2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Sensing the Sounds of Silence: A Pilot Study on the Detection of Model Mice of Autism Spectrum Disorder from Ultrasonic Vocalisations

Kun Qian[1,2], *Senior Member, IEEE*, Tomoya Koike[2], *Student Member, IEEE*, Kota Tamada[3], Toru Takumi[3], Björn W. Schuller[4,5], *Fellow, IEEE*, and Yoshiharu Yamamoto[2], *Member, IEEE*

***Abstract*— Studying the animal models of human neuropsychiatric disorders can facilitate the understanding of mechanisms of symptoms both physiologically and genetically. Previous studies have shown that ultrasonic vocalisations (USVs) of mice might be efficient markers to distinguish the wild type group and the model of autism spectrum disorder (mASD). Nevertheless, in-depth analysis of these 'silence' sounds by leveraging the power of advanced computer audition technologies (e. g., deep learning) is limited. To this end, we propose a pilot study on using a large-scale pre-trained audio neural network to extract high-level representations from the USVs of mice for the task on detection of mASD. Experiments have shown a best result reaching an unweighted average recall of 79.2 % for the binary classification task in a rigorous subject-independent scenario. To the best of our knowledge, this is the first time to analyse the sounds that cannot be heard by human beings for the detection of mASD mice. The novel findings can be significant to motivate future works with according means on studying animal models of human patients.**

I. Introduction

Autism spectrum disorder (ASD), *aka* autism spectrum condition (ASC), is considered as a developmental brain disease [1], [2], which is a common and heterogeneous neuropsychiatric disorder that involves deficit in social interaction, speech and nonverbal communication, repetitive behaviour or restricted interest [3]. Autism is thought to be a kind of heritable neuropsychiatric disorder [4], which means the genetic factors contribute significantly to its etiology [5]. Based on the conserved human/mouse linkage, previous studies [6], [7] had generated the mice with a 6.3 Mb duplication of mouse chromosome 7, which was mirroring the human chromosome 15q11-13 duplication (known to be the most frequent cytogenetic abnormality in autism [8]). More interestingly, it was observed that, abnormal vocalisations were found in the model ASD (mASD) mice displaying poor social interaction and behavioural inflexibility [6].

On the one hand, computer audition (CA) and its related audio signal processing and machine learning (ML) and/or deep learning (DL) technologies have been increasingly applied to the field of healthcare [9], e. g., snore sound [10], heart sound [11], and even the ongoing COVID-19 pandemic [12]. On the other hand, according studies on analysing the sounds that cannot be heard by human beings, e. g., ultrasonic vocalisations (USVs) generated by mice, are extremely limited. Specifically, recent studies using CA based methods for analysing the USVs of mice were focusing on human annotated simple behaviours (e. g., courtship [13]) or gender distinction [14]. In contrast, using USVs to detect mASD mice is lacking. Motivated by the previous successful achievements in analysing human speech for ASD detection as a task [15]–[17], we introduce and explore the capacity of advanced CA methods for detection of mASD mice in this pilot study.

The main contributions of this work can be summarised as: First, to the best of our knowledge, it is the first study on using CA for analysing USVs of mice for the detection of mASD. Second, we introduce novel large scale pre-trained deep convolutional neural network models to the field of mice USVs. The models were pre-trained by a large-scale audio data set, rather than the widely used image data for a better fit. Third, we observe that, the CA based method appears available to analyse mental diseases biologically. Last but not least, we demonstrate that, the models trained by audible data (that can be heard by humans), are still effective to extract high-level representations from non-audible data (that cannot be heard by humans). The remainder of this paper will be organised as follows: We firstly introduce the related work by giving a brief literature review. Then, Section III describes the data and methods used in this study. Subsequently, the experimental results are shown in Section IV followed by a discussion in Section V. Finally, this work is concluded in Section VI.

II. Related Work

It is worth noting that, in the recent five years, leveraging the state-of-the-art ML/DL methods for analysing mice's

This work was partially supported by the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, and the Grants-in-Aid for Scientific Research (No. 19F19081 and No. 20H00569) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. Kun Qian and Tomoya Koike contributed equally to this work. Kun Qian is the *Corresponding author*.

[1]Kun Qian is with the Group on Audition for Intelligent Medicine (AIM), Institute of Engineering Medicine, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing 100081, China. qiantum@hotmail.com

[2]Kun Qian, Tomoya Koike, and Yoshiharu Yamamoto are with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. {qian, tommy, yamamoto}@p.u-tokyo.ac.jp

[3]Kota Tamada and Toru Takumi are with the Graduate School of Medicine, Kobe University, 7-5-1 Kusunoki-cho, Chuo-ku, Kobe 650-0017, Japan. {tamada, takumit}@med.kobe-u.ac.jp

[4,5]Björn W. Schuller is with GLAM – the Group on Language, Audio & Music, Imperial College London, 180 Queens' Gate, Huxley Bldg., London SW7 2AZ, UK, and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Eichleitnerstr. 30, Augsburg 86159, Germany. schuller@ieee.org

USVs has increasingly attracted efforts from a broad community of neuroscience, psychiatry, and computer science. A simple task on distinguishing two types, i. e., "part of a call" or "not part of a call" was presented in [18]. The authors claimed that their proposed wavelet transformation based scalograms can be superior to the traditional Fourier transformation (FT) based spectrograms in modelling a mouse's pitch perception [18]. Coffey *et al.* introduced a deep learning paradigm called DeepSqueak, to classify mice volcalisations into five categories: *Split*, *Inverted U*, *Short Rise*, *Wave*, and *Step* [19]. In their study, the cutting-edge regional convolutional neural network architecture (Faster-RCNN) [20] was used, which was shown to be effective for nuanced explorations of the interplay between vocalisations and behaviours, even in noisy environmental conditions [19]. Vogel *et al.* studied the classic ML paradigm, which used human hand-crafted acoustic features and ML models for classifying nine types of mice USVs, i. e., *complex*, *two components*, *upward*, *downward*, *chevron*, *short*, *composite*, *frequency step*, and *flat* [21]. They indicated that, a random forest outperformed a support vector machine, which can achieve a promising result of approximately 85.0 % classification accuracy [21]. Moreover, Ivanenko *et al.* investigated the capacity of a deep neural network (DNN) for classifying sex and strain from the mice USVs [14]. They claimed that, a sufficient nonlinear combination of features extracted from the spectrograms of the mice USVs can facilitate the classification of emitter's sex and/or strain. Sangiamo *et al.* combined the sound source localisation technology with the ML/DL based classification system to analyse the mice behaviour types in [13]. They found a clear pattern linking particular social behaviours and vocal communication in male mice [13]. A study on learning the dictionary of the mice USVs were proposed in [22]. The authors proposed a hybrid approach between sparse subspace clustering and more traditional clustering techniques, and found that the subspace similarity is a better similarity than cosine similarity to compare USVs [22].

However, the existing studies aforementioned cannot answer the question on if CA based methods can be feasible for detection of the mASD mice. To this end, we conduct this pilot study on using advanced DL models to extract high-level representations from the USVs of mice. Furthermore, we use this paradigm to classify the groups of the wild type (WT) control and the mASD.

III. Materials and Methods

A. Data Collection and Protocols

The detailed data collection protocol can be found in [7]. The mice USVs data collection environment is briefly illustrated in Fig. 1. Each pup was separated from its dam and placed into a plastic tray at P 8 (postnatal day 8) and P 12 (postnatal day 12). The microphone (416 H, Avisoft Bioacoustics) was located 10 cm above the bottom of the field (see Fig. 1). The recording time was set to 5 min at a sampling rate of 300 kHz. After recording on P8, each pup was labelled by tattoo in tail, which used non-toxic ink to identify each mouse. The investigators were blind to the genotypes of the mice. Totally, we have collected 168 recordings of USVs. Among of these recordings, 88 are labelled as 'WT', and 80 are labelled as 'mASD'.

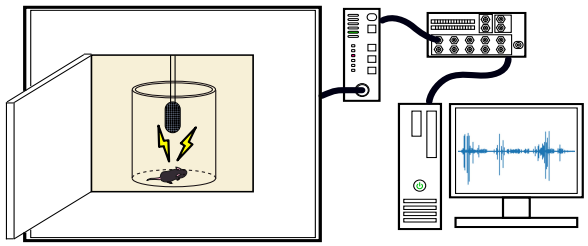
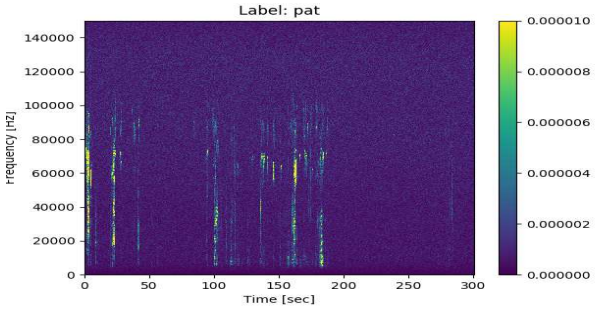
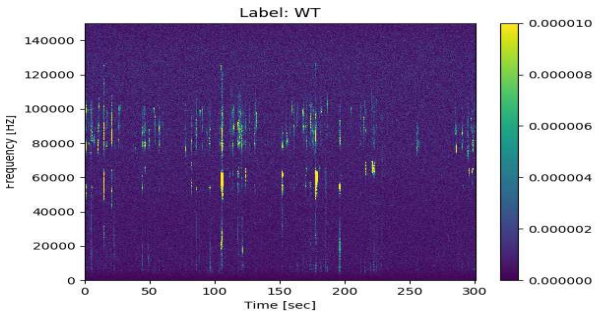


Fig. 1. An overview of the mice USVs data collection. The source of this figure is from [23]. Permission was received from Elsevier.



(a) mASD



(b) WT

Fig. 2. The example spectrograms of the USVs generated from the model mice of ASD and the mice of wild type (WT) control.

The spectrograms can be achieved via the short-time Fourier transformation (STFT) [24], which are widely used to analyse audio data in the time-frequency domain. Fig. 2 illustrates examples of spectrograms of the USVs generated from the mASD mice and the WT mice.

B. Deep Transfer Learning Models

The general framework of the proposed method is depicted in Fig. 3. Kong *et al.* proposed novel deep transfer learning pre-trained models, the large-scale pre-trained audio neural networks for audio pattern recognition (PANNs) [25], which were validated successfully in our previous CA for healthcare (CA4H) applications of heart sound classification [26] and speech under facial mask detection task [27]. PANNs were pre-trained by the large-scale Audio Set [28], instead of the up-to-now conventionally used image data, which makes

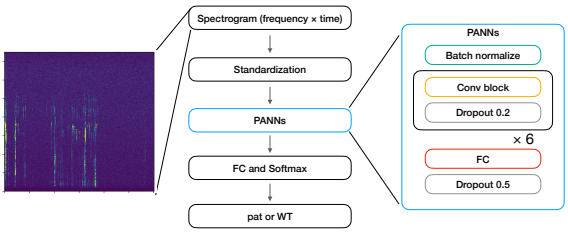


Fig. 3. The proposed framework for detection of mASD from the USVs of mice. Firstly, the original USVs are transformed into spectrograms based by STFT. Then, the pre-trained DL model, i. e., PANNs (CNN 14) [25] are used to extract high-level representations from the data. Finally, a softmax layer is used to make predictions on the data as mASD or WT.

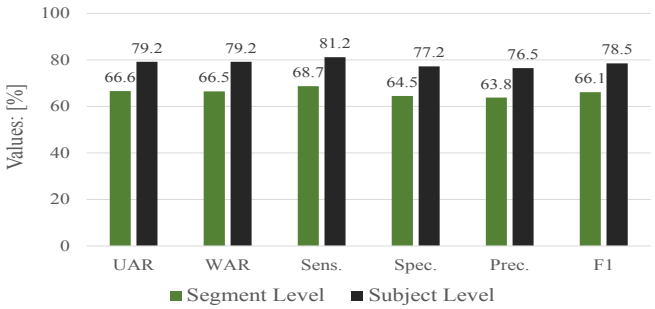


Fig. 4. The evaluation metrics (in [%]) of the experimental results for both the segment level and the subject level based instances. All the results are averaged values of a 5-fold cross validation.

them more suitable to extract high-level representations from the USVs.

One of the structures of PANNs, called CNN 14, is composed of 6 layers of convolutional blocks and 2 fully connected layers. A convolutional block has 2 layers of 3×3 convolutional filters, a batch normalisation layer, and a Rectified Linear Unit (ReLU), followed by 2×2 pooling layers. Input layer and output layer are changed accordingly to fit USV data and label.

The loss function for fine-tuning PANNs' CNN 14 is binary cross-entropy or log loss which is defined as:

$$Log\,Loss = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right], \qquad (1)$$

where n is the number of instances, $\hat{y}_i$ is the predicted probability of genetically modified mouse type, y_i is 1 (mASD) or 0 (WT).

C. Evaluation Metrics

Considering the data imbalance between mASD and WT mice, we use the unweighted average recall (UAR) [29], i. e., the averaged *recall* of the two classes, as the main evaluation metric. In addition, the widely used *accuracy*, i. e., weighted average recall (WAR), sensitivity (Sens.), specificity (Spec.), precision (Prec.), and F1 measure (F1) are used as complementary metrics for evaluating the proposed model's performance.

IV. Experimental Results

A. Setup

We use Python based scripts via PyTorch (Version–1.5.1) to run all the experiments in this study. All the original USVs are transformed to spectrograms via STFT. The original long USVs recordings (duration: 5 minutes) are chunked into shorter clips with a duration of 30 seconds and an overlap of 15 seconds between the neighbouring segments. We use a 5-fold cross validation strategy to train and validate the models. The final results are the averaged values of 5 times independent experiments: In each experiment, four folds of the segmented instances are used to train the model while the remaining fold of the segmented instances is used to validate the model. In order to avoid over-optimistic results, a rigorous subject-independent method is applied to the fold partitioning. For the final validation, we use two methods, i. e., segment level and subject level. For segment level validation, the final predictions are based on the segment based instances. In contrast, for subject level validation, the final predictions are made based on a *majority voting* of the segment level instances belonging to one certain subject. All the extracted high-level representations are standardised to eliminate the effects of outliers.

TABLE I

Confusion matrices (normalised: in [%]) of the two considered methods (segment level vs subject level). All of the instances are included by excluding each subject within a 5-fold subject-independent cross-validation. **mASD**: model Autism Spectrum Disorder; **WT**: Wild Type.

(a) Segment Level

Pred ->	**WT**	**mASD**
WT	64.6	35.4
mASD	31.3	68.7

(b) Subject Level

Pred ->	**WT**	**mASD**
WT	77.3	22.7
mASD	18.8	81.2

B. Results

The evaluation metrics (in [%]) are shown in Fig. 4. We can see that, all the results based on the subject level strategy are better than the counterparts via segment level strategy. When looking at the UARs, both of the two strategies (66.6 % vs 79.3 %) have shown an effective capacity in the detection of mASD, i. e., much higher than the according chance level of 50.0 % UAR for two classes. The normalised confusion matrices are given in Table I. It is found that, *majority voting* can improve the recalls for both mASD and WT USVs. Specifically, the recall of WT can be improved from 68.7 % to 81.2 %, which lends the model a higher specificity (see Fig. 4).

V. Discussion

As a first study on using CA technologies to analyse the USVs for detection of model mice with ASD, we demonstrate that, the models pre-trained by audible data can also be used for extracting high-level representations from non-audible data, particularly for a healthcare related task. The experimental results are encouraging and promising. Furthermore, a majority voting strategy can significantly

($p < .001$ in a one-tailed Student's t-test [30]) improve the final performance of the model. The sensitivity can reach a high result surpassing 80.0 % (see Fig. 4), which is already comparable or even better than the previous study on analysing human speech for ASD detection [17].

The limitations and perspectives of this pilot study are: First, the explainability of the proposed DL models is lacking. In a next step, one needs to explore the visualisation of learnt features from the USVs data by the DL models, which will be benefiting the understanding of the mechanisms why and how the models work well for the considered mASD mice detection task. Second, we will investigate and compare different topologies of the pre-trained models in the USVs analysis work. Moreover, a combination (fusion) of the models might improve the performance of the models. Third, traditional human hand-crafted features carrying important information about the pathological vocalisations should be studied for comparison in this context. One should also study how to combine the state-of-the-art DL methods with the classic ML models to achieve better results. Finally, we need to note that, the whole spectrum of the USVs are used in this study. Namely, some audible events (e. g., the movement sounds of the mice) might be included in the analysed data. Future work can be done by excluding these parts of the USVs data.

VI. Conclusion

In this study, we investigated using advanced CA based methods for analysing USVs of mice, particularly for the task of mASD detection. The novel PANNs were firstly introduced into the field of mice USVs analysis and demonstrated to be efficient to reach a UAR of 79.2 % in a rigorous subject-independent scenario for the detection of model mice of ASD. This promising result shows that, DL models can extract high-level representations from the sounds beyond human hearing capacity, and these features can be useful for detecting vocalisations of the model ASD mice sharing biological/genetic backgrounds with the human patients.

Acknowledgment

The authors would like to thank all the colleagues involved in the data collection work.

References

[1] M. K. Belmonte *et al.*, "Autism as a disorder of neural information processing: directions for research and targets for therapy," *Molecular Psychiatry*, vol. 9, no. 7, pp. 646–663, 2004.

[2] E. DiCicco-Bloom *et al.*, "The developmental neurobiology of autism spectrum disorder," *Journal of Neuroscience*, vol. 26, no. 26, pp. 6897–6906, 2006.

[3] C. Lord *et al.*, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, 2018.

[4] D. H. Geschwind and P. Levitt, "Autism spectrum disorders: Developmental disconnection syndromes," *Current Opinion in Neurobiology*, vol. 17, no. 1, pp. 103–111, 2007.

[5] J. Vorstman *et al.*, "Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism," *Molecular Psychiatry*, vol. 11, no. 1, pp. 18–28, 2006.

[6] J. Nakatani *et al.*, "Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism," *Cell*, vol. 137, no. 7, pp. 1235–1246, 2009.

[7] N. Nakai *et al.*, "Serotonin rebalances cortical tuning and behavior linked to autism symptoms in 15q11-13 CNV mice," *Science Advances*, vol. 3, no. 6, pp. e1 603 001: 1–13, 2017.

[8] A. Hogart *et al.*, "Chromosome 15q11–13 duplication syndrome brain reveals epigenetic alterations in gene expression not predicted from copy number," *Journal of medical genetics*, vol. 46, no. 2, pp. 86–93, 2009.

[9] K. Qian *et al.*, "Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, no. 5, pp. 1–4, 2020.

[10] K. Qian *et al.*, "Can machine learning assist locating the excitation of snore sound? A review," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2020, in press.

[11] F. Dong *et al.*, "Machine listening for heart status monitoring: Introducing and benchmarking HSS–the heart sounds Shenzhen corpus," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2082–2092, 2020.

[12] K. Qian, B. W. Schuller, and Y. Yamamoto, "Recent advances in computer audition for diagnosing COVID-19: An overview," in *Proc. LifeTech*. Nara, Japan: IEEE, 2021, pp. 185–186.

[13] D. T. Sangiamo, M. R. Warren, and J. P. Neunuebel, "Ultrasonic signals associated with different types of social behavior of mice," *Nature Neuroscience*, vol. 23, no. 3, pp. 411–422, 2020.

[14] A. Ivanenko *et al.*, "Classifying sex and strain from mouse ultrasonic vocalizations using deep learning," PLoS *Computational Biology*, vol. 16, no. 6, pp. 1–27, 2020.

[15] E. Marchi *et al.*, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Proc. INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 115–119.

[16] M. Schmitt *et al.*, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *Proc. ITG*. Paderborn, Germany: VDE, 2016, pp. 264–268.

[17] F. B. Pokorny *et al.*, "Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach," in *Proc. INTERSPEECH*. Stockholm, Sweden: ISCA, 2017, pp. 309–313.

[18] A. A. Smith and D. Kristensen, "Deep learning to extract laboratory mouse ultrasonic vocalizations from scalograms," in *Proc. BIBM*. Kansas, MO, USA: IEEE, 2017, pp. 1972–1979.

[19] 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, no. 5, pp. 859–868, 2019.

[20] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] 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, no. 1, pp. 1–10, 2019.

[22] J. Wang *et al.*, "Bringing in the outliers: A sparse subspace clustering approach to learn a dictionary of mouse ultrasonic vocalizations," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 3432–3436.

[23] T. Takumi *et al.*, "Behavioral neuroscience of autism," *Neuroscience & Biobehavioral Reviews*, vol. 110, pp. 60–76, 2020.

[24] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Speech Processing (1st Edition)*. Upper Saddle River, NJ, USA: Pearson Higher Education, Inc., 2010.

[25] Q. Kong *et al.*, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[26] T. Koike *et al.*, "Audio for audio is better? An investigation on transfer learning models for heart sound classification," in *Proc. EMBC*. Montréal, Canada: IEEE, 2020, pp. 74–77.

[27] T. Koike *et al.*, "Learning higher representations from pre-trained deep models with data augmentation for the ComParE 2020 challenge mask task," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 2047–2051.

[28] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. New Orleans, LA, USA: IEEE, 2017, pp. 776–780.

[29] K. Qian, *Automatic General Audio Signal Classification*. Munich, Germany: Technical University of Munich, 2018, Doctoral Thesis.

[30] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.