OMP-SSC	0.258 / 0.265	0.205 / 0.190
	40	k-means	0.364 / 0.356	0.189 / 0.181
		CS + SC	0.386 / 0.374	0.174 / 0.166
		LASSO-SSC	**0.453 / 0.403**	**0.158 / 0.138**
		OMP-SSC	0.208 / 0.217	0.203 / 0.212
	60	k-means	0.366 / 0.359	0.196 / 0.188
		CS + SC	0.398 / 0.387	0.183 / 0.174
		LASSO-SSC	**0.461 / 0.417**	**0.161 / 0.141**
		OMP-SSC	0.192 / 0.202	0.223 / 0.206

4.4. Performance measures

Given the lack of a ground truth, we compare our results both qualitatively and quantitatively. First, we compare the low-dimensional representation of the USVs using t-SNE [27], for the case $K = 40$ clusters. For k-means, to reduce the computational complexity, we first use PCA to reduce the dimensions to 100 before applying t-SNE. For the other methods, we use the embedding produced by spectral clustering as input for t-SNE. Since the outlier data set does not have an embedding, we only plot the inliers data sets.

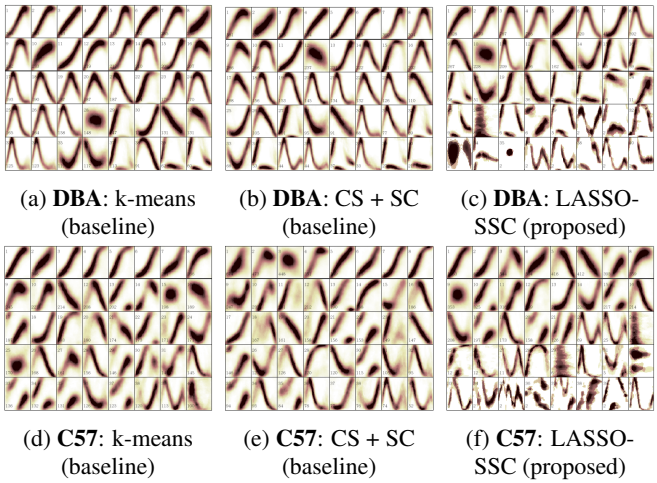
We also compute the harmonic mean of cosine distance between the centroids of the clusters:

$$\bar{d}_{\cos}(\boldsymbol{C}) = \left(\frac{1}{K(K-1)} \sum_{i \neq j} \frac{1}{1 - \cos(\boldsymbol{c}_i, \boldsymbol{c}_j)} \right)^{-1}.$$

Since the harmonic mean is skewed towards small values, a higher harmonic mean of cosine distances can be interpreted as consistently having more diverse centroids. We also compute the standard deviation $\sigma(d_{\cos}(\boldsymbol{c}))$ of cosine distances between centroids. Lower standard deviations indicate that centroids are different from each other more consistently.

[1] `www.github.com/usc-sail/mupet-subspace-clustering`

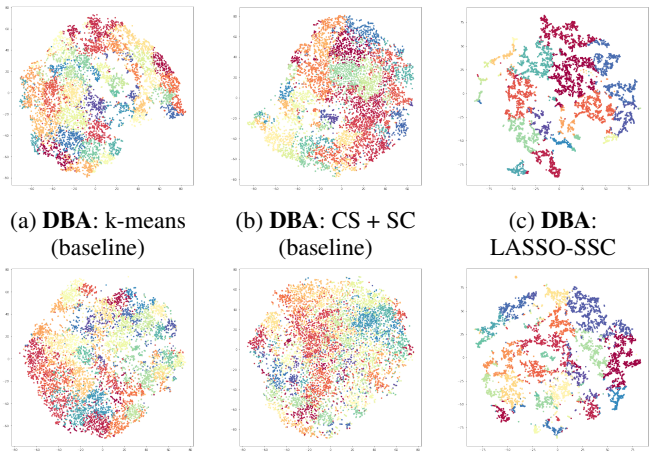
[2] `https://github.com/mvansegbroeck/mupet/wiki/MUPET-wiki` (sample files for DBA and C57).



(a) **DBA**: k-means (baseline) (b) **DBA**: CS + SC (baseline) (c) **DBA**: LASSO-SSC (proposed)

(d) **C57**: k-means (baseline) (e) **C57**: CS + SC (baseline) (f) **C57**: LASSO-SSC (proposed)

Fig. 3: Centroids of clusters of inliers. In each figure, the top left window contains the cluster with most USVs, and decreases towards the right.



(a) **DBA**: k-means (baseline) (b) **DBA**: CS + SC (baseline) (c) **DBA**: LASSO-SSC

(d) **C57**: k-means (baseline) (e) **C57**: CS + SC (baseline) (f) **C57**: LASSO-SSC

Fig. 4: t-SNE visualizations of the clusters in 2 dimensions. The embedding computed from the subspace similarity matrix $\boldsymbol{A}$ is able to better discriminate among different clusters compared to the cosine similarity between feature vectors.

5. RESULTS

We present the quantitative results in Table 1. We observe that the LASSO version of subspace clustering (LASSO-SSC) achieves the best results for all number of clusters K tested in both strains with and without outliers, while the OMP-SSC method yields the worst results across all experiments. When comparing the clustering results on the inlier data set, the best number of clusters is $K = 20$ for DBA. However, for C57 the highest mean cosine distance is achieved for $K = 60$ clusters, while the minimum standard deviation for the cosine distance is achieved at $K = 20$ clusters. When including the outliers, we find that the $\bar{d}_{\cos}(\boldsymbol{C})$ of LASSO-SSC decreases more than the other methods. This is because the outliers highly affect several clusters with a small number of samples produced by LASSO-SSC. However, LASSO-SSC achieves the best result for all the cases with outliers. LASSO-SSC has higher $\bar{d}_{\cos}(\boldsymbol{C})$ because it tends to cluster similar clusters in other methods into one big cluster. Therefore, it has less similar centroids, which leads to higher harmonic means of cosine distances between centroids.

We show a qualitative representation in Fig. 3, where the centroids for each cluster are depicted (top left is the cluster containing the largest amount of USVs; the number gets smaller towards the right). k-means and CS + SC have many centroids with similar shapes (for example, several shapes look very similar in the first rows of Fig. 3(a) and (b) and Fig. 3(d) and (e)), while we observe more variability in the centroids for both DBA and C57 using LASSO-SSC. This can be interpreted as our proposed method clustering together the USVs that correspond to a same class or cluster, yielding clusters with less variability of shapes between USVs. Due to the smaller variability within clusters, we see more clusters with noise (by the end of Fig. 3(c,f)), which are clustered together with different shapes by k-means and CS + SC, producing more noisy clusters.

We show different qualitative results in Fig. 4. These plots show a 2-dimensional representation of the space in which the USVs were clustered, and the colors represent different clusters. We observe that for both strains DBA and C57, the 2-dimensional representations for LASSO-SSC have more distance between clusters than the k-means and CS + SC methods. We omit OMP due to the poor performance shown in Table 1.

6. DISCUSSION

One difficulty in clustering task is that some kinds of vocalizations have very few instances, which makes it hard to cluster them into a category. Sparse subspace clustering has a better ability to discover those categories than k-means, which contributes to the diversity of the clusters that can be seen in Fig. 3. Obtaining higher harmonic mean of cosine distances in the proposed method also indicates that subspace clustering leads to clusters with more diversity between each other. These results confirm our original hypothesis that the information in the frequency domain of the USVs is well-modeled as low-dimensional subspaces in a high-dimensional space.

7. CONCLUSION

In this paper, we propose a pipeline to cluster mouse vocalizations based on a two-step approach using sparse subspace clustering. Our method outperforms the previous methods evaluated both through qualitative plots (using t-SNE, Fig. 4) and our proposed performance measure based on the harmonic mean. Moreover, we find that subspace similarity is a better similarity than cosine similarity to compare USVs. Our approach provides more diverse and cleaner clustering results than previous algorithms.

For future work, we consider several avenues, including: (1) different ways to assess the quality of the clusters, but also looking at the performance of each clustering technique in mice behavior analysis tasks, and (2) including temporal information to aid the clustering.

8. ACKNOWLEDGEMENTS

We thank the USC Viterbi School of Engineering and the Feng Deng Foundation in Tsinghua for their support.

9. REFERENCES

[1] Gillian D Sewell, "Ultrasound in rodents," *Nature*, vol. 217, no. 5129, pp. 682–683, 1968.

[2] Christine V Portfors and David J Perkel, "The role of ultrasonic vocalizations in mouse communication," *Current Opinion in Neurobiology*, vol. 28, pp. 115–120, 2014.

[3] J Chabout, A Sarkar, D. B. Dunson, and E. D. Jarvis, "Male mice song syntax depends on social contexts and influences female preferences," *Frontiers in Behavioral Neuroscience*, vol. 9, no. 76, pp. 76, 2014.

[4] Hui Juan Cao and Jian Ping Liu, "A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (mus musculus domesticus)," *Plos One*, vol. 9, no. 5, pp. e97244–e97244, 2014.

[5] Jacqueline N Crawley, "Translational animal models of autism and neurodevelopmental disorders," *Dialogues in Clinical Neuroscience*, vol. 14, no. 3, pp. 293, 2012.

[6] 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, no. 1, pp. 17–27, 2011.

[7] Markus Wöhr and Rainer K. W. Schwarting, "Affective communication in rodents: ultrasonic vocalizations as a tool for research on emotion and motivation," *Cell and Tissue Research*, vol. 354, no. 1, pp. 81–97, 2013.

[8] Maria Luisa Scattoni, Laura Ricceri, and Jacqueline N Crawley, "Unusual repertoire of vocalizations in adult BTBR T+ tf/J mice during three types of social encounters," *Genes, Brain and Behavior*, vol. 10, no. 1, pp. 44–56, 2011.

[9] Gustavo Arriaga, Eric P Zhou, and Erich D Jarvis, "Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds," *PloS one*, vol. 7, no. 10, pp. e46610, 2012.

[10] Zachary D Burkett, Nancy F Day, Olga Peñagarikano, Daniel H Geschwind, and Stephanie A White, "VoICE: A semi-automated pipeline for standardizing vocal analysis across models," *Scientific Reports*, vol. 5, pp. 10237, 2015.

[11] Maarten Van Segbroeck, Allison T. Knoll, Pat Levitt, and Shrikanth Narayanan, "MUPET-Mouse Ultrasonic Profile ExTraction: A Signal Processing Tool for Rapid and Unsupervised Analysis of Ultrasonic Vocalizations," *Neuron*, vol. 94, no. 3, pp. 465–485, 2017.

[12] Kevin R Coffey, Russell G Marx, and John F Neumaier, "Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," *Neuropsychopharmacology*, vol. 44, no. 5, pp. 859, 2019.

[13] James M. Reno, Bryan Marker, Lawrence K. Cormack, Timothy Schallert, and Christine L. Duvauchelle, "Automating ultrasonic vocalization analyses: The WAAVES program," *Journal of Neuroscience Methods*, vol. 219, no. 1, pp. 155–161, 2013.

[14] David J. Barker, Christopher Herrera, and Mark O. West, "Automated detection of 50-kHz ultrasonic vocalizations using template matching in XBAT," *Journal of Neuroscience Methods*, vol. 236, pp. 68–75, 2014.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candés, "Robust subspace clustering," *Annals of Statistics*, vol. 42, no. 2, pp. 669–699, 2013.

[17] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2790–2797.

[18] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, Nov 2013.

[19] Mahdi Soltanolkotabi and Emmanuel J. Candés, "A geometric analysis of subspace clustering with outliers," *Annals of Statistics*, vol. 40, no. 40, pp. 2012, 2011.

[20] René Vidal, "A tutorial on subspace clustering," 2012.

[21] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[22] Andrew Y Ng, Michael I Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[23] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[24] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *Departmental Papers (CIS)*, p. 107, 2000.

[25] Allison T Knoll, Kevin Jiang, and Pat Levitt, "Quantitative trait locus mapping and analysis of heritable variation in affiliative social behavior and co-occurring traits," *Genes, Brain and Behavior*, vol. 17, no. 5, pp. e12431, 2018.

[26] Chong You, Daniel Robinson, and René Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3918–3927.

[27] Laurens van der Maaten and Geoffrey Hinton, "Viualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.