Misophonia Sound Recognition Using Vision Transformer

B. Bahmei, E. Birmingham, and S. Arzanpour

Abstract— **Misophonia is a condition characterized by an abnormal emotional response to specific sounds, such as eating, breathing, and clock ticking noises. Sound classification for misophonia is an important area of research since it can benefit in the development of interventions and therapies for individuals affected by the condition. In the area of sound classification, deep learning algorithms such as Convolutional Neural Networks (CNNs) have achieved a high accuracy performance and proved their ability in feature extraction and modeling. Recently, transformer models have surpassed CNNs as the dominant technology in the field of audio classification. In this paper, a transformer-based deep learning algorithm is proposed to automatically identify trigger sounds and the characterization of these sounds using acoustic features. The experimental results demonstrate that the proposed algorithm can classify trigger sounds with high accuracy and specificity. These findings provide a foundation for future research on the development of interventions and therapies for misophonia.**

Keywords—Misophonia, Sound Classification, Transformer Models, Deep Learning

I. INTRODUCTION

Misophonia is a relatively new and understudied condition characterized by an abnormal emotional response to specific sounds that are often repetitive and human-produced (e.g., chewing, snoring, tapping, sniffing, etc.), eliciting excessive and inappropriate negative reactions even at low amplitudes [1], [2]. People who suffer from misophonia experience increased sympathetic nervous system arousal, accompanied by emotional distress in response to these sounds, which are known as trigger sounds. The condition can have a significant impact on an individual's daily life, leading to difficulties in social and professional settings [2].

Most of the studies on misophonia have been case studies aiming to uncover the nature of the trigger sounds, and physical responses to those sounds [3]–[5]. However, there are few publications and studies to evaluate treatments for misophonia [6]–[8]. In the absence of these studies, it is very challenging for families and clinicians to provide appropriate care to individuals who are suffering from misophonia. Identifying and categorizing sounds associated with misophonia can assist in creating treatments and therapies for people who have the condition [9]. Upon detecting the trigger sounds, further considerations can be evaluated to ameliorate the condition such as notifying of sound's presence, filtering out the sound, masking the sound, and so on.

In recent years, there has been an increasing interest in using machine learning algorithms to automatically classify environmental sounds such as Support Vector Machine (SVM) [10], and hidden Markov models (HMM) [11]. Recently, deep learning techniques have been introduced to enhance the recognition performance of environmental sounds [12], [13]. Deep neural networks can automatically learn and extract features from the raw data, which reduces the need for manual feature engineering, and allows the model to capture more complex patterns and dependencies in the data [12]. Deep learning algorithms such as Convolutional Neural Networks (CNNs) [13], [14], Recurrent Neural Networks (RNNs) [15], [16], and their combination [17] have been shown to be highly effective in accurately identifying environmental sounds, with high sensitivity and specificity. Although CNNs have been used to classify sounds, they are less successful at processing long sequences of audio data.

Lately, there has been a growing trend of utilizing attention mechanisms to concentrate on the essential aspects of the sound being analyzed are designed to handle long sequences of data. Attention-based models [18], particularly those using Transformers, have been gaining popularity in recent years. Transformers are a type of neural network that relies solely on attention mechanisms. This makes them well-suited for parallel computations and the incorporation of global context, leading to more accurate results. As a result, they have become a popular choice in various fields such as Natural Language Processing (NLP), computer vision, and more recently, areas related to sound.

In the field of audio classification, there have been a number of transformer models proposed [19]. Some of the proposed models have investigated the benefits of using Bidirectional Encoder Representations from Transformers (BERT) models [20]. The BERT models take a given token and the position embeddings as input, to address the problem of sound classification at the edge. Similarly, an Audio Spectrogram Transformer (AST) is proposed in [21] that is completely based on attention-based models. In another study, the use of AST is explored with a Many-to-Many Audio Spectrogram Transformer (M2M-AST), which can output sequences with different resolutions for multi-channel audio inputs [22]. To simplify the training process, the drop token technique is introduced in combination with a Video–Audio–Text Transformer (VATT) model, which achieved competitive results [23]. All these studies achieved significant performance in the task of audio classification.

This paper proposed a sound classifier for misophonia to recognize trigger sounds in the environment. The purpose of this study is to propose a trigger sound recognition system for the sufferer of misophonia which can provide a foundation for future research on the development of interventions and

B. Bahmei (bbahmei@sfu.ca), School of Mechatronics System Engineering, Simon Fraser University, Surrey, BC, Canada

E. Birmingham (elina_birmingham@sfu.ca), Faculty of Education, Simon Fraser University, Burnaby, BC, Canada

S. Arzanpour (arzanpour@sfu.ca), School of Mechatronics System Engineering, Simon Fraser University, Surrey, BC, Canada

therapies. In this paper, a standard transformer model which initially is applied to images, named Vision Transformer (ViT) is considered and modified for the misophonia sound classification [24]. The model is trained and evaluated on selected sounds from the ESC-50 dataset [25] which are commonly reported as trigger sounds in misophonia. The experimental results indicate that the proposed model has the capability to accurately identify the target trigger sounds in the environment. To summarize, the main contributions of this paper are as follows:

1) A ViT model is considered and modified to achieve a very high-level of classification accuracy in sound recognition.

2) The classification model is applied on selected trigger sounds for misophonia to introduce the first misophonia sound recognition system.

The composition of the paper is as the following: In Section II, the methods including the ViT, and dataset are discussed. Section III provides details about the experimental results. Finally, the conclusions are presented in Section IV.

II. METHODS

In this section, the methods for ViT, and the dataset are explained.

A. *Vision Transformer (ViT)*

The transformer model, first introduced in 2017, uses an attention mechanism to generate representations of its inputs and outputs [18]. The transformer model has two main components, an encoder, and a decoder. The encoder converts an input sequence of symbol representations into a sequence of continuous representations, and the decoder generates an output sequence of symbols one at a time. Additionally, the model is autoregressive, meaning that it uses previously generated symbols as input when generating the next one at each step.

The architecture used here is based on the ViT model proposed in [24]. This model breaks down an image into fixed-size patches, accurately embeds each one, and incorporates positional embedding as input to the transformer encoder. The transformer encoder embeds information globally across the entire image, and during training, the model learns to encode the relative location of the image patches to rebuild the image's structure. Furthermore, a classification token is added to learn the information extracted by the transformer encoder for the classification task.

A schematic view of the designed model in this paper is depicted in Figure 1. The input to the model is a 2D spectrogram. A spectrogram is a logarithmic frequency scale and is considered one of the most common and effective features for audio recognition [13]. In order to extract patches that are required for a transformer, the raw audio signal will be framed into 18 frames with a length of 32 milliseconds.

The frames have a 50% overlap to avoid missing information at the edges. Afterward, the spectrograms of each frame are extracted, stacked together as the input patches, and fed to the transformer encoder as the input. The output of the transformer encoder known as encoded patches is flattened and fed to the fully connected layers for the classification task.

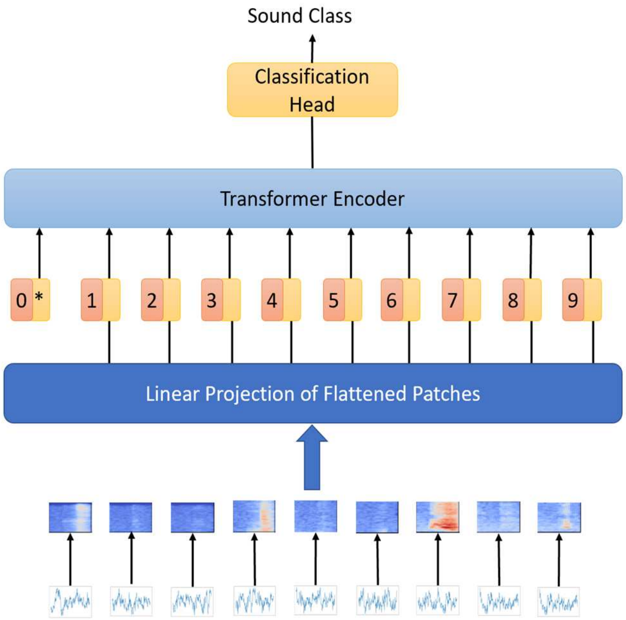

Figure 1 Transformer model schematic view

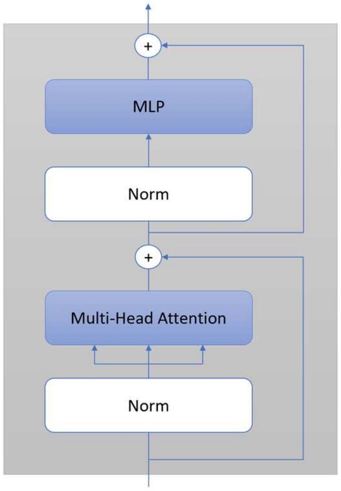

Figure 2 Transformer encoder block

The output of the model is the one-hot representation of sound classes.

The transformer encoder block is presented in Figure 2. The transformer encoder includes:

- The Multi-Head Self Attention (MSA) Layer, or also known as Multi-Head Attention (MHA) is a key component of the transformer encoder. It allows the model to attend to different positions of the input sequence simultaneously, by performing multiple self-attention operations with different weight matrices, also known as heads. These attention heads help to train both local and global dependencies in the input. This allows the model to learn and capture more complex patterns in the input data and improve its performance.
- The Multi-Layer Perceptron (MLP) Layer, also known as the Position-wise Fully Connected Feed-

Forward Network, is another key component of the transformer encoder. It is a simple feed-forward neural network that is applied to each position of the input sequence independently and in parallel. The role of this layer is to learn and capture more complex patterns in the input data that the self-attention layer might have missed.

- Layer Normalization (LN), also known as Layer Norm, is a technique used to normalize the activations of the neurons in a neural network layer. It is typically applied before each block, such as the MHA and MLP layers, as it does not introduce any new dependencies between the training images. This helps to improve the training time and overall performance of the model.

The classification head in Figure 1 is implemented using MLP with four fully connected layers. In the classification head, the ReLU activation function is used after each layer except the last one. For the last layer, the SoftMax activation function is applied. There is batch normalization after each layer. The ADAM optimizer [26] is used in order to update network weights.

B. Dataset

The ESC-50 dataset is a popular dataset for sound classification which includes 50 classes consisting of animal, human, natural, and urban sounds. Seven sounds are selected from this dataset as the commonly reported trigger sounds for misophonia including breathing, snoring, drinking, keyboard typing, clock ticking, mouse-clicking, and coughing. For each class, there are 40 audio recordings, each lasting 4 seconds in duration. In this paper, as chewing sounds are commonly reported in the literature as a trigger sound for misophonia, 40 samples of chewing sounds were collected from freesound.org and added to the dataset. In total, the dataset includes eight trigger sounds as the target output sounds.

III. RESULTS AND DISCUSSION

In this section, the simulation results of the proposed technique are presented. Considering that there is no specific publication about misophonia sound recognition, it is not possible to directly compare our work with others in this specific area. However, several experiments are conducted to evaluate the performance and accuracy of the proposed method.

The model is trained over 100 epochs using a batch size of 32. Since there are few misophonia sounds in the dataset, applying k-fold cross-validation would provide even fewer training and validation sets, which might result in overfitting and reduced generalization performance. Therefore, the dataset is split into 80% for training, 15% for validation, and 5% for testing. The data samples are shuffled before feeding to the model. Figure 3 shows the overall accuracy and loss of the training and validation set of the model on each training iteration.

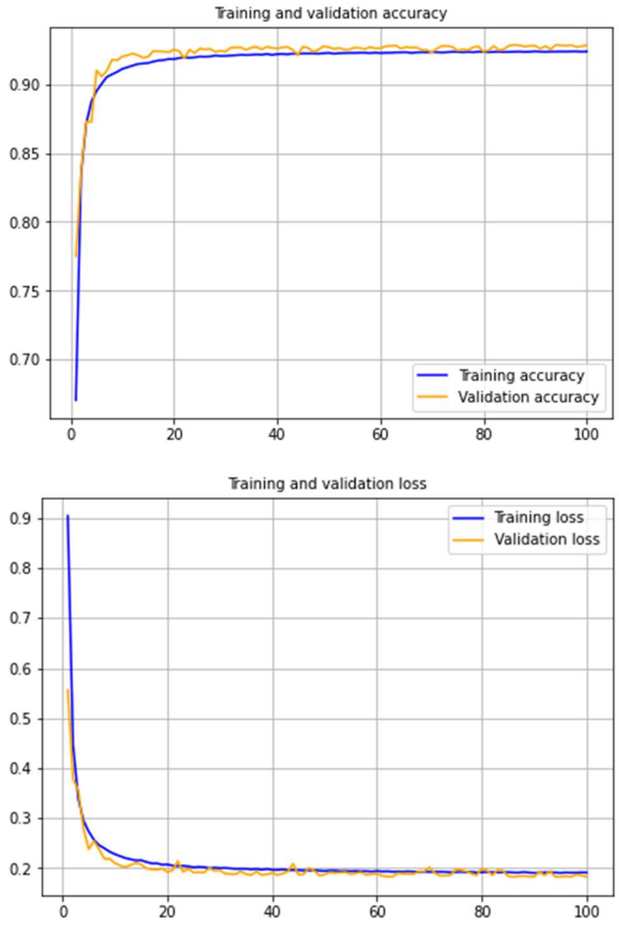

Figure 3 . Overall accuracy and loss

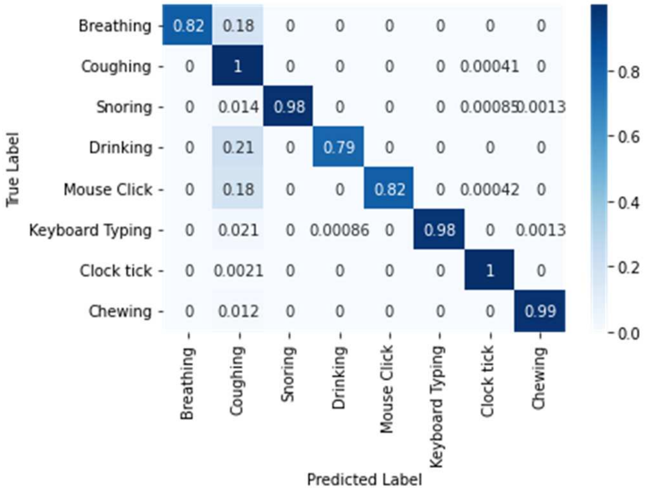

Figure 4 Confusion Matrix. The diagonal elements represent the percentage of instances for which the predicted label is equal to the true label (TP and TN), while off-diagonal elements are those that are mislabeled by the classifier (FP an FN).

For this study, the categorical cross-entropy loss is considered. This figure shows that the validation accuracy and loss of the model reach 92.29% and 0.1956, respectively which shows the performance of the model to recognize trigger sounds. It also indicates that the learning process is quite consistent and there is no bias or variance during the training.

In addition, a confusion matrix is also used to evaluate the performance of the classifier presented in Figure 4. It shows the percentage of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) predictions made by a model. It can be seen from Figure 4 that the most difficult classes to classify were drinking, breathing, and mouse-clicking. They have been misclassified as coughing in some samples. However, it is noteworthy that the coughing, chewing and clock ticking sounds are almost not misclassified. From the confusion matrix, some performance metrics including precision, recall and F1 score are computed and shown in Table 1 to evaluate the classification results.

Table 1 Performance metrics for the model, showing precision, recall, and F1 score for each class

	Precision	**Recall**	**F1 score**
Breathing	1	0.97	0.98
Coughing	0.92	1	0.95
Snoring	1	0.98	0.99
Drinking	1	0.96	0.98
Mouse Click	1	0.97	0.98
Keyboard Typing	1	0.98	0.99
Clock tick	1	1	1

IV. CONCLUSION

In this paper, a vision transformer-based deep learning model is evaluated for misophonia sound classification. The accuracy of the model indicates that the system can recognize the trigger sounds accurately in the environment. It is a foundation and starting point for designing intervention techniques and therapies for the sufferer of misophonia. This is the first study conducted for detecting trigger sounds for misophonia using artificial intelligence techniques. Further investigation can be conducted to use these results.

ETHICS STATEMENT

This paper does not include any experimental procedures involving human subjects or animals.

ACKNOWLEDGMENT

This project is supported by funding from the Kids Brain Health Network (KBHN).

REFERENCES

[1] N. E. Scheerer, T. Q. Boucher, B. Bahmei, G. Iarocci, S. Arzanpour, and E. Birmingham, "Family Experiences of Decreased Sound Tolerance in ASD," *J. Autism Dev. Disord.*, 2021, doi: 10.1007/S10803-021-05282-4.

[2] J. J. Brout *et al.*, "Investigating Misophonia: A Review of the Empirical Literature, Clinical Implications, and a Research Agenda," *Front. Neurosci.*, vol. 0, no. FEB, p. 36, Feb. 2018, doi: 10.3389/FNINS.2018.00036.

[3] M. Edelstein, D. Brang, R. Rouw, and V. S. Ramachandran, "Misophonia: Physiological investigations and case descriptions," *Front. Hum. Neurosci.*, vol. 7, no. JUN, p. 296, Jun. 2013, doi: 10.3389/FNHUM.2013.00296/BIBTEX.

[4] H. Tinnitus, M. G. Editors, D. F. Duddy, . D Au, and L. A. Flowers, "Treatments for Decreased Sound Tolerance (Hyperacusis and Misophonia)," *Au.D. Semin Hear*, vol. 35, pp. 105–120, 2014, doi: 10.1055/s-0034-1372527.

[5] E. Boucher, T. Q., Scheerer, N. E., Iarocci, G., Bahmei, B., Arzanpour, S., & Birmingham, "Misophonia, hyperacusis, and the relationship with quality of life in autistic and non-autistic adults," 2021.

[6] R. L. Schneider and J. J. Arch, "Case study: A novel application of mindfulness- and acceptance-based components to treat misophonia," *J. Context. Behav. Sci.*, vol. 6, no. 2, pp. 221–225, Apr. 2017, doi: 10.1016/J.JCBS.2017.04.003.

[7] A. E. Schröder, N. C. Vulink, A. J. van Loon, and D. A. Denys, "Cognitive behavioral therapy is effective in misophonia: An open trial," J. Affect. Disord., vol. 217, pp. 289–294, Aug. 2017, doi: 10.1016/J.JAD.2017.04.017.

[8] A. Schröder, N. Vulink, and D. Denys, "Misophonia: Diagnostic Criteria for a New Psychiatric Disorder," PLoS One, vol. 8, no. 1, Jan. 2013, doi: 10.1371/JOURNAL.PONE.0054706.

[9] B. Birmingham, E., Arzanpour, S., Bahmei, "System and Method for Ambient Noise Detection, Identification and Management," WO/2021/119806, 2021.

[10] S. Sameh and Z. Lachiri, "Multiclass support vector machines for environmental sounds classification in visual domain based on log-Gabor filters," undefined, vol. 16, no. 2, pp. 203–213, Jun. 2013, doi: 10.1007/S10772-012-9174-0.

[11] Y. T. Peng, C. Y. Lin, M. T. Sun, and K. C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," Proc. - 2009 IEEE Int. Conf. Multimed. Expo, ICME 2009, pp. 1218–1221, 2009, doi: 10.1109/ICME.2009.5202720.

[12] K. J. Piczak, "Environmental sound classification with convolutional neural networks," IEEE Int. Work. Mach. Learn. Signal Process. MLSP, vol. 2015-November, Nov. 2015, doi: 10.1109/MLSP.2015.7324337.

[13] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: 10.1109/LSP.2017.2657381.

[14] S. Adapa, "Urban Sound Tagging using Convolutional Neural Networks," pp. 5–9, Sep. 2019, doi: 10.33682/8axe-9243.

[15] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," Adv. Neural Inf. Process. Syst., pp. 892–900, Oct. 2016, Accessed: Jan. 11, 2022. [Online]. Available: https://arxiv.org/abs/1610.09001v1.

[16] T. H. Vu and J.-C. Wang, "Acoustic Scene and Event Recognition Using Recurrent Neural Networks," 2016.

[17] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network For Environmental Sound Classification," IEEE Signal Process. Lett., 2022, doi: 10.1109/LSP.2022.3150258.

[18] A. Vaswani et al., "Attention is All you Need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[19] P. Remagnino et al., "Transformers for Urban Sound Classification—A Comprehensive Performance Evaluation," mdpi.com, 2022, doi: 10.3390/s22228874.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.

[21] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 1, pp. 56–60, Apr. 2021, doi: 10.48550/arxiv.2104.01778.

[22] S. Park, Y. Jeong, T. L.- DCASE, and undefined 2021, "Many-to-Many Audio Spectrogram Tansformer: Transformer for Sound Event Localization and Detection.," dcase.community, Accessed: Jan. 20, 2023. doi: https://dcase.community/documents/workshop2021/proceedings/DCASE2021Workshop_Park_39.pdf

[23] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2022-September, pp. 2753–2757, 2021, doi: 10.21437/INTERSPEECH.2022-227.

[24] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, doi: 10.48550/arxiv.2010.11929.

[25] K. J. Piczak, "ESC: Dataset for environmental sound classification," in MM 2015 - Proceedings of the 2015 ACM Multimedia Conference, Oct. 2015, pp. 1015–1018, doi: 10.1145/2733373.2806390.

[26] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., Dec. 2014, doi: https://arxiv.org/abs/1412.6980v9.