



Please cite the Published Version

Ridley, Harrison, Cunningham, Stuart , Darby, John, Henry, John  and Stocker, Richard (2024) The Affective Audio Dataset (AAD) for non-musical, non-vocalized, audio emotion research. IEEE Transactions on Affective Computing. pp. 1-12.

DOI: <https://doi.org/10.1109/TAFFC.2024.3437153>

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/635285/>

Usage rights:  In Copyright

Additional Information: © 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

The Affective Audio Dataset (AAD) for non-musical, non-vocalized, audio emotion research

Harrison Ridley, Stuart Cunningham, John Darby, John Henry, and Richard Stocker

Abstract—The Affective Audio Dataset (AAD) is a new and novel dataset of non-musical, non-anthropomorphic sounds intended for use in affective research. Sounds are annotated for their affective qualities by sets of human participants. The dataset was created in response to a lack of suitable datasets within the domain of audio emotion recognition.

A total of 780 sounds are selected from the BBC Sounds Library. Participants are recruited online and asked to rate a subset of sounds based on how they make them feel. Each sound is rated for arousal and valence. It was found that while evenly distributed, there was bias towards the low-valence, high-arousal quadrant, and displayed a greater range of ratings in comparison to others.

The AAD is compared with existing datasets to check its consistency and validity, with differences in data collection methods and intended use-cases highlighted. Using a subset of the data, the online ratings were validated against an in-person data collection experiment with findings strongly correlating. The AAD is used to train a basic affect-prediction model and results are discussed.

Uses of this dataset include, human-emotion research, cultural studies, other affect-based research, and industry use such as audio post-production, gaming, and user-interface design.

Index Terms—Affect, arousal, audio emotion recognition, dataset, sound, valence.

I. INTRODUCTION

SOUND design for film often makes use of audio for narrative or functional purposes, but also to elicit an affective response in a film’s audience. Following research that discussed the use of audio in this way with film sound professionals [1], the authors investigated existing sound datasets with affective labelling.

Whilst there are a small number of affective audio datasets in the field, almost all include musical and/or human vocalizations, which may skew data or not be of use when editing sound for movie and TV scenes. For example, IADS-2 [2] and its more recent derivative IADS-E [3]. Both datasets include sounds of musical instruments and human vocalization, such as moaning or burping. In audio post-production music has typically been considered and chosen for scenes in advance, and will be the key emotional driver. Human-vocalized sounds are not typically used in any manner other than to drive the story through dialogue [1], [4].

H. Ridley, S. Cunningham and R. Stocker are with the School of Computer and Engineering Sciences, University of Chester, UK (e-mail: 1723786@chester.ac.uk; s.cunningham@chester.ac.uk; r.stocker@chester.ac.uk).

J. Darby and J. Henry are with the Department of Computing and Mathematics, Manchester Metropolitan University, UK (e-mail: j.darby@mmu.ac.uk; john.henry@mmu.ac.uk).

IADS-2 contains 167 sounds, relatively few compared to the International Affective Picture System (IAPS), which contains 1,182 pictures [5] and less still when compared to MER datasets, such as AMG1608 [6] or DEAM [7] with 1608 and 1802 sounds, respectively. IADS-E [3] intended to increase the distribution of sounds in the bi-dimensional affective space of arousal and valence, since IADS-2 is skewed towards high-arousal, low-valence sounds. IADS-E did achieve a wider spread of data, however, it was found that results were still skewed towards high-arousal, low-valence, much in the same way as IADS-2. Further, IADS-E made significant use of sounds of a musical nature ($N = 170$) and of human origin ($N = 74$), equating to 26% of its total samples.

Accounting for these limitations, and compounded by the relative scarcity of datasets for AER in general, it was decided that creating a novel dataset of affect-labelled audio files that purposefully exclude sounds identified as being musical or human vocalizations would be beneficial to the affective audio research community.

This article describes a sound affect-rating data collection, similar to those conducted by the aforementioned datasets, with particular care taken to omit musical and human sounds. These are omitted to control for potential bias and to make the dataset more usable in the sound-for-moving image industries and related research fields. A novel dataset of 780 affect-labelled, non-musical, non-anthropomorphic sounds are presented as the culmination of this data-collection and its potential uses are explored in the sections that follow.

II. BACKGROUND AND RELATED WORK

A. Music and Audio Emotion Recognition

Music Emotion Recognition (MER) is a well-established research discipline, investigating the relationship between musical stimuli and human affective responses, predicting responses to stimuli for use in a multitude of fields that may include music recommendation, retrieval, generative composition, psychotherapy and more [8], [9]. MER makes extensive use of Machine Learning (ML) techniques to predict emotional responses [10]–[14]. To support this there are many datasets available for training ML algorithms [15]–[19] and these are exclusively music-based.

Audio Emotion Recognition (AER) is similar to MER but differs in focus, by studying non-musical sounds. AER is less established and is an emerging field of research. Where MER may be useful in generative composition, music recommendation, etc., AER may be useful in User Interface (UI) design,

audio post-production for film, TV, radio, advertising, computer gaming, speech recognition [20], and has overlapping uses with MER, such as psychotherapy.

Whilst many researchers are making use of ML techniques within AER, there are limited datasets with large numbers of either sounds or human annotator participants available for use. One notable AER dataset is the Emotional Sound Database [21], consisting of 390 sound clips. The Emotional Sound Database made use of musical instruments and human sounds, although it was annotated by only four participants. The researchers made use of a regressor to predict the arousal ($r = 0.61$) and valence ($r = 0.49$) values of sounds with varying performance across the underlying classes of sound in the database.

Drossos *et al.* [22], [23] used arousal and valence data from the IADS-2 dataset to train ML algorithms to classify sounds into quadrants in the arousal-valence space. They then went further to extract typical audio features to use in training and validation of both support vector machines and Artificial Neural Networks (ANN) to predict a more accurate placement in the quadrant. Cunningham *et al.* [24] also made use of IADS-E to successfully train regressors and ANNs to predict affective response for arousal ($R^2 = 0.644$) and valence ($R^2 = 0.654$). Further, Choi *et al.* [25] were able to successfully classify sounds from IADS-2 into the emotional factors of ‘happiness’, ‘sadness’ and ‘negativity’ with an 88.9% accuracy rate.

B. Models of Emotion

There are two primary models of emotion utilized in the field, *continuous* and *discrete* models [26].

Discrete models are those such as Ekman’s [27]–[29] and Roseman’s [30] that describe defined, labelled emotions and are often based on facial expressions. They are useful in organizing the emotional tendencies of people and simplifying data [31].

Continuous models of emotion, such as Russell’s Circumplex Model of Affect [32], which measures arousal and valence, and Thayer’s Model, which measures stress and energy [33], do not label emotions but give numerical values for each dimension. Such models propose that all affective states can be determined using these neurophysical attributes [26]. The continuous approach allows detachment from labelled emotions and study of individual dimensions as and when necessary, such as in research conducted by Drossos *et al.* [23], where the authors examined only the relationship between arousal and rhythmic qualities of sounds.

III. METHOD

A. Sound Stimuli

The sounds were manually selected by the principal author from the 33,066 sounds in the BBC Sounds Library (<https://sound-effects.bbcrewind.co.uk>) by applying selection criteria that sounds must be: non-musical; non-human; and that they could reasonably be utilized in a film or TV production. Sounds needed to be short enough to not evolve in their complexity or content (for example, simpler sounds such as

TABLE I
PERCENTAGE OF AAD SOUNDS CATEGORIZED BY GROUP

Category	Contribution	Examples
Daily Life	23%	Washing Dishes, Footsteps
Industry/Machinery	19%	Printing Press, Circular Saw
Animals	13%	Sheep Baa-ing, Dog Bark
Transport/Aircraft	12%	Tyre Screech, Plane Flyby
Atmosphere/Weather	11%	Wind, Thunderclap
Electronics	9%	Phone Ringing, Mains Hum
Bells/Alarms/Clocks	4%	Church Bell, Fire Alarm
Military/Destruction	4%	Gunshots, Rubble Falling
Sports and Toys	3%	Tennis Hits, Swimming
Fire	1%	Crackling Fire, Gas Burner
Other	1%	Heartbeat, Comedy Boing

footsteps or a car engine running should be used). They should all be of similar loudness in relation to one another (so that peak volume was not a factor in participants ratings) and be easily distinguishable when played through numerous audio reproduction devices. Selected sounds were categorized by the principal researcher to enable users of the dataset to understand the content of sounds used. Table I shows the distribution of sounds across these groups. The categories are loosely based on those already utilized by the BBC Sounds Library, with some merged to condense the table.

Where necessary sounds were truncated in length to be as close to six seconds as reasonably possible (some, such as thunderclaps, were left longer to enable capturing of the whole instance). The longest sound was nine seconds and the shortest was three seconds, with the average duration 6.27 seconds. The target duration was set at six seconds to enable comparison with other datasets such as IADS-2, IADS-E [2], [3], and Emo-Soundscapes [34], in which similar averages were realized. Other comparable research, such as Lopes, Liapis and Yannakakis’ ‘Modelling Affect for Horror Soundscapes’ [35], and the Emotional Sound Database [18] contain sounds with a duration range of five to ten seconds (Horror Soundscapes), and average of 3.5 seconds (Emotional Sound Database) and so the AAD is in-line with ad-hoc standards in the field. Other datasets, such as AudioSet [36] and AMG1608 [6], contain sounds of a longer average duration (10 and 30 seconds, respectively). However, they contain little non-musical, non-anthropomorphic sounds, so a target duration of six seconds was decided for the AAD.

Due to a mixture of stereo and mono files in the source data, all stereo files were summed to mono to enable playback on most devices and to eliminate stereo spread/movement as a factor in participant ratings. All files were then LUFS (Loudness Units Full Scale)-normalized using Adobe Audition to -23 LUFS (as per ITU-R BS.1770-4 recommendation for measuring programme loudness and true-peak audio levels [37]) and exported to 44.1 kHz, 16-bit uncompressed WAV files. No further manipulation of the audio files was undertaken. Figure 1 summarizes the steps taken in preparing the audio for rating, illustrating the selection, processing, and output steps.

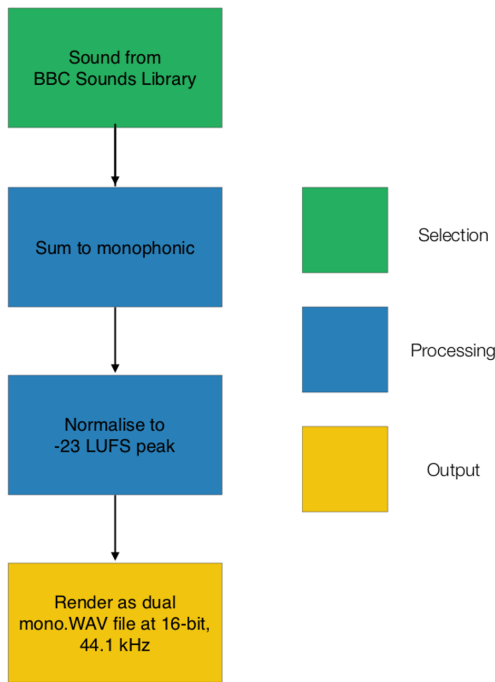


Fig. 1. Flowchart depicting audio standardisation workflow.

B. Participants

Participants were recruited using the Amazon Mechanical Turk (AMT) online ‘crowdworking’ platform [38]. AMT allows the researchers to set parameters for the ‘workers’ to satisfy. AMT has been used to collect data for affective audio research purposes [6], [39] and has been determined to be a suitable alternative to traditional methods [40]. Workers were selected for participation if they met the inclusion criteria of: residing in a country considered to have a ‘Western’ culture; were aged 18 or over; and could make use of headphones or external speakers. As the research aims to aid Western sound research, workers residing in countries typically not considered to be of Western culture were excluded from the data collection.

Participants were required to have an AMT ‘master’ qualification, meaning that they had consistently demonstrated a high degree of success in their tasks across a variety of work in AMT in the past. This characteristic was chosen to reduce the risk of abuse of the AMT task.

All participants were given a small monetary reward through AMT for satisfactorily completing their batch of ratings.

Following approval from the researchers’ institutional ethics committee, a total of 867 participants took part in the data collection and submitted ratings that were used in the formation of the dataset. To ascertain the diversity of participants, demographic data was collected. Gender was relatively evenly split, with 49% identifying as female, 50% identifying as male and the remaining participants identifying as non-binary or preferred not to give their gender identity. 97.92% of participants resided in the USA, 0.57% in the UK, 0.35% in Italy, 0.46% in Brazil, 0.46% in Canada, and 0.23% in Mexico. Participants had an average age of 43.2 years with a range of

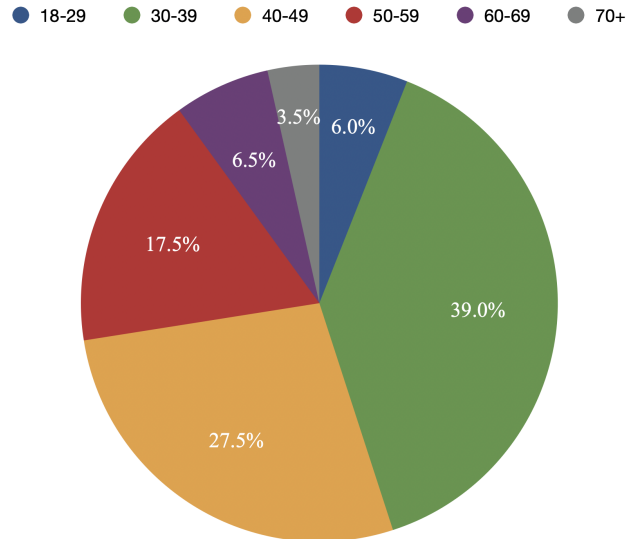


Fig. 2. Pie-chart showing age distribution of participants.

18-74. The complete age range distribution is shown in the pie chart (Figure 2).

Participants were given an outline of the research, its intended purpose and how the data they provided would be used. They were made aware of the anonymization process, their rights as a participant and how the data would be stored in accordance with the laws in force at the time of its collection. Participants were told how to raise a complaint with the researchers and/or the University or withdraw from the research at any point during the data collection, or in the future, should they wish to do so.

Participants took part in the data collection remotely using their own equipment, which it was recognized may vary, especially in terms of audio reproduction and playback. As such, participants were asked to wear headphones whenever possible to reduce any effect that different acoustic environments and related background noise may have on their rating. Overall, 19.77% of participants did not use headphones. The researchers acknowledge that this may have some effect on their ability to hear, and thus rate, the sounds effectively [41].

C. Measuring Emotion

The complete set of sounds would take a long time for any one participant to rate and be at high risk of inducing fatigue. As such, the sounds were divided into batches of 60, each to be rated by 40 participants. It was originally planned that 900 sounds would be rated in this manner to match the number of sounds more closely with that of IADS-E. However, time constraints meant that the final two batches of data collection did not proceed, and the researchers continued with the 780 sounds annotated up to this point.

A dedicated web site for the data collection was designed by the research team, making use of the Affective Slider [42] as the tool for users to give a response for arousal and valence by using a simple, two slider interface. The Affective Slider allows values between 0 and 1, with granularity of two decimal places (effectively 101 possible ratings for each

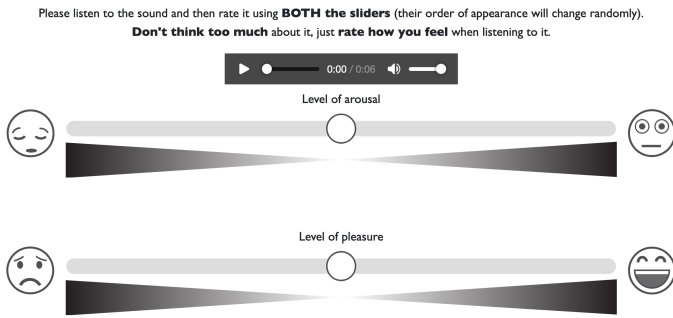


Fig. 3. Example of the Affective Slider interface, as used in data collection.

of arousal and valence). This was chosen as an alternative to the Self-Assessment Manikin (SAM) [43] used in other research, such as IADS-2 and IADS-E. The Affective Slider allows for more granularity and has been shown to be quick and easy to use, without the need for detailed instructions to participants [42]. Further, the Affective Slider has been found to give comparable results to SAM while presenting a more modern interface that compensates for the general contemporary desensitization toward highly arousing content [42]. Participants were asked to rate how each of the 60 sounds made them feel at that time. They were asked to do so by selecting a point on each of the Affective Slider scales for arousal and valence for each sound, as in Figure 3.

D. Procedure

Participants were instructed to complete the whole task (rating 60 sounds) in one sitting. They were introduced to the project, given information to enable them to contact the researchers or University should they wish to for any reason, shown instructions for the task, and then asked to complete a practice rating task to familiarize themselves with the interface and to allow them to set their volume to a comfortable level.

Once the participants understood the task and use of the Affective Slider, they were asked to listen to each sound in its entirety (average 6.27 seconds) and rate how it made them feel (induced emotion) using the sliders for arousal and valence. Participants were free to listen to each sound as many times as they wished. When the participants were satisfied with their answer, they could continue to the next sound at their leisure. In total there were 870 participants, with each individual allowed to complete no more than two batches of sounds over the duration of the data collection period.

The sounds in each batch were presented in a dynamically randomized sequence to prevent the order of sounds having any effect on the ratings [44]. The mean time taken to complete a batch over all participants was 16 minutes.

E. Data Validation

1) *In-Person Validation*: To verify the viability of the data collected utilizing the AMT platform, a single batch was repeated with in-person participants. The research team invited students and staff at their institution to take part in a controlled environment (a usability lab). The method replicated that of the AMT version with minor technical differences in execution.

TABLE II
COMPARISON OF IN-PERSON AND AMT COLLECTED DATA

	In-Person		AMT	
	Arousal	Valence	Arousal	Valence
Mean	0.57	0.45	0.54	0.42
SD	0.24	0.23	0.22	0.21

Data collection took place using a local emulator to host the website to rate sounds instead of hosting the website online. Participants all used the same audio equipment to play back sounds: a MacBook Pro paired with a Universal Audio Volt 4 audio interface and a pair of Beyerdynamic DT 770 Pro headphones (80 Ohm version). 27 participants were asked the same preliminary questions and filtered in the same way as the AMT workers, except for having to hold an AMT master qualification. Participants were given equivalent vouchers to use at University food outlets instead of a monetary reward.

The data was processed in the same manner as the online data collection with the ratings provided from each participant subsequently being averaged in the arousal and valence response for each sound to provide an overall summary.

The results of the in-person set were compared to the same batch in the online data collection and found to be strongly correlated. Pearson's correlation coefficient was calculated, showing a strong positive correlation of $r = 0.740$ for arousal and $r = 0.882$ for valence, with strong statistical significance ($p < 0.00001$) in both dimensions. Mean values were compared for the in-person and AMT batches, shown in Table II. The table shows a small positive shift in overall ratings of both arousal and valence when conducted in person, albeit with a slightly wider spread. Generally, it can be concluded that in-person rating and online rating gave very similar results.

2) *Emotion Classification from Sound Descriptors*: The BBC Sound Library gives text descriptors for each sound and it was considered this may be a point of validation. One batch was selected to test the descriptors affective content against the rated arousal and valence. Each of the sounds' descriptors was classified to determine the most likely of Ekman's Basic Emotions [45] using a fine-tuned version of the DistilRoBERTa-base for Emotion Classification model [46].

Figure 4 shows the sounds from this validation set in arousal-valence space, with colors representing each of Ekman's Basic Emotions that they were classified as. The locations of the basic emotions [47] on the arousal-valence plane are super-imposed on the chart to help identify how well the text-classified sounds match their AAD.

The audio ratings were compared against the quadrants on the arousal-valence space that the text classifier predicted they should sit in. Results showed that the text classifier quadrant agreed with the audio rating for 30% of the sounds, with 'disgust' being the most accurate of the classifications at 50% agreed and 'sadness' being least accurate with no agreements (although only two sounds were categorized as 'sadness' in this batch). 'Joy', which had the most text-classified sounds in the batch (23 sounds), only agreed in the case of four sounds. This short validation exercise suggests that the text data may not be very useful in assisting the emotion prediction.

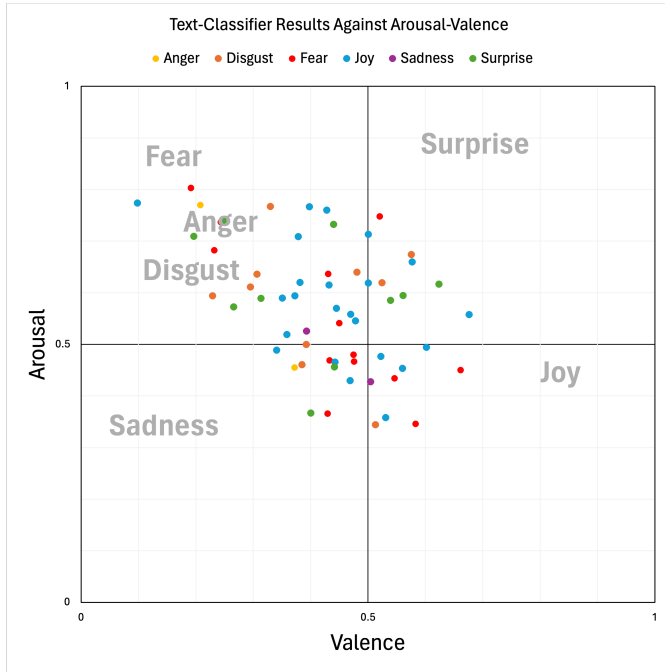


Fig. 4. Scatter chart of average arousal and valence of sounds in batch 1, and their text-predicted basic emotions.

3) Arousal and Valence Prediction from Sound Descriptors:

To further explore the potential of the AAD sounds' text descriptors as predictors of the emotions they would induce, a more capable, large language model was used. One restriction of the fine-tuned version of the DistilRoBERTa-base model used in the previous sub-section is that it performs emotion *classification*, whereas AAD represents emotion using continuous values of arousal and valence, thus being intended for *regression-oriented* ML tasks.

In this case, the ChatGPT API (<https://platform.openai.com/docs/api-reference/introduction>) was provided with the descriptors for all 780 sounds in the AAD and prompted to respond with arousal and valence values for each. The GPT-4 model was called via the API in system mode and provided with the prompt:

You will be provided with a text string associated with an audio sample. Based only on each text string, output the values of induced affect for the arousal and valence dimensions between 0 (low) and 1 (high). Respond in the following format: x, y where x and y are the continuous values of arousal and valence that must be between 0 (low) and 1 (high).

The GPT-4 default hyper-parameters were used (maximum tokens = 256, diversity via nucleus sampling 'top-p' = 0.5, frequency penalty = 0, presence penalty = 0), with the exception of temperature, which was reduced so that the model is likely to generate more deterministic, consistent predictions (temperature = 0.1). However, recent studies in other application domains (semantic tasks, medical examinations, legal bar examinations, and multiple choice question problem-

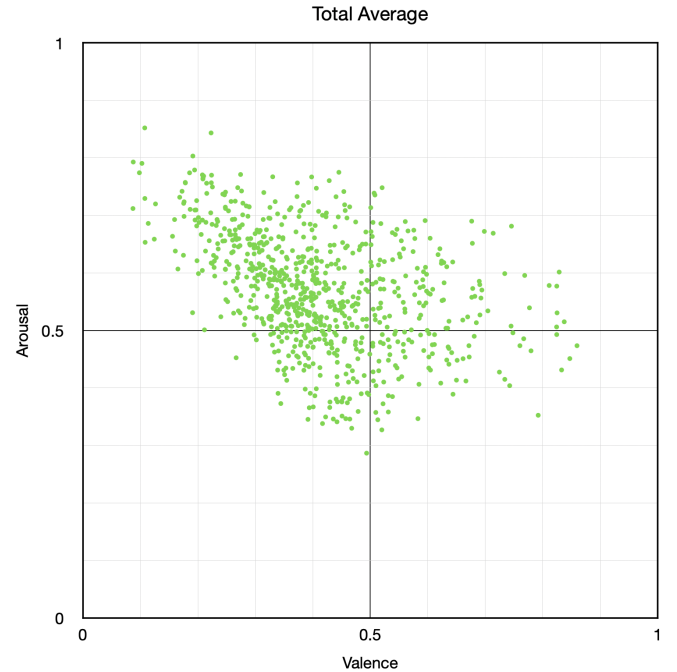


Fig. 5. Scatter chart of average arousal and valence of sounds in the AAD.

solving tasks), indicate that output from GPT models is often unaffected by variations in temperature [48]–[51].

The arousal and valence output generated by GPT-4 was compared to the mean AAD annotations obtained via the AMT workers. Pearson's correlation coefficient and Mean Absolute Error (MAE) were calculated for the data. This showed a weak positive correlation for arousal $r = 0.335, p < 0.00001, MAE = 0.20$ and a moderate correlation for valence $r = 0.402, p < 0.00001, MAE = 0.16$. Overall, this supports findings of the experimentation with the DistilRoBERTa-base model, namely that the text descriptions do not seem to be especially useful in helping predict the emotions induced by sounds in the AAD.

A minor limitation encountered was that, during pilot testing of this approach, the GPT-4 model was found to produce arousal and valence scores with only one decimal place of accuracy, even when prompted to generate ratings to two or three places. As such, there is a loss of precision that accounts for some of the difference between its arousal and valence values and those of the AAD. An interesting side note is that the GPT-4 API was also asked to generate *perceived* ratings, which were identical to those for *induced* emotion in 77.82% of arousal and 77.31% of valence ratings.

IV. RESULTS AND COMPARISONS

Arousal and valence ratings for each sound in the AAD have been averaged and are displayed in Figure 5.

As in previous studies [2], [3], [34], [52], Cronbach's alpha was calculated for the AAD sounds using all participants' ratings, yielding $\alpha = 0.885$ for arousal and $\alpha = 0.854$ for valence, showing good level of reliability [53]. However, in comparison to previous AER datasets, such as IADS-E, AAD's internal reliability is slightly lower. It is notable,

TABLE III
AROUSAL AND VALENCE DESCRIPTIVE STATISTICS OF AAD

	Arousal	Valence
Mean	0.57	0.41
SD	0.10	0.14
CV	17.51%	33.43%
Mean (scaled 1-9)	5.56	4.28
SD (scaled 1-9)	1.80	2.12

however, that IADS-E's Cronbach's alpha was calculated using approximately 10% of their participants' data, whereas the α -scoring for this work was calculated using ratings from *all* participants.

The Coefficient of Variation (CV) and Standard Deviation (SD) were calculated for arousal and valence of the AAD and are shown in Table III. Mean arousal and valence scores of AAD sounds are compared alongside those of IADS-2 and IADS-E in Table IV, with their standard deviations. To compare absolute ratings across all three datasets, some of which used 9-point SAM scales, the authors have calibrated the AAD ratings to a 9-point scale.

The scaled CV values for arousal and valence ($CV_{arousal} = 32.37\%$, $CV_{valence} = 49.53\%$) are notably higher in this set of ratings than that of IADS-E ($CV_{arousal} = 21.04\%$, $CV_{valence} = 31.23\%$), showing wider variance in responses than in other datasets. It also shows a similar pattern, as found in IADS-E and IADS-2, that arousal ratings are more agreeable than valence ratings between participants.

A. Limitations

Comparing the scaled SD scores and CV values to that of IADS-E and IADS-2, as in Tables III and IV, it is evident that the ratings in this research may be less consistent than those of previous studies. IADS-E reported SD values of 1.00 and 1.54 for arousal and valence, whereas IADS-2 reported values of 1.16 and 1.76 respectively. While not directly comparable due to the difference in focus of the sounds in this research and numbers of participants involved, it does suggest that remote data collection may not be as reliable as in-person methods. On balance, the mean figures for arousal and valence are close to other studies, with greater variance in arousal than valence. However, there may be other factors causing this, such as:

- The AAD sounds are different than in other datasets. For example, IADS-2 and IADS-E use sounds that are the same with IADS-E adding more sounds to the original set of sounds in IADS-2. Whereas AAD uses sounds sourced only from the BBC Sounds Library.
- The participants may be more diverse in AAD as the only limitation was that participants must have considered themselves to have normal hearing. IADS-E, for example, included participants that were exclusively students within a Japanese university.
- The number of participants providing ratings is greater in AAD than many other datasets. For example, IADS-E had 207 participants, whereas AAD had 870 online participants and 27 in-person participants.
- When examining the arousal-valence relationship, it is evident that sounds in the low-valence space are often

TABLE IV
COMPARISON OF AROUSAL AND VALENCE RELIABILITY VALUES BETWEEN AAD, IADS-E AND IADS-2

Dimension		AAD (Scaled)	IADS-E	IADS-2
Arousal	Mean	5.56	5.85	5.84
	SD	1.8	1.00	1.16
	CV	32.37%	21.04%	19.86%
Valence	Mean	4.28	4.40	4.78
	SD	2.12	1.54	1.76
	CV	49.53%	31.23%	35.82%

TABLE V
DISTRIBUTION OF SOUNDS' MEAN RATINGS IN THE AROUSAL-VALENCE SPACE

	Low Valence	High Valence	Total
High Arousal	477	109	586
Low Arousal	121	73	194
Total	598	182	780

rated with a higher arousal, whereas the spread of arousal ratings in the high valence space is more even. This is consistent with previous research [2], [3], [54] and indicates that negative stimuli are generally perceived as more arousing than positive ones.

- Table V shows the spread of ratings within each quadrant of the arousal-valence space. It is clear that most AAD sounds belong to the low-valence, high-arousal class.

Further comparisons of participants, split by gender identity were carried out. Correlation coefficients were calculated for arousal and valence comparing female (49% of participants) with male (51% of participants) responses. Results were: arousal $r = 0.781$ and valence $r = 0.858$ with both significant $p < 0.0001$. The strong positive correlation indicates that both genders agree on the induced emotions for sounds in the AAD. Figures 6, 7, and 8 display the non-binary, male and female ratings graphically, showing a very similar distribution across the gender identities. It is of note that as a small percentage ($< 1\%$) of participants identified as non-binary, each participant identifying as such rated a different batch of sounds. Therefore, it was not possible to average results and raw data is displayed for the non-binary identifying participants.

The relationship between arousal and valence values for each sound was also investigated, with a correlation coefficient of $r = -0.427$, $p < 0.0001$, indicating a moderate trend towards becoming less arousing the higher the valence. This is similar to other research [3] in which it was noted that the lowest scoring valence sounds typically had a higher arousal rating.

The overall arousal-valence averages for all participants and sounds are shown in Figure 5, which also shows the clustering of sounds in the high-arousal, low-valence quadrant.

Table VI compares the content of AAD with other datasets in the fields of emotional audio and Table VII compares AAD to emotional speech datasets. Both tables show the total number of audio instances, and Table VI also shows the total number of non-musical and non-anthropomorphic sounds in each. This was omitted in Table VII as only AAD has this data. Where data was not available in the accompanying

TABLE VI
COMPARISON OF AAD AND OTHER EMOTIONAL AUDIO DATASETS

Dataset	No. of Sounds	% Non-musical, non-anthropomorphic	Average Duration (seconds)	Ratings per sound	Categorical or Dimensional	Dimensions	Categories
AAD	780	100%	6.3	40	Dimensional	Arousal Valence	n/a
IADS-E [3]	930	74%	6.7	22	Dimensional and Categorical	Arousal Valence Dominance	Happiness Sadness Fear
IADS-2 [2]	167	69%	6.7	100	Dimensional	Arousal Valence Dominance	n/a
AudioSet [36]	1,789,621	44%	10.0	Not defined	Categorical	n/a	Non-emotional (e.g. music, glass)
Horror Soundscapes [35]	97	Not defined	5-10 (range)	10	Dimensional	Arousal Valence Tension	n/a
Emotional Sound Database [18]	390	65%	3.5	4	Dimensional	Arousal Valence	n/a
AMG1608 [6]	1,608	0%	30.0	15-32	Dimensional	Arousal Valence	n/a
Emo-Soundscapes [34]	1,213	Not defined	6.0	Not defined	Dimensional	Arousal Valence	n/a

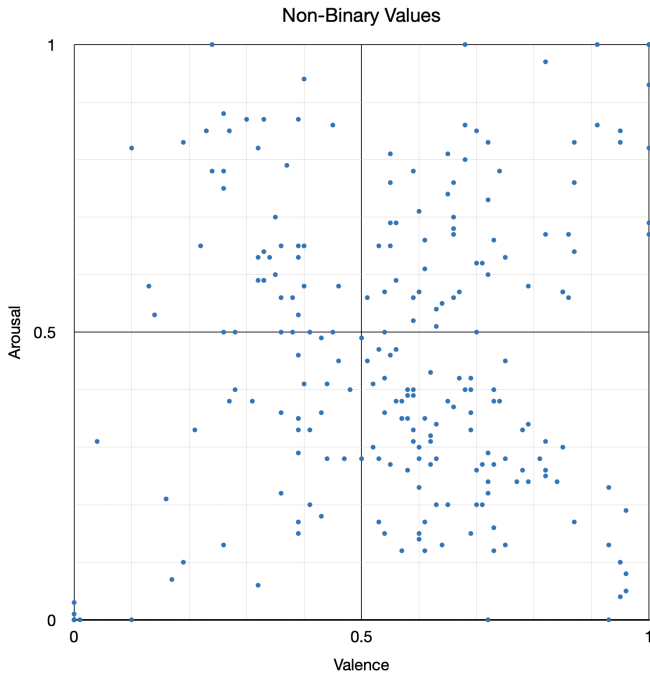


Fig. 6. Scatter chart of raw arousal and valence of each sound in AAD as rated by non-binary participants.

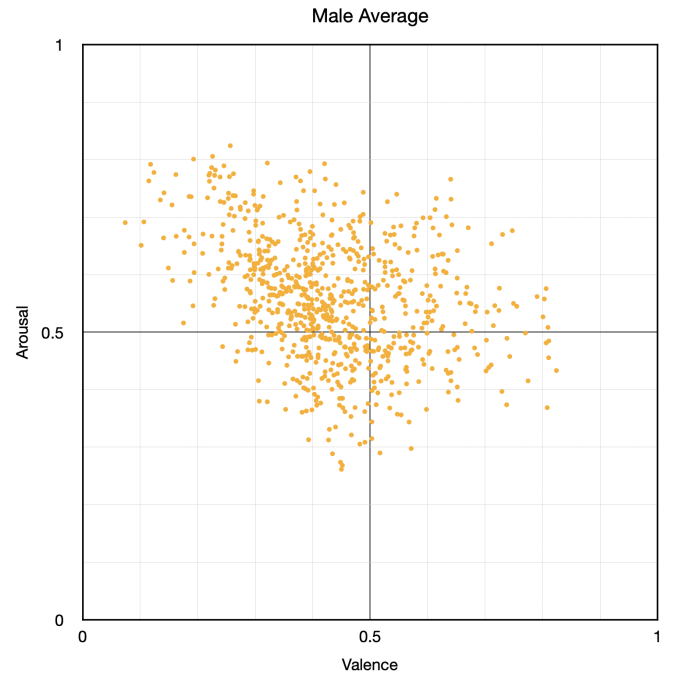


Fig. 7. Scatter chart of raw arousal and valence of each sound in AAD as rated by male participants.

manual or research paper, ‘not defined’ denotes this. For example, there was no reference found to ratings per sound in Emo-Soundscapes [34], and so ‘not defined’ is placed in the associated cell.

Comparing the total number of sounds, AAD is a smaller dataset in the field, but is the only one that contains exclusively non-musical, non-anthropomorphic content. Emotional speech datasets were included to show the disparity in total data

collected between these fields. Typically, emotional speech datasets contain more samples than emotional audio datasets.

It is noteworthy that AudioSet [36] groups sounds into more than one category and therefore the number of non-musical, non-anthropomorphic sounds may be less than shown. Data for this table was collated using the respective datasets’ published categories, and AudioSet’s non-musical, non-anthropomorphic sounds total was assumed using their ‘music’ category only.

TABLE VII
COMPARISON OF AAD AND EMOTIONAL SPEECH DATASETS

Dataset	No. of Sounds	Average Duration (seconds)	Ratings per sound	Categorical or Dimensional	Dimensions	Categories
AAD	780	6.3	40	Dimensional	Arousal Valence	n/a
CMU-MOSEI [55]	23,500	7.3	Not defined	Categorical	n/a	Happiness, Sadness, Anger, Disgust, Surprise, Fear
CREMA-D [56]	7,442	2.6	10	Categorical	n/a	Happiness, Sadness, Anger, Disgust, Fear, No Emotion
RAVDESS [57]	7,356	1.8	319	Categorical	n/a	Neutral, Calm, Happy, Sad, Angry, Fearful
ESD [58]	3,500	3.0	Not defined	Categorical	n/a	Neutral, Happy, Sad, Angry, Surprise
SEWA [59]	1,990	79.0	5	Dimensional	Arousal Valence	n/a
RECOLA [60]	23	300.0	2	Both	Arousal Valence	Agreement, Disagreement, Dominance, Engagement, Performance, Rapport

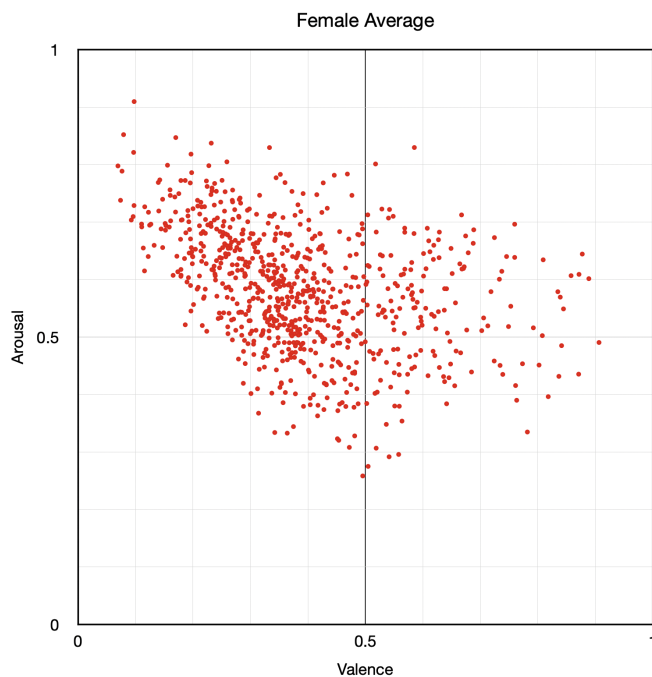


Fig. 8. Scatter chart of raw arousal and valence of each sound in AAD as rated by female participants.

It should also be highlighted that Emo-Soundscapes has a total of 1,213 sounds derived from a source of 613 sounds [34] and annotates for *perceived*, not *induced*, emotion.

B. Testing AAD for Audio Emotion Recognition

Previous research [24] found that utilising an Artificial Neural Network (ANN) gave good results in predicting arousal and valence in sound. Based on this previous success, an ANN was built to provide a baseline platform for preliminary testing of the AAD dataset for use in AER. Audio features utilized in the previous research were extracted from the 780 sounds in AAD and stored. Extracted features included 20 Mel-Frequency Cepstral Coefficients (MFCCs), zero-crossing

rate, spectral centroid, spectral bandwidth, rolloff, and Root Mean Square (RMS) loudness values of the sounds.

The dataset was separated into training and testing segments, according to an 80/20 split. A systematic approach to testing multiple ANNs variants was taken to evaluate the effectiveness of varying amounts of hidden layers and neurons at predicting arousal and valence values. As in previous research [24], tests were undertaken on an ANN consisting of one hidden layer with 2, 4, 8 and 16 neurons. Further, ANNs consisting of 4, 8, and 16 hidden layers with 128, 256 and 512 neurons in each layer were also tested. The results from these variations are shown in Table VIII.

Comparing the various iterations of the ANN, the single-layer variants performed poorly, with negative R^2 values as low as -10.750 for arousal prediction and -4.867 for valence prediction (1-layer, 4-neurons), indicating that the models were a very poor fit for the data. However, of the variants that produced positive R^2 values, the 8- and 16-layer versions performed best for in both Root Mean Square Error (RMSE) and R^2 metrics. For arousal prediction, the 8-layer, 512-neuron ANN gave the lowest RMSE and the highest R^2 value. For valence prediction, the 16-layer, 512-neuron ANN gave both the lowest RMSE and highest R^2 values.

Scatter charts depicting the best-performing arousal and valence predictor outputs are shown in figures 9 and 10. The best fit lines clearly shows arousal prediction to be more accurate than that of valence.

As in previous research [21], [24], valence proved more difficult to predict than arousal using the features extracted from the AAD sounds. However, while previous research suggested that varying the amount of neurons did little to improve R^2 values, in this case a general trend of improving R^2 values with increasing neurons in each layer presented itself, and is shown in Table VIII.

The R^2 values across testing are generally low, but are consistent with other research [21], [24] in the AER field. As this ANN was designed only to produce a baseline performance metric using the AAD, no fine-tuning or advanced weighting of features was undertaken. Further tuning of the ANN and

TABLE VIII
ANN TEST: RMSE AND R^2 FOR AAD (BEST-PERFORMING VALUES IN BOLD).

Layers	Neurons	Arousal RMSE	Arousal R^2	Valence RMSE	Valence R^2
1	4	0.327	-10.750	0.322	-4.867
1	8	0.151	-1.505	0.216	-1.627
1	16	0.017	-2.039	0.187	-0.969
2	128	0.156	-1.665	0.167	-0.581
2	256	0.129	-0.832	0.149	-0.253
2	512	0.287	-0.289	0.143	-0.148
4	128	0.091	0.080	0.139	-0.099
4	256	0.102	-0.138	0.135	-0.039
4	512	0.082	0.263	0.129	0.055
8	128	0.092	0.065	0.133	-0.006
8	256	0.085	0.199	0.134	-0.018
8	512	0.081	0.279	0.130	0.044
16	128	0.092	0.067	0.132	0.003
16	256	0.087	0.158	0.132	0.013
16	512	0.082	0.259	0.128	0.073

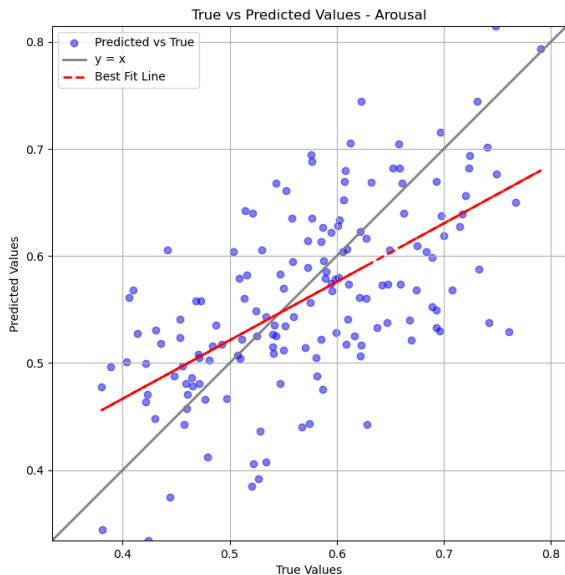


Fig. 9. True vs Predicted arousal using 8-layer ANN with 512 neurons per layer.

investigation into feature importance should be carried out to understand if the ANN accuracy can be improved, alongside application of other ML techniques.

V. CONCLUSIONS AND FUTURE WORK

In this article 780 sounds were collected based on their suitability for use in film and/or TV soundtracks. Their affective qualities have been rated and standardized in batches by 870 participants. Results showed that a relatively large dataset of sounds, accurately annotated for affective qualities via dimensional scales can be provided to the wider affective audio research community. Although male, female, and non-binary annotators had some variance in affective rating, the difference was minor. Further research should take into consideration

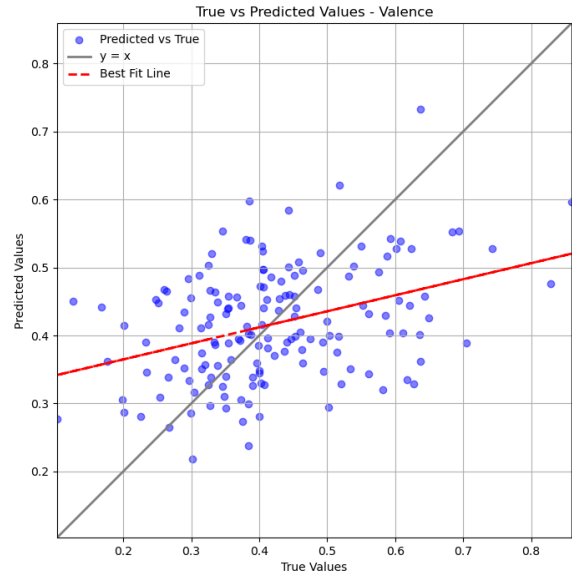


Fig. 10. True vs Predicted valence using 16-layer ANN with 512 neurons per layer.

the gender-identity variation in ratings in affective responses wherever possible and appropriate.

It is well known that real-use cases of emotion evocation in film and TV will be caused by a combination of the visual *and* sonic elements [4], [54], [61], [62] as well as any other sensory modalities. The effective use of affective audio may enable post-production teams to enhance or guide the emotive responses of audiences as they see fit for their story or project [24].

While it is important to remember the intended use of the AAD is somewhat different to others before it, the authors have recognized limitations to this research, including:

- The distribution of sound stimuli in the arousal-valence space is not even and is heavily skewed towards the high-arousal, low-valence space. This may present issues of imbalance or cause limitations in using the AAD in AER applications.
- Participant ratings are more variable than that of comparable datasets. This may be caused by multiple factors including the stimulus materials (no human vocalization or musical sounds were used in AAD), the data collection methodology is different (data was collected remotely in AAD, against in-person in others), and cultural differences may have an effect, as IADS-E identified a significant difference in valence ratings based on cultural background [3].
- No capture of basic emotion responses was undertaken as part of this research, as in others. It may, however, be possible to determine basic emotions for each sound by mapping their mean arousal-valence scores to pre-defined values [63].

The authors plan future research using this dataset to involve

the training and fine-tuning of affect-recognition models, such as the ANNs described in this article, to predict the affective responses of listeners to non-musical, non-vocalized sounds. The use of other ML predictors alongside ANNs are also suggested, alongside deep learning methods.

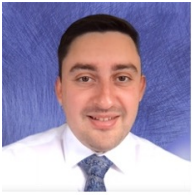
It is the principal author's intent that using an effective audio emotion prediction model, an affect prediction tool may then be created and tested for use in the audio post-production profession. Through this research the AAD may enable further comparison of results across other auditory disciplines, as well as in research and professional applications, such as AER and film, media studies, advertising, and more. Other potential uses of this dataset may include, but are not limited to: psychology and sound; user interface development and design; sonic interaction; game sound; audio processing; and cultural studies.

REFERENCES

- [1] H. Ridley, S. Cunningham, and R. Picking, "Developing an affective audio toolbox for audio post-production," in *ITNG 2022 19th International Conference on Information Technology-New Generations*. Springer, 2022, pp. 371–378.
- [2] M. Bradley and P. Lang, "The international affective digitized sounds (2-nd edition; iads-2): Affective ratings of sounds and instruction manual gainville," *The Center for Research in Psychophysiology: Gainesville, FL, USA*, 2007.
- [3] W. Yang, K. Makita, T. Nakao, N. Kanayama, M. G. Machizawa, T. Sasaoka, A. Sugata, R. Kobayashi, R. Hiramoto, S. Yamawaki *et al.*, "Affective auditory stimulus database: An expanded version of the international affective digitized sounds (iads-e)," *Behavior Research Methods*, vol. 50, pp. 1415–1429, 2018.
- [4] E. Weis and J. Belton, *Film Sound: Theory and Practice*. New York: Columbia University Press, 1985.
- [5] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, no. 39-58, p. 3, 1997.
- [6] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 693–697.
- [7] M. Soleymani, A. Aljanaki, and Y. Yang, "Deam: Mediaeval database for emotional analysis in music," *Geneva, Switzerland*, 2016.
- [8] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, p. 166335, 2022.
- [9] X. Han, F. Chen, and J. Ban, "Music emotion recognition based on a neural network with an inception-gru residual structure," *Electronics*, vol. 12, no. 4, p. 978, 2023.
- [10] X. Hu and Y.-H. Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.
- [11] S. Mo and J. Niu, "A novel method based on ompgw method for feature extraction in automatic music mood classification," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 313–324, 2017.
- [12] P. Saari, G. Fazekas, T. Eerola, M. Barthelet, O. Lartillot, and M. Sandler, "Genre-adaptive semantic computing and audio-based modelling for music mood annotation," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 122–135, 2015.
- [13] A. Roda, S. Canazza, and G. De Poli, "Clustering affective qualities of classical music: Beyond the valence-arousal plane," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 364–376, 2014.
- [14] J. Yang, "A novel music emotion recognition model using neural network technology," *Frontiers in psychology*, vol. 12, p. 760060, 2021.
- [15] E. Y. Koh, K. W. Cheuk, K. Y. Heung, K. R. Agres, and D. Herremans, "Merp: a music dataset with emotion ratings and raters' profile information," *Sensors*, vol. 23, no. 1, p. 382, 2022.
- [16] J. S. Gómez-Cañón, N. Gutiérrez-Páez, L. Porcaro, A. Porter, E. Cano, P. Herrera-Boyer, A. Gkiokas, P. Santos, D. Hernández-Leo, C. Karreman *et al.*, "Trompa-mer: an open dataset for personalized music emotion recognition," *Journal of Intelligent Information Systems*, vol. 60, no. 2, pp. 549–570, 2023.
- [17] E. M. Schmidt and Y. E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *ISMIR*, vol. 11. Miami (Florida), USA, 2011, pp. 777–782.
- [18] E. Çano and M. Morisio, "Moodylyrics: A sentiment annotated lyrics dataset," in *Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence*, 2017, pp. 118–124.
- [19] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audiovisual emotion database," in *22nd international conference on data engineering workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.
- [20] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.
- [21] B. Schuller, S. Hantke, F. Wening, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 341–344.
- [22] K. Drossos, A. Floros, and N.-G. Kanellopoulos, "Affective acoustic ecology: Towards emotionally enhanced sound events," in *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, 2012, pp. 109–116.
- [23] K. Drossos, R. Kotsakis, G. Kalliris, and A. Floros, "Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal," in *IISA 2013*. IEEE, 2013, pp. 1–6.
- [24] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021.
- [25] Y. Choi, S. Lee, S. Jung, I.-M. Choi, Y.-K. Park, and C. Kim, "Development of an auditory emotion recognition function using psychoacoustic parameters based on the international affective digitized sounds," *Behavior research methods*, vol. 47, pp. 1076–1084, 2015.
- [26] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.
- [27] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [28] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, p. 550–553, 1992.
- [29] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.
- [30] I. J. Roseman, "Cognitive determinants of emotion: A structural theory," *Review of personality & social psychology*, 1984.
- [31] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press, 2004.
- [32] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
- [33] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1990.
- [34] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2017, pp. 196–201.
- [35] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 209–222, 2017.
- [36] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [37] ITU-R, "Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Union Radiocommunication Assembly, Geneva, CH, Standard, Oct. 2015.
- [38] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision making*, vol. 5, no. 5, pp. 411–419, 2010.
- [39] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A comparative study of collaborative vs. traditional musical mood annotation," in *ISMIR*, vol. 104, 2011, pp. 549–554.

- [40] J. H. Lee and X. Hu, "Generating ground truth for music mood classification using mechanical turk," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 2012, pp. 129–138.
- [41] T. X. Seow and T. U. Hauser, "Reliability of web-based affective auditory stimulus presentation," *Behavior research methods*, vol. 54, no. 1, pp. 378–392, 2022.
- [42] A. Betella and P. F. Verschure, "The affective slider: A digital self-assessment scale for the measurement of human emotions," *PLoS one*, vol. 11, no. 2, p. e0148037, 2016.
- [43] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
- [44] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford University Press, 2007.
- [45] P. Ekman, *The Handbook of Cognition*, T. Dalgleish and T. Power, Eds. John Wiley and Sons Ltd., 1999.
- [46] J. Hartmann, A. R. Murthy, and K. M. A. Kumar. (2021) MS Windows NT kernel description. [Online]. Available: <https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/>
- [47] S. Havaldar, B. Singhal, S. Rai, L. Liu, S. C. Guntuku, and L. Ungar, "Multilingual language models are not multicultural: A case study in emotion," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 202–214. [Online]. Available: <https://aclanthology.org/2023.wassa-1.19>
- [48] L. Zhang, M. Wang, L. Chen, and W. Zhang, "Probing gpt-3's linguistic knowledge on semantic tasks," in *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2022, pp. 297–304.
- [49] M. Rosoł, J. S. Gasior, J. Łaba, K. Korzeniewski, and M. Młyńczak, "Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination," *Scientific Reports*, vol. 13, no. 1, Nov. 2023.
- [50] E. Martínez, "Re-evaluating gpt-4's bar exam performance," *Artificial Intelligence and Law*, pp. 1–24, 2024.
- [51] M. Renze and E. Guven, "The effect of sampling temperature on problem solving in large language models," *arXiv preprint arXiv:2402.05201*, 2024.
- [52] S. Sundaram and R. Schleicher, "Towards evaluation of example-based audio retrieval system using affective dimensions," in *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010, pp. 573–577.
- [53] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011.
- [54] P. Susini, O. Houix, and N. Misdariis, "Sound design: an applied, experimental framework to study the perception of everyday sounds," *The New Soundtrack*, vol. 4, no. 2, pp. 103–121, 2014.
- [55] P. P. Liang, R. Salakhutdinov, and L.-P. Morency, "Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion," in *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, vol. 1, no. 2, 2018, p. 3.
- [56] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [57] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS one*, vol. 13, no. 5, p. e0196391, 2018.
- [58] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [59] J. Kossaiifi, R. Walecki, Y. Panagakis, J. Shen, M. Scmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1022–1040, 2021.
- [60] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, J. Barnes, O. De Clercq, and R. Klinger, Eds. IEEE, Apr. 2013, pp. 1–8.
- [61] D. Sonnenshein, "Sound design: The expressive power of music, voice and sound effects in cinema. studio city," 2002.
- [62] F. Winckel, *Music, Sound and Sensation: A Modern Exposition*. Courier Corporation, 1967.
- [63] D. Griffiths, S. Cunningham, J. Weinel, and R. Picking, "A multi-genre model for music emotion recognition using linear regressors," *Journal of New Music Research*, vol. 50, no. 4, pp. 355–372, 2021.

VI. AUTHOR BIOGRAPHIES



Harrison Ridley is a PhD student at University of Chester, in the field of affective audio and machine learning. He has a bachelor's degree in music technology (Hons) from Wrexham Glyndŵr University (2016). Other than his primary research field, he has a keen interest in affective computing, audio for film, audio in the natural environment, and audio conservation.



Stuart Cunningham is a Senior Lecturer, specializing in User Experience (UX), at the University of Chester. He received his BSc and MSc awards from the University of Paisley in the fields of Computer Networks (2001) and Multimedia Communications (2003) respectively. Following this he was awarded the PhD from the University of Wales (2009) in audio data compression. His research interests cover a range of computing and creative hybrids, including audio compression, affective technologies, human-computer interaction, and sonic interaction.



John Darby is a Senior Lecturer in the Department of Computing and Mathematics at Manchester Metropolitan University. He holds an undergraduate degree in Computational Physics (BSc (Hons.) from the School of Physics and Astronomy at The University of Edinburgh (2003) and an MSc in Mobile and Distributed Computer Networks from Leeds Metropolitan University (2004). He received his PhD from Manchester Metropolitan University in 2010. His research is around Computer Vision approaches for the tracking and analysis of human movements.



John Henry is a Senior Lecturer in Computer Games at Manchester Metropolitan University with research interests in the Internet of Things applications for good, including healthcare, and the combination of game experiences with the Internet of Things for good. His research background includes developing simulations for raising awareness and measuring student engagement at the university level of study through a Serious Game that embedded the Internet of Things.



Richard Stocker is a Senior Lecturer at the University of Chester, with research interests in the simulation and formal verification of human-agent-robot teamwork. Richard received his BSc in Computer Science (2005), MSc in Advanced Computer Science (2008), and PhD on the topic of Towards the Formal Verification of Human-Agent-Robot Teamwork (2013) from the University of Liverpool.