Exploring Pre-trained models on Ultrasound Modeling for Mice Autism Detection with Uniform Filter Bank and Attentive Scoring

Yuchen Song[1], *Yucong Zhang*[1,2], *Ming Li**[1,2]

[1]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center, Duke Kunshan University, Kunshan, China
[2]School of Computer Science, Wuhan University, Wuhan, China

ming.li369@dukekunshan.edu.cn

Abstract

Genetically engineered mice, whose behaviors resemble those of individuals with Autism Spectrum Disorder (ASD), serve as valuable models for studying ASD through ultrasound vocalization (USV) analysis. In this paper, we investigate the effectiveness of pre-trained models in learning features of the USVs by fine-tuning. To bridge the gap between the pre-trained model and the inductive bias of the ultrasonic signal, we design a uniformly-spaced filter bank to reduce the dimension in the frequency domain. The extracted filter-bank energies of the ultrasonic spectrogram form a pseudo-spectrogram for pre-trained models. In the back-end, we employ an attentive frame-wise scoring method for classification, resulting in a comprehensive judgment. Experimental results demonstrate the effectiveness of our approach, achieving a segment-level Unweighted Average Recall (UAR) of 0.729 and a subject-level UAR of 0.882 on the validation set provided by the MADUV 2025 Challenge.

Index Terms: mice autism spectrum disorder detection, ultrasound vocalizations, bioacoustic feature analysis

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurological and developmental disorder characterized by deficits in social communication and interaction, as well as restricted and repetitive behaviors [1, 2]. The etiology of ASD is multifactorial, involving a combination of genetic and environmental factors that disrupt typical brain development. Experimental animal models are indispensable for advancing our understanding of the underlying causes and developmental trajectories of human diseases, including ASD [3, 4]. These models allow researchers to explore mechanisms that cannot be easily studied in humans and are essential for evaluating potential treatments. Generally, ASD animal models can be classified into two main categories: those induced by environmental factors and those generated through genetic manipulations [3]. In particular, genetically induced ASD models, particularly mouse models, have been instrumental in elucidating the genetic basis of the disorder. Mice share significant genetic similarities with humans, including conserved genomic regions associated with ASD [5]. By using genetic engineering techniques, researchers can create mouse models that replicate key aspects of human ASD symptoms, enabling them to test potential treatments and identify genetic alterations that may also be relevant to human populations [3, 6].

Among the various studies using mouse models, one key area of focus is analyzing mice's behavioral patterns through their Ultrasonic Vocalizations (USVs) and examining how these vocalizations relate to social behavior [7]. Numerous studies have demonstrated significant differences between the USVs of genetically engineered mice and those of wild-type mice [8, 9]. Building on these biological insights, the MADUV 2025 Challenge [10] introduces a novel interdisciplinary research field that integrates speech processing techniques with biology and neuroscience. The challenge provides audio data from both ASD and wild-type mice, aiming to advance speech-technique-aided classification and diagnosis of ASD in animal models.

While previous studies have explored ASD detection through the analysis of human vocalizations [11, 12, 13], this challenge takes a groundbreaking step by focusing on non-human vocalizations. A key difficulty lies in the fact that most existing speech processing technologies are designed for human speech, which primarily occupies lower frequency ranges. In contrast, mice produce USVs at relatively high frequencies, posing a unique challenge.

In this paper, we explore the effectiveness of audio pre-trained models on the mice ASD task of mice ultrasound vocalization processing. In order to bridge the gap between the ultrasound spectrogram and the audio spectrogram, we introduce a set of uniform filter banks to model USVs. The obtained filter bank energies serve as pseudo-spectrogram input for fine-tuning the BEATs [14] and CED [15], which are pre-trained on AudioSet [16], as the backbone models. We also introduce an attention-based frame-wise scoring method for the classification task to deal with the frequency characteristics that exhibit significant temporal fluctuations in mice USVs. We achieved a segment UAR of 0.729 and a subject UAR of 0.882 on the validation set of the MADUV 2025 challenge. We release our code at [1].

2. Related works

Previous studies in humans ASD have identified prosodic differences between individuals with and without ASD, revealing that individuals with ASD tend to exhibit a slower speech rate [17], a more melodic intonation style [18], and higher pitch values along with abnormal high-frequency components [19].

These findings underscore the distinct acoustic characteristics associated with ASD, suggesting that vocal patterns could serve as potential biomarkers for diagnosis. Building on this, researchers have increasingly leveraged machine learning and speech processing techniques to automatically detect ASD using human speech. In 2013, the Interspeech Computational Paralinguistics Challenge introduced tasks aiming at studying autism and its manifestations in speech [20]. Asgari et al. [21] proposed a novel speech feature extraction algorithm that im-

[*] Corresponding author: Ming Li.

[1] https://github.com/EEugeneS/Method-for-MADUV-2025-Challenge

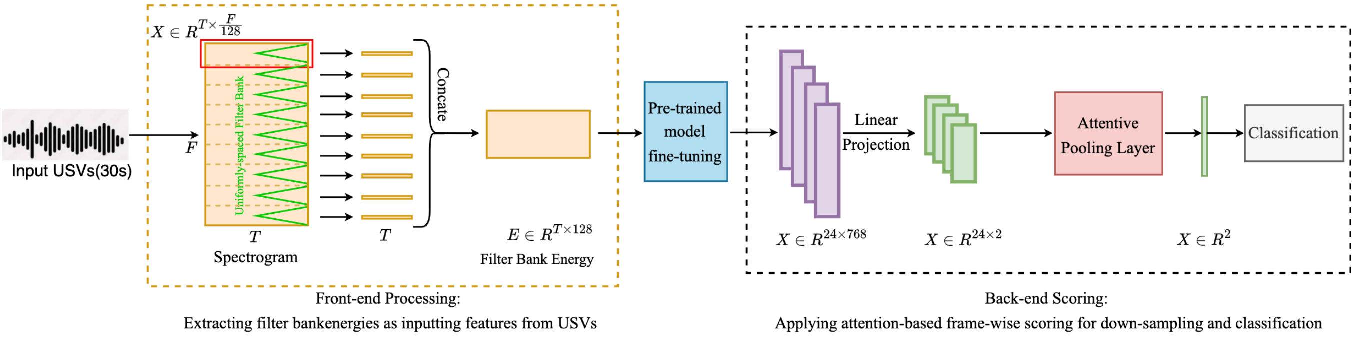
$X \in R^{T\times \frac{F}{128}}$
Uniformly-spaced Filter Bank
Input USVs(30s)
F
T
Spectrogram
T
Concate
$E \in R^{T\times 128}$
Filter Bank Energy
Front-end Processing:
Extracting filter bankenergies as inputting features from USVs
Pre-trained model fine-tuning
Linear Projection
Attentive Pooling Layer
Classification
$X \in R^{24\times 768}$
$X \in R^{24\times 2}$
$X \in R^{2}$
Back-end Scoring:
Applying attention-based frame-wise scoring for down-sampling and classification

Figure 1: *The overview of our proposed framework. F and T stands for the dimensions for frequency and time domain in the spectrogram respectively. E denotes the filter bank energy.*

proved upon existing methods by estimating harmonic model parameters and deriving features such as Harmonic-to-Noise Ratio (HNR), shimmer, and jitter. By combining these features with standard acoustic features and employing Support Vector Machines (SVMs) for classification, their approach achieved a 2.3% and 2.8% improvement in UAR over baseline results for detecting ASD and classifying individuals into four subtypes. Later, Baird et al. [22] explored multi-class classification of vocalizations from children with varying autism severity using a novel dataset. Their research applied feature extraction methods based on the Interspeech Computational Paralinguistics Challenge [23] feature sets and used SVM for classification. Additionally, they investigated the use of convolutional neural networks (CNNs) to analyze spectrogram representations of autistic speech data, further expanding the application of deep learning techniques in ASD detection. Recently, Chi et al. [24] compared the performance of Random Forest, CNNs, and a fine-tuned Wav2Vec 2.0 model for autism detection, demonstrating the significant advantage of deep learning models over traditional machine learning approaches.

Mice USVs, which have been found to be related to characteristics such as sex, health, and health conditions [25], have been used as a model for understanding human communication [26]. However, mouse vocalizations differ significantly from human speech, as they occur in the ultrasonic frequency range. Fonseca et al. [27] developed a software called VocalMat, which applies image processing and machine learning for multi-class labeling to detect mice USVs in spectrograms. Premoli et al. proposed a CNN-based framework for automatic mice vocalization classification, with experimental results showing that, compared to traditional supervised machine learning methods like SVM and Random Forest, considering the entire time/frequency information of the spectrograms led to a significant performance improvement [28]. In a similar study [29] conducted on the challenge dataset, Qian et al. explored the inaudible portion of the USVs for ASD diagnosis using a large-scale pre-trained audio model, PANNs (CNN14) [30]. Their approach achieved a segment-level UAR of 0.666 and a subject-level UAR of 0.792 for this binary classification task.

3. Methodology

We fine-tune pre-trained models to learn feature representations of ultrasound vocalizations obtained by a novel uniformly spaced filter bank. Specifically, the filter bank reduces the frequency-domain dimensionality of ultrasound signals, extracting filter-bank energies to construct a pseudo-spectrogram, which then serves as input for model fine-tuning. Additionally, we also propose an attention-based frame-wise scoring method in the back-end to capture essential frequency changes along frames. As illustrated in Figure 1, the input audio is first transformed into a spectrogram representation, which is then preprocessed by our proposed uniformly spaced filter bank and concatenated into a low-dimensional filter bank energy feature, which is then encoded using a pre-trained model for fine-tuning. The framework further incorporates an output module designed to down-sample the encoded embeddings and perform classification. In Section 3.1, we explain the proposed uniformly spaced filter bank method. In Sections 3.2, we introduce the pre-trained BEATs model [14] and briefly discuss the fine-tuning strategies. Section 3.3 presents the proposed attention-based frame-wise scoring method and provides a detailed explanation of its design and implementation.

3.1. Uniformly Spaced Filter Bank

As shown in Figure 1 our method involves transforming high-frequency ultrasonic audio into a spectrogram and applying Uniformly Spaced Filter Bank to extract filter-bank energies as a pseudo low-dimensional spectorgram. Specifically, by applying the filter bank, the full frequency range is partitioned into 128 equal sub-band images, and the mean value within each images is computed along the frequency-domain. This process results in a compact feature representation of shape $[t, 128]$, where t represents the time dimension.

The key advantage of this approach is its ability to significantly reduce the dimensionality of the ultrasonic spectrogram while retaining crucial frequency information. By downsampling the ultrasonic signals, this method enables us to fine-tuning models that are pre-trained on human-audible sound using USVs. Additionally, by averaging within each sub-band image, it smooths out local variations, mitigating the effects of noise and redundant high-frequency components.

In Section 4.5, we compare the performance of different window length parameters when extracting the spectrogram during the feature extraction.

3.2. Pre-trained Models and Fine-tuning

In this section we will provide a brief introduction of the pre-trained models explored and employed in our system, and the fine-tuning strategies.

BEATs [14] is a self-supervised learning framework for audio pre-training that utilizes an acoustic tokenizer to learn meaningful audio representations. Inspired by NLP pre-training methods, BEATs first trains an acoustic tokenizer to discretize audio signals into token-like units, following the human auditory perception by focusing on high-level audio semantics. BEATs achieves a state-of-the-art results in various audio understanding tasks, such as speech, music and environmental sound classification.

CED [15] is an augmentation and knowledge distillation method designed for audio tagging. It uses the Vision Transformer (ViT) as its backbone model, leveraging its scalability and ability to handle varying input lengths. By transferring the advantages of ensemble models to a more efficient, lightweight student model, CED significantly enhances audio tagging accuracy while reducing computational complexity.

In our framework, we perform full model fine-tuning on both pre-trained models rather than training from scratch, given the limited dataset size. Fine-tuning is particularly effective in this scenario, as it leverages the knowledge already learned by the pre-trained models, reducing the need for large amounts of data and mitigating the risk of overfitting [31].

In Section 4.5, we compare the performance of these two pre-trained models.

3.3. Attention-based Frame-wise Scoring Method

For the classification task, we propose an attention-based frame-wise scoring method to adaptively down-sample the feature embeddings obtained from the backbone model. As illustrated in Figure 1, our approach consists of a dropout layer, a linear projection layer, and an attentive pooling layer.

Specifically, given an input embedding $X \in \mathbb{R}^{B\times T\times F}$, where B, T, and F denote batch size, time steps, and frequency feature dimensions, respectively, we first apply a linear transformation to project the feature dimension from F to 2, resulting in $X' \in \mathbb{R}^{B\times T\times 2}$. This transformation condenses the frequency domain information into two representative logits per frame.

Next, we employ an attention-based pooling layer [32] to adaptively assign importance scores α_t to different frames, ensuring the model focuses on the most informative time steps. The final representation is computed as a weighted sum of transformed frame-level embeddings:

$$X_{\text{final}} = \sum_{t=1}^{T} \alpha_t X''_t, \quad X_{\text{final}} \in \mathbb{R}^{B\times 2} \tag{1}$$

This process effectively aggregates the sequence into a compact form that serves as the input for the classification task. By leveraging attention mechanisms, our method enhances key temporal patterns while filtering out noise and irrelevant variations. Given the temporal variability in mice vocalizations—where frequency characteristics can fluctuate significantly over time—our approach dynamically weights frames based on their contribution to classification. This allows us to capture essential frequency changes while reducing the influence of less relevant information, ultimately leading to more robust and accurate predictions.

In Section 4.5, we present experiments that demonstrate the effectiveness of the scoring method.

4. Experiments

4.1. Dataset

The MADUV 2025 challenge provides a dataset consisting of recordings from 84 mouse subjects, including 44 wild-type and 40 ASD model type [10]. Each subject was recorded once at Postnatal Day 8 (P08) for five minutes using high-precision microphones at a 300 kHz sampling rate, resulting in approximately 7 hours of audio.

Table 1: *Statistics on number of audio segments in the dataset*

	Train	Valid	Test	Total
All	51	17	16	84
With ASD	27	7	6	40
Wild Type	24	10	10	44

To ensure balanced distributions of ASD model type, the dataset was divided into training (51 subjects), validation (17 subjects), and test (16 subjects) sets. The test set was further segmented into 160 non-overlapping 30-second clips, with labels removed and shuffled for evaluation. In contrast, training and validation recordings remain unsegmented. The dataset distribution is summarized in Table 1.

4.2. Implementation Details

For data processing, we follow the baseline approach in [10], segmenting each audio recording in the training and validation sets into 30-second clips with a 15-second overlap. This segmentation ensures that the model captures sufficient context while preserving relevant temporal information. Spectrograms are then computed from the segmented audio using an NFFT of 300k, a hop length of 150k, and a window length of 300k. Next, we apply our proposed uniformly spaced filter bank method to extract filter bank energies with dimensions of $[t, 128]$ and $[t, 64]$ from each spectrogram. The 128-dimensional representation aligns with the BEATs [14] structure, while the 64-dimensional representation is suited for the CED [15] framework.

For the pre-trained model, we load the BEATs$_{\text{iter3}}$ checkpoints for both backbone models from their official GitHub repositories. During training, we follow the challenge baseline paper and use the binary cross-entropy loss function, optimizing the model with the ADAM optimizer [33]. A warm-up scheduler is applied, starting with an initial learning rate of $2e-5$ and gradually increasing to a target learning rate of $1e-4$ over 1000 warm-up steps. The batch size is set to 250, and the model is trained for 200 epochs.

4.3. Metrics

Following the baseline paper of the MADUV challenge [10], Segment Unweighted Average Recall (UAR) and Subject UAR are used as the criterion, with the following definition:

$$\text{UAR} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{2}$$

where C is the total number of classes, TP_c represents the number of true positive samples for class c, and FN_c denotes the false negatives for class c.

4.4. Performance Comparison

We compare the performance of our proposed model with the challenge baseline model, as shown in Table 2. On the validation set, our model achieves a segment UAR of 0.729 and a subject UAR of 0.882, representing relative improvements of 0.047 and 0.069, respectively, over the best-performing baseline. These results highlight the effectiveness of our approach

Table 2: *Comparison of performance between the proposed and baseline models on the validation set*

Model	Feature	Validation Set Segment UAR	Subject UAR
baseline [10]	full	0.675	0.813
baseline [10]	ultra	0.664	0.819
baseline [10]	audi	0.682	0.813
Ours	-	**0.729**	**0.882**

in capturing more discriminative features and improving classification accuracy.

4.5. Ablation Study

Table 3: *Comparison of the best performance of BEATs and CED pre-trained models on the validation set*

Pre-trained Model	Segment UAR	Subject UAR
BEATs	0.729	0.882
CED	0.643	0.750

4.5.1. Pre-trained Models

Table 3 presents a comparison between pre-trained BEATs and CED models. BEATs significantly outperform CED, achieving 0.729 and 0.882 in segmental and subject UAR, respectively, compared to 0.643 and 0.750 for CED. The performance advantage of BEATs can be attributed to its self-supervised learning (SSL) approach, which enables it to learn robust audio representations directly from data, making it more adaptable to ultrasound audio. In contrast, CED relies on knowledge distillation from pre-trained teacher models, which may not generalize well to non-speech audio like ultrasound signals.

Table 4: *Comparison of performance between different input feature dimensions and different types of pooling layer on the validation set using pre-trained BEATs*

Feature Dim	Pooling Layer	Segment UAR	Subject UAR
[181,128]	AttAvgPool	0.706	0.764
[91,128]	AttAvgPool	0.716	0.764
[61,128]	AttAvgPool	**0.729**	**0.882**
[61,128]	AvgPool	0.688	0.750

4.5.2. Input Feature Dimension

Lines 1-3 in Table 4 present a comparison of the performance across different input feature dimensions on the validation set. The variations in time axis of the feature dimensions are the result of setting different window length, 100k, 200k and 300k respectively, when extracting spectrogram from raw audio data. From the results, it is evident that using an input size of $[61, 128]$ yields the best performance. This improvement can be attributed to the fact that the $[61, 128]$ is obtained from a spectrogram extracted with the same window length, 300k as the baseline paper [10]. In contrast, when the input dimension deviates from this size, the feature distribution becomes misaligned, which in turn hampers the effectiveness of fine-tuning.

4.5.3. Pooling Layer

Lines 3 and 4 in Table 4 compare two pooling layers for downsampling embeddings from the backbone model. Our proposed Attentive Average Pooling (AttAvgPool) significantly outperforms standard Average Pooling (AvgPool), despite both using the same input dimension of $[61, 128]$. This highlights the benefits of adaptive frame-wise weighting, allowing the model to emphasize the most relevant frames for classification. By dynamically adjusting frame importance, our method effectively handles the temporal variability in mice vocalizations, demonstrating its superiority in enhancing classification performance.

Table 5: *Comparison of performance between different spectrogram augmentation mask size on the validation set using pre-trained BEATs*

TimeMask	FreqMask	Segment UAR	Subject UAR
5	10	0.713	0.826
10	15	**0.729**	**0.882**
15	20	0.683	0.833
20	25	0.669	0.694

4.5.4. Spectrogram Augmentation

As shown in Figure 1, we apply spectrogram augmentation [34] to enhance model robustness by randomly masking spectrogram regions during training. Table 5 compares model performance with different time and frequency mask sizes on the validation set. The best results—Segment UAR of 0.729 and Subject UAR of 0.882—are achieved with a time mask of 10 and a frequency mask of 15, suggesting that moderate masking helps the model learn more generalized features. However, larger masks (e.g., 15, 20 and 20, 25) degrade performance due to excessive information loss, while smaller masks (e.g., 5, 10) lead to overfitting, highlighting the need for balanced augmentation.

5. Conclusion

In this work, we explore the effectiveness of pre-trained models for ultrasound vocalization (USV) analysis in genetically engineered mice, a crucial step toward improving Autism Spectrum Disorder (ASD) detection. To address the gap between models pre-trained on audible data and the high-dimensionality feature of ultrasonic signals, we design a uniformly spaced filter bank to reduce frequency-domain dimensionality, enabling more effective feature extraction. The resulting filter bank energy features are used to fine-tune pre-trained models, while an attentive frame-wise scoring method provided a robust classification framework. Experimental results demonstrate the superiority of our approach, achieving segment-level and subject-level UARs of 0.729 and 0.882 on the MADUV 2025 Challenge validation set. These findings highlight the potential of our method in advancing automated ASD analysis in mouse models and pave the way for further research on pre-trained model adaptation for bioacoustic signal processing.

6. Acknowledgement

This research is funded by the DKU foundation project ”Interdisciplinary Signal Processing Technologies”. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

7. References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The lancet*, vol. 392, no. 10146, pp. 508–520, 2018.

[2] M. Varghese, N. Keshav, S. Jacot-Descombes, T. Warda, B. Wicinski, D. L. Dickstein, H. Harony-Nicolas, S. De Rubeis, E. Drapeau, J. D. Buxbaum *et al.*, "Autism spectrum disorder: neuropathology and animal models," *Acta neuropathologica*, vol. 134, pp. 537–566, 2017.

[3] Z. Ergaz, L. Weinstein-Fudim, and A. Ornoy, "Genetic and non-genetic animal models for autism spectrum disorders (asd)," *Reproductive Toxicology*, vol. 64, pp. 116–140, 2016.

[4] E. Ey, C. S. Leblond, and T. Bourgeron, "Behavioral profiles of mouse models for autism spectrum disorders," *Autism research*, vol. 4, no. 1, pp. 5–16, 2011.

[5] L. van der Weyden and A. Bradley, "Mouse chromosome engineering for modeling human disease," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, no. 1, pp. 247–276, 2006.

[6] E. Lee, J. Lee, and E. Kim, "Excitation/inhibition imbalance in animal models of autism spectrum disorders," *Biological psychiatry*, vol. 81, no. 10, pp. 838–847, 2017.

[7] 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.

[8] J. Nakatani, K. Tamada, F. Hatanaka, S. Ise, H. Ohta, K. Inoue, S. Tomonaga, Y. Watanabe, Y. J. Chung, R. Banerjee *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.

[9] M. Woehr, "Ultrasonic vocalizations in shank mouse models for autism spectrum disorders: detailed spectrographic analyses and developmental profiles," *Neuroscience & Biobehavioral Reviews*, vol. 43, pp. 199–212, 2014.

[10] Z. Yang, M. Song, X. Jing, H. Zhang, K. Qian, B. Hu, K. Tamada, T. Takumi, B. W. Schuller, and Y. Yamamoto, "Mad-uv: The 1st interspeech mice autism detection via ultrasound vocalization challenge," *arXiv preprint arXiv:2501.04292*, 2025.

[11] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, P. Robinson, I. Davies, O. Golan, S. Friedenson, S. Tal, S. Newman *et al.*, "Asc-inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions," in *Proc. of IDGEI*, Chania, Greece, 2013.

[12] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[13] A. Mohanta and V. K. Mittal, "Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features," *Computer Speech & Language*, vol. 72, p. 101287, 2022.

[14] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.

[15] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *Proc. of ICASSP*. IEEE, 2024, pp. 291–295.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. of ICASSP*. IEEE, 2017, pp. 776–780.

[17] S. P. Patel, K. Nayar, G. E. Martin, K. Franich, S. Crawford, J. J. Diehl, and M. Losh, "An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives," *Journal of Autism and Developmental Disorders*, vol. 50, pp. 3032–3045, 2020.

[18] S. Wehrle, F. Cangemi, K. Vogeley, and M. Grice, "New evidence for melodic speech in autism spectrum disorder," in *Proc. of Speech Prosody*, vol. 2022, 2022, pp. 37–41.

[19] E. Lyakso, O. Frolova, and A. Grigorev, "A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders," in *Proc. of SPECOM*, 2016, pp. 43–50.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of ICASSP*, 2013.

[21] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *Proc. of Interspeech*, 2013, pp. 191–194.

[22] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. of Interspeech*, 2017, pp. 849–853.

[23] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proc. of Interspeech*, 2017, pp. 3442–3446.

[24] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowd-sourced semistructured speech recordings: machine learning model comparison study," *JMIR pediatrics and parenting*, vol. 5, no. 2, p. e35406, 2022.

[25] S. M. Zala, D. Reitschmidt, A. Noll, P. Balazs, and D. J. Penn, "Sex-dependent modulation of ultrasonic vocalizations in house mice (mus musculus musculus)," *PloS one*, vol. 12, no. 12, p. e0188647, 2017.

[26] K. Yao, M. Bergamasco, M. L. Scattoni, and A. P. Vogel, "A review of ultrasonic vocalizations in mice and how they relate to human speech," *The Journal of the Acoustical Society of America*, vol. 154, no. 2, pp. 650–660, 2023.

[27] A. H. Fonseca, G. M. Santana, G. M. Bosque Ortiz, S. Bampi, and M. O. Dietrich, "Analysis of ultrasonic vocalizations from mice using computer vision and machine learning," *Elife*, vol. 10, p. e59161, 2021.

[28] M. Premoli, D. Baggi, M. Bianchetti, A. Gnutti, M. Bondaschi, A. Mastinu, P. Migliorati, A. Signoroni, R. Leonardi, M. Memo *et al.*, "Automatic classification of mice vocalizations using machine learning techniques and convolutional neural networks," *PloS one*, vol. 16, no. 1, p. e0244636, 2021.

[29] K. Qian, T. Koike, K. Tamada, T. Takumi, B. W. Schuller, and Y. Yamamoto, "Sensing the sounds of silence: A pilot study on the detection of model mice of autism spectrum disorder from ultrasonic vocalisations," in *Proc. of EMBC*. IEEE, 2021, pp. 68–71.

[30] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "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.

[31] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[32] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[33] D. P. Kingma, J. A. Ba, and J. Adam, "A method for stochastic optimization. arxiv 2014," *arXiv preprint arXiv:1412.6980*, vol. 106, p. 6, 2020.

[34] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," in *Proc. of ICASSP*. IEEE, 2020, pp. 6879–6883.