



# Supervised machine learning for audio emotion recognition

## Enhancing film sound design using audio features, regression models and artificial neural networks

Stuart Cunningham<sup>1</sup> · Harrison Ridley<sup>1</sup> · Jonathan Weinel<sup>2</sup> · Richard Picking<sup>3</sup>

Received: 10 December 2019 / Accepted: 5 March 2020 / Published online: 22 April 2020  
© The Author(s) 2020

### Abstract

The field of Music Emotion Recognition has become an established research sub-domain of Music Information Retrieval. Less attention has been directed towards the counterpart domain of Audio Emotion Recognition, which focuses upon detection of emotional stimuli resulting from non-musical sound. By better understanding how sounds provoke emotional responses in an audience, it may be possible to enhance the work of sound designers. The work in this paper uses the International Affective Digital Sounds set. A total of 76 features are extracted from the sounds, spanning the time and frequency domains. The features are then subjected to an initial analysis to determine what level of similarity exists between pairs of features measured using Pearson's  $r$  correlation coefficient before being used as inputs to a multiple regression model to determine their weighting and relative importance. The features are then used as the input to two machine learning approaches: regression modelling and artificial neural networks in order to determine their ability to predict the emotional dimensions of arousal and valence. It was found that a small number of strong correlations exist between the features and that a greater number of features contribute significantly to the predictive power of emotional valence, rather than arousal. Shallow neural networks perform significantly better than a range of regression models and the best performing networks were able to account for 64.4% of the variance in prediction of arousal and 65.4% in the case of valence. These findings are a major improvement over those encountered in the literature. Several extensions of this research are discussed, including work related to improving data sets as well as the modelling processes.

**Keywords** Affect · Arousal · Audio emotion recognition · Audio features · Emotion · IADS · Regression · Neural networks · Valence

### 1 Introduction

This article extends our previous work, where we presented promising, but initial, results of the use of supervised

machine learning techniques in the field of Audio Emotion Recognition (AER), with the intention of using such information to help sound designers navigate large sound databases and select the most emotionally effective material [15]. In this section, the underpinning concepts of affect recognition in sound are introduced. The section begins by explaining emotion recognition tasks, models and approaches before describing the data set employed in our work. The importance of emotional sound is highlighted with a particular emphasis on its application in film and other visual media.

#### 1.1 Affective computing and audio

Affective computing is an interdisciplinary research field concerned with the emotional interaction between technology and humans [48]. The field of Music Emotion Recognition (MER) is one such subset of this broad field and has received considerable attention from the research community in recent years [16, 32, 43, 52, 52, 54, 66]. In this

---

✉ Stuart Cunningham  
s.cunningham@mmu.ac.uk

Harrison Ridley  
HARRISON.RIDLEY@stu.mmu.ac.uk

Jonathan Weinel  
weinelj@lsbu.ac.uk

Richard Picking  
r.picking@glyndwr.ac.uk

<sup>1</sup> Manchester Metropolitan University, Manchester, M1 5GD, UK

<sup>2</sup> London South Bank University, London, SE1 0AAB, UK

<sup>3</sup> Wrexham Glyndŵr University, Wrexham, LL11 2AW, UK

article, however, we turn our focus to the area of AER, which deals with affect in non-musical sound. The AER field has received less attention in the literature, although we make the case that it is equally as relevant. This is particularly true, for example, in the task of sound design for media such as television, computer games and film, where sound effects are typically coupled with music to direct the perception of the audience [11]. For the purposes of this article, we define sound effects as including the many layers of audio, whether they be natural or artificial in origin, and sounds other than music, that are found in media, namely ambience, Foley, dialogue, and human non-verbal utterances.

Gerhard [25] provided a considered and useful taxonomy of sounds that can be appropriated in the context of the research that we plan and report upon here. In particular, the taxonomy contains four classes of hear-able sound: noise, natural sounds, artificial sounds, speech, and music, each with a number of sub-categories. The sounds that are used in this article provide a comprehensive number of samples that fall broadly within Gerhard's categories of natural sounds, artificial sounds, non-verbal speech, and the sounds created by instruments, which Gerhard allocates to the music class as "...sound made [by] humans using instruments...".

Traditionally, affective computing makes use of theoretical models of emotion. The most common models encountered are either categorical or dimensional. Categorical models use qualitative descriptions, commonly text-based, to identify discrete emotions, whereas dimensional models use quantitative values on one or more dimensions. An example of a categorical model can be seen in the work of Ekman [22] or Panksepp [47], whereas dimensional models may be seen in those of Thayer [63] or Russell [53].

The research documented in this article adopts the use of the latter: Russell's circumplex model of affect, which is a two-dimensional Cartesian emotion space consisting of axes relating to arousal (vertical) and valence (horizontal) [53]. Such an approach is typical in the field of emotion recognition. Although our work focuses upon the affective analysis of audio, it is worth making the observation that, in the field of Music Emotion Recognition (MER), it is typically reported that models for the prediction of the arousal dimension tend to outperform those of valence [32, 43, 52, 54, 54].

Models for the prediction of emotion in media make use of the coefficient of determination  $R^2$  as a performance metric. It is based upon a set of dependent variables  $Y$  output from a regression model and calculated knowing the set of independent input variables  $X$  and is calculated as

$$R^2 = \left( \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 \quad (1)$$

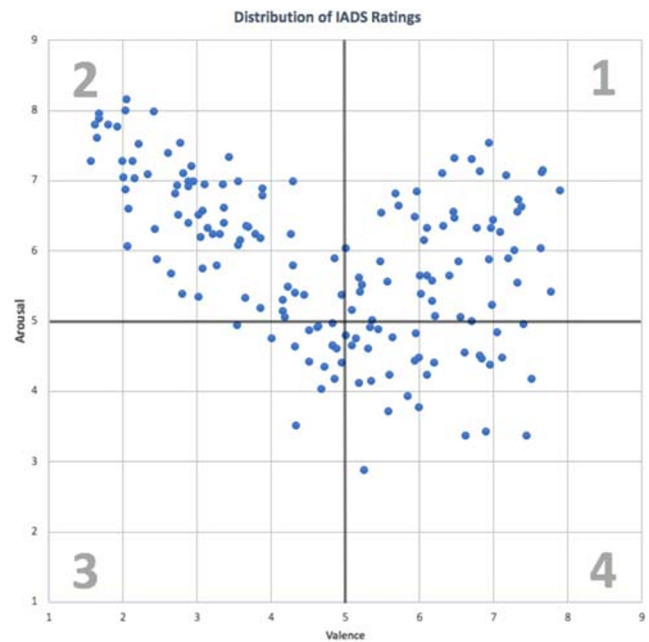


Fig. 1 IADS mean values in arousal and valence space

where  $\bar{x}$  and  $\bar{y}$  are the means of sets  $X$  and  $Y$ .

In the field of MER, the upper range of  $R^2$  values for arousal is approximately 80 to 85% and approximately 60 to 70% for valence [20, 32, 39]. An aim of our work is to determine if similar levels of performance can be achieved in AER.

## 1.2 The IADS data set

An existing corpus of validated, emotionally annotated sounds exists in the form of the International Affective Digitized Sound (IADS) system [8]. The IADS set provides 167 varied sounds and their associated emotional ratings, obtained through the Self-Assessment Manikin (SAM) approach on the dimensions of arousal, valence and dominance [6]. Ratings are presented for each dimension using a 9-point scale and each sound has been rated by a minimum of 100 participants. The mean duration of the 167 sound samples in the IADS set is 6.014 s ( $\sigma = 0.017$  s). The overall distribution of the IADS ratings in arousal and valence space is illustrated in Fig. 1 and broken down by quadrant in Table 1. By extracting data relating to the

Table 1 Quadrant distribution of IADS ratings

| Quadrant | Number of sounds |
|----------|------------------|
| Q1       | 51               |
| Q2       | 70               |
| Q3       | 16               |
| Q4       | 30               |

arousal and valence dimensions therein, we make use of the IADS in our attempt to create computational models of emotional response to sound.

Nardelli et al. [44] used the valence and arousal ratings from IADS and heart rate variability (HRV) data collected from participants who were listening to the rated sound. The participants' HRV was measured as they listened to various "arousal sessions", compiled of similarly arousing sounds. Using the data collected, the researchers could automatically recognise emotions elicited by affective sounds with considerable accuracy (84%), in both valence and arousal. This study effectively proves that the IADS data set can be useful for research into audio and emotions, particularly when attempting to predict emotional affect.

### 1.3 Affective audio

Whilst Picard [48] argues that human-computer interactions can be improved through the design of systems that represent, recognise, respond to or have emotions, these concerns are also significant for a variety of media, namely games, audio-visual art and film, which increasingly are embedded within computer systems. For instance, Weinel's work [64] on altered states of consciousness argues that the design of various electronic music and audio-visual media allows the transmission of affective properties to audiences. He argues that these affective properties are combined with representational properties, which frame the emotive aspects of the media with different forms of conceptual meaning. According to this argument, non-diegetic music is a central feature of media that elicits states of positive or negative valence and arousal (following Russell's circumplex model of affect [53]). For instance, Gabrielsson and Lindstrom's [24] meta-study of music and emotion revealing musical features, including rhythm, melody, pitch and tonality, may often be associated with specific affective responses. Weinel [64] argues that these features are primarily involved in the production of affect, whilst non-diegetic sounds or images may frame these with representational meaning. Yet he also notes that there is inevitably some overlap between these broad categories.

Considering this overlap, affective properties are elicited in conjunction with other representational aspects of these media, such as diegetic audio, or visual representations, which may suggest places, spaces and narratives. Sound effects, for instance, may reference real or imaginary locations (through soundscapes, for instance), and suggest sequences of events. The primary role of these is often to reveal the diegesis, conjuring these spaces for the audience. Yet diegetic sound may also have emotional resonances for the listener. For instance, following R. Murray Schafer's [56] discussion, we may consider how sounds like alarms or dogs barking have representational

or symbolic meaning, and also give rise to emotional responses. Such emotional responses may include aspects that are culturally shared, and those that are highly individualised and subjective [56]. For example, the sound of an alarm ringing is widely understood to indicate forms of alert, corresponding with high arousal, calling those who can hear it to action in some form—whether to take action to prevent the breakout of a fire, or, in a musical context where this sound effect is often used by DJs at raves, to trigger ecstatic dance [51]. Such sounds can be understood as cross-cultural, relating to shared cultural knowledge and semantic memory (in terms of Schacter and Tulving's [55] theory of semantic and episodic memory). Yet alarms can also trigger highly individualised responses, for example in persons living with post-traumatic stress syndrome (PTSD), the sound may trigger traumatic autobiographical episodic memories [21].

### 1.4 Affective sound design for film and other media

It has been suggested [5] that sound design can "actively shape how we perceive the image". The shaping of an audience's experience through the use of sound can be performed by the elicitation of emotions, amongst other techniques. This section shall discuss some methods of purposefully shaping an audience's response as well as look at examples of previous research into the matter.

The use of the "affective qualities" of sound may communicate "dramatic tone, atmosphere and mood" [17], whilst also describing the fictional world, giving it a "particular toning" [34]. The use of sound in this way, to create a more detailed and believable world, is useful for filmmakers to envelop their audience within the fictional world, to experience it alongside the characters that inhabit it [5, 9].

We hypothesise that by using affective audio in cinema, a filmmaker may be able to ensure their audience feels a particular emotion at specific points in the film. Some research into how this can be implemented has already been undertaken. Most notable is the work of Hillman and Pauletto [29, 30], which concluded that a "Four Sound Areas framework" in which sound design is broken down into four areas (logical, abstract, temporal and spatial) would afford a more flexible approach to "emotive sound design".

The induction and perception of emotion by hearing non-speech, non-musical sounds are just as intuitive as their speech and musical counterparts [65]. Consider the elicitation of tension and alarm upon hearing the sound of a pulsing evacuation warning, the excitement and delight of the sound of a bottle cork proudly popping open, the calm and contentment of hearing the sea lapping gently ashore, or the sadness and depression of hearing the life sputter out of a car engine. Representations of such sounds are frequent in

many forms of popular media and may be present alongside music and speech or entirely on their own, at the discretion of the sound design and production team [3, 4, 27].

Sound in the real-world is known to cause affect. The everyday soundscapes that we experience may change our mood. For example, for someone from the countryside, the soundscape when visiting a busy city may cause distress or unease as they are not accustomed to the sounds [33, 56]. Further, the notion of “*Acoustic Violence*” coined by Miyara [42] and described as sound that is invasive, exudes power or prominence, or is not wanted, may also cause affect. Consider the sound of a newly built airport, and the affect it has on the local community. The airport exerts its acoustic power over the local community and comes to define it. Over time, the local people may consider the sound to become a part of the local soundscape. At this point, it is no longer a violent invasion but a keynote [56] — a background sound that is part of life. One could now argue that to take this sound away is in itself acoustic violence, changing the soundscape that local residents live with. The unwanted sound is the acoustic violence in either case. This scenario looks at both sides of interpretation, and such techniques may be used as plot devices.

Successful sound design does not only access the representational aspects of sound but also taps into this capability to elicit culturally shared affective properties. In films and video games, the diegetic soundscape extends the spatial representations of these media beyond the limited space of the screen, allowing the construction of believable environments. Yet the affective properties of sound furnish these narrative spaces with cues and triggers for mood and emotion. For example, in *Wall Street* [61], one scene cuts from Bud Fox’s budget New York apartment to Gordon Gekko strolling a beach at dawn. As Gekko tempts Fox with a business proposition that could fulfil his wildest capitalist fantasies, the contrast between the mundane and the affluent sublime are underscored by the contrasting sounds of street noise and ocean waves lapping on the beach; the full frequency sound of Gekko’s voice speaking into his expensive mobile phone, and the low-fidelity simulacrum which comes through Fox’s landline. In Michael Mann’s films such as *Miami Vice* [40], we similarly find sublime tropical environments contrasted with dense industrial landscapes or seedy urban sprawl. Here, sounds of waves or palm trees rustling in the wind similarly create a tangible, affective, aural sense of the sublime, which contrasts with the noise of the action sequences. Considering the latter, in the main shootout sequence of *Heat*, after the main crew exit from a bank robbery, for approximately 5 min we hear no music—only machine gun fire, shattering glass, squealing tires, and occasionally screams and shouts. Here, it is not music that gives a sense

of adrenaline and excitement, but rather the sound effects that delivers a high arousal affective experience for the audience.

Further, the building of soundscapes for film has been documented as a method for creating a mood, atmosphere or otherwise eliciting audience emotion. A noticeable example of positive use of such technique can be heard in *American Graffiti* [38]. In street scenes where teenagers are driving their cars, the background sound is filled with happy crowd noises, laughs and giggles, radios playing and so on [37]. All of this adds to the party atmosphere of the film, and may subtly elicit euphoric and sometimes nostalgic feelings in the audience, as it never over emphasises anything.

In video games, sound design often serves similar purposes as it does with film but in an interactive context. However, this scope for interactivity also gives rise to the use of sound for other goals, such as playing-along and dynamic sound manipulation game mechanics. The diegetic soundscape serves to make spatial environments convincing navigable spaces, leading to presence and immersion through what Cajella [10] refers to as spatial player involvement. Yet the affective properties of diegetic sound can also be understood as creating an affective sensorium, facilitating Cajella’s concept of affective player involvement. Through the combination of the two, we can think of games as providing interactive affective spaces, which may denote zones of safety and danger and reinforce rewarding and un-rewarding actions. Thereby, affect also contributes to ludic involvement, since it gives sensory cues regarding the relative success or behaviour of the player’s actions in the virtual world. With video games as with other audio-visual media, it is not only non-diegetic music but also diegetic sound that contributes towards the audience experience of affect, which in turn plays a pivotal role in the overall experience of the media.

The use of affective sound design in film and linear media has not advanced much since its most prominent use in the auteur renaissance of the 1970s and 1980s, particularly with directors such as Lucas and Coppola, and sound editors such as Murch and Burt [59]. In sound design for interactive media, such as videogames, the technical tools and capabilities have advanced significantly. Notably in games, the dynamic, real-time manipulation of sound for affective purposes is often seen, but even this relies upon the selection of appropriate sonic assets by the sound designer in the first instance [31]. A core intention of the work that follows in this paper is to be able to empower and enhance sound designers and their work, particularly for application in film. We envisage that computational models for AER will enable sound designers to create work with greater emotional impact and to evaluate their designs prior to audience trials.

## 2 Related work

This section begins by depicting existing research studies into the manifestation of human emotional responses to non-musical sounds. Following this, recent research specifically into AER is chronicled, highlighting the techniques employed and performance of the models created.

### 2.1 Audio affect identification

Existing research in the field of emotion recognition for audio-visual media, most often short video clips, represents the combination of audio and visual modes [14, 23, 28, 45, 46, 58]. Current research works in emotion recognition involve non-musical audio, but are typically focused upon human speech and utterances in the audio component, and therefore less concerned with what might be deemed sound effects and are more oriented towards human affect expression. Nevertheless, such current research examples have exploited machine learning approaches in emotion recognition classification tasks and have made use of techniques that included state vector machines, Bayesian methods, neural networks, linear discriminant analysis and so forth.

In noting that the presence of film when studying affective responses makes it difficult to isolate the sonic aspects, work by Bradley and Lang [7] gathered affective reaction data from acoustic stimuli. The study involved the playback of 60 sounds to test subjects, who were asked to rate how they felt when listening to the sounds based on arousal, valence and dominance using a scale of 1 to 9. The study found that the results followed a similar pattern to studies using the International Affective Picture System (IAPS) data set, with extreme ratings of pleasure having extreme ratings of arousal, and neutral levels of pleasure having low arousal ratings.

A study by Redondo et al. [50] replicated the original IADS experiments with the intention of finding differences in ratings based on cultural differences between American (original study) participants and Spanish participants. The study found that whilst Spaniards rated sounds in a very similar way to Americans, there were some, if only minor, differences. It found that Americans tend to rate sounds with more positive valence than Spaniards and as less activate in the arousal scale, but with a wider range. The study also noted that some specific sounds in the data set may be affected by cultural differences, giving the examples of American Football, which is seldom played in Spain, and the sounds of bombs. At the time of the writing of their study, the authors noted that explosive devices had been recently used in terror attacks in Spain.

In a similar manner to previously mentioned studies, research by Stevenson and James [60] aimed to predict the

arousal, valence and dominance for a set of sounds after categorising them into one of five emotions: happiness, anger, sadness, fear and disgust. Participants rated each of the IADS sounds on a scale of 1–9 for each emotion. The data from this experiment was used to label each sound in the IADS data set with one or more emotions. The conclusion was that valence and arousal were only effectively predicted in the fear emotion, for both positive and negative stimuli. The study acknowledged that even though the results obtained were not entirely useful for predicting responses, the categorisation of sounds it produced may be beneficial to future research.

### 2.2 Audio emotion recognition

Sundaram and Schleicher [62] conducted experiments modelling the affective response of listeners to a range of sound recordings. What makes their work novel was their use of recordings that might be considered complex, in that they did not represent a single attributable source. Instead, they were recordings of outdoor spaces and real environments, meaning each sound contained multiple, often overlapping, acoustic sources. The authors advocated a move away from the use of categorical models of emotion. Primarily, this was based upon the difficulties associated with using categorical approaches for sounds with multiple acoustic sources, and was also supported by the assertion that alternate approaches are already robustly employed in the field of experimental psychology. Therefore, their work makes use of a dimensional model, with ratings being produced on arousal, valence and dominance axes by using the Self-Assessment Manikin [6]. The work makes use of Latent Perceptual Indexing (LPI) to produce affective values for sounds using twelve Mel-Frequency Cepstral Coefficient (MFCCs) audio features. Sounds rated as similar in terms of their affect were also comparable in terms of their latent similarity index.

Drossos et al. [18] utilised the arousal and valence ratings for samples in the IADS set, which they described as being representative of sound events based upon a set of criteria defined from the literature. First, they performed an initial classification upon all sounds in the IADS set, determined by the quadrant location of each sound, as shown in Fig. 1. This meant that the process became a classification task, rather than the prediction of continuous variables representing arousal and valence. A range of typical audio features were then extracted and used in a series of training and validation exercises using support vector machines (SVM) and artificial neural networks (ANNs). Classification accuracy using these methods was reported at 43.7% for arousal and 36.5% for valence. The aforementioned findings must be considered in the context of a theoretical 25% allocation by chance, further skewed

by the distribution of the sounds, as evidenced in Fig. 1. The authors submit that traditional approaches used in MER may not be equally applicable to AER tasks.

The IADS set was used in another work by Drossos et al. [19] that examined rhythmic attributes of sound and their relationship to arousal. The authors elected to follow the approach of using a dimensional approach to dealing with arousal values, thereby avoiding the complexities associated with detaching dimensional components from categorical descriptors of affect. The approach employed seven different window lengths during the audio analysis. Prior to features being extracted, the data set was split into two groups, allocating the samples into a low or high arousal class. Six audio features were extracted along with statistics describing the shape of their distribution. Three approaches to classification were adopted: artificial neural systems, logistic regression and K-nearest-neighbour. The overall approach was shown to yield strong outcomes in performance, with the lowest outcome being 71.26% accuracy with ANS and 88.37% using logistic regression. However, these results must be contextualised against being a classification task of two categories, where chance would result in a theoretical outcome of 50% accuracy, which would then be further skewed by the distribution of the sounds between the two classes. This reflects sounds being attributed to either a low or high arousal class, not one where dimensional output is sought.

Schuller et al. [57] also recognised the value of researching AER in work that explored “realistic acoustic environment conditions”, which they classified into eight different subsets, such as animals, musical instruments, people and vehicles. Acknowledging the general lack of existing work and resources in the field of AER, the authors elected to construct their own data set, known as the Emotional Sound Database, sourced from an online sound repository. The sounds were annotated by a group of four participants, which was arguably a limitation of the corpus as a valid ground truth. A large number of audio features were extracted from the sounds and modelled with a regression approach, yielding results that equated to a  $R^2$  of 37.21% in the prediction of arousal and 24.01% for valence.

The potential for machine learning techniques to be employed in AER, specifically using the IADS set, was previously further verified by Choi et al. [12], who consider the audio clips in IADS to broadly represent those that would be encountered in daily life. Using a small set of audio features, namely loudness, sharpness, roughness and fluctuation strength, Choi and colleagues were able to demonstrate a better-than-chance discriminant function able to classify audio clips from IADS into one of three emotional factors: happiness, sadness and

negativity. Although Choi et al. describe another example of a classification approach to AER, rather than one of regression, their findings are notably robust. Such strength is due to an additional human-testing phase, involving 140 students, that was employed to help validate such a methodology. To this end, a new set of participant ratings was produced for the IADS set according to the three categories and accompanied by an additional test data set of 62 sounds, which were also subjected to the same procedure. Choi et al. found that when a reduced IADS set of 82 sounds were used as training data, a mean overall classification accuracy of 88.9% was reported for IADS sounds and 63.04% for the test data set.

### 3 Audio emotion recognition in IADS

This section explains our empirical work towards the recognition of emotion in the IADS set. It describes the methods of analysis, creation of models using regression and neural networks, and the performance of each model. The steps followed are graphically summarised in Fig. 2. The audio features used are defined and then extracted from the sounds in the IADS set. Resultant feature information was then subjected to an initial analysis to determine the similarity that existed between pairs of features, before being used as inputs to a multiple regression model to determine their weighting and relative importance. Following this, the features are used as inputs to the machine learning approaches of regression modelling and artificial

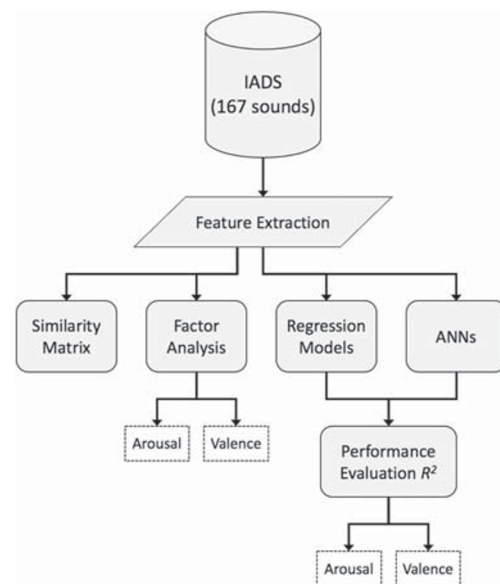


Fig. 2 Audio emotion recognition in IADS - workflow

neural networks so as to determine their ability to predict the emotional dimensions of arousal and valence.

### 3.1 Analysis method

All 167 sounds from IADS were employed within the analysis. Before work commenced, the sounds were peak-normalised, as advocated by other researchers [19], to control the loudness. The normalisation replicated the conditions reported by the originators of IADS in their participant study [8]. Audio features were extracted using the Matlab 2018b software and the Matlab Audio Analysis Library [26], using the settings of a 50-ms window with a 50% overlap. The adopted values have been shown effective in other works relating to emotion analysis in the domain of acapella singing [16]. The set of 35 features from the Audio Analysis Library were extracted, which comprise zero-crossing rate, energy, energy entropy, spectral centroid, spectral centroid, spectral entropy, spectral flux, spectral rolloff, the first 13 MFCCs, harmonic ratio, fundamental frequency and 12 chroma vectors. For each feature, mean and standard deviation were calculated.

In addition to these features, it was decided to incorporate other higher level data relating to the mode, harmonicity, distribution of energy and rhythmic elements, the latter being recognised as of value in prediction of emotional arousal in audio samples [19]. The following features were obtained by making use of the MIRToolbox [35] and included for analysis: inharmonicity, low energy, mode, tempo and pulse clarity. Finally, the location of the peak amplitude level, expressed in seconds, was added to give an indication of the attack envelope of each sound. Consequently, a total of 76 features were obtained for subsequent analysis.

Regression analysis was performed on the response variables from the IADS mean arousal and mean valence using a range of models, in order to find the one that performed the best in terms of minimising the root-mean-square error (RMSE) and producing the strongest  $R^2$  value. RMSE is the standard error of the regression model where values tending to zero indicate better performance, and is calculated by analysing the outputs from the model  $f(x)$ , for each instance  $i$ , with respect to the known set of outputs  $Y$  as

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - f(x_i))^2}{n}}. \quad (2)$$

Variations were performed using five and tenfold cross-validation (CV) with and without dimension reduction via principal component analysis (PCA), which explained 95% of the variance.

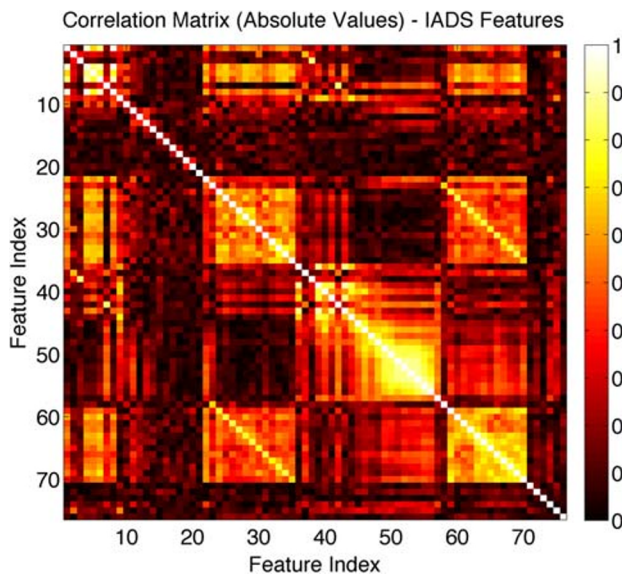
For the ANN experiments, a shallow, two-layer feed-forward ANN was configured, with one hidden layer. The IADS data were divided into segments for the purposes of training (70%), validation (15%) and testing (15%) of the ANNs. The results reported in the next section relate specifically to the performance on the test data subsets. Training used the Levenberg-Marquardt algorithm [36, 41]. Unlike statistical regression, the outcome of training a neural network is a non-deterministic process, ostensibly due to the random allocation of weights to nodes in the network and the potential of the performance metric being used in the training process becoming trapped in local minima [2, 13].

To address this, a systematic approach to the training and evaluation of multiple ANNs was adopted to determine how beneficial ANNs consisting of a different number of neurons, with a differing number of training phases, would be in the prediction of arousal and valence. The approach used involved evaluating ANNs alongside varying the number of neurons in the hidden layer between 1 and 30 and exponentially increasing the number of training iterations at each neuron size, starting at 1 training iteration and finishing at 1000 training iterations. For each training cycle, the resultant best performing network determined in terms of its  $R^2$  metric.

An identical approach was followed as one network was created for the output of emotional arousal and another for valence. Despite the fact that it is possible to produce a single network with two distinct outputs, at this stage it was decided to deal with each dimension separately, as has become common practice in MER [32, 66]. Separation of the affective space also allows the performance of the ANN to be easily examined in terms of each of the aforementioned dimensions.

### 3.2 Results: feature analysis

Prior to applying the machine learning techniques to the IADS features and arousal and valence ratings to create predictive models, an investigation was undertaken on the set of extracted audio features. This analysis was performed to provide insight into any correlations and dependencies between the 76 features and their respective weightings in being able to predict arousal and valence ratings for the IADS contents. To achieve this, Pearson's correlation coefficient was calculated for each pairing of features to form a correlation matrix. Each pairing was then tested for statistical significance and correlation strength to identify similar features. Following this, and by using the audio features as independent variables and the arousal and valence ratings as dependent variables, the predictive power



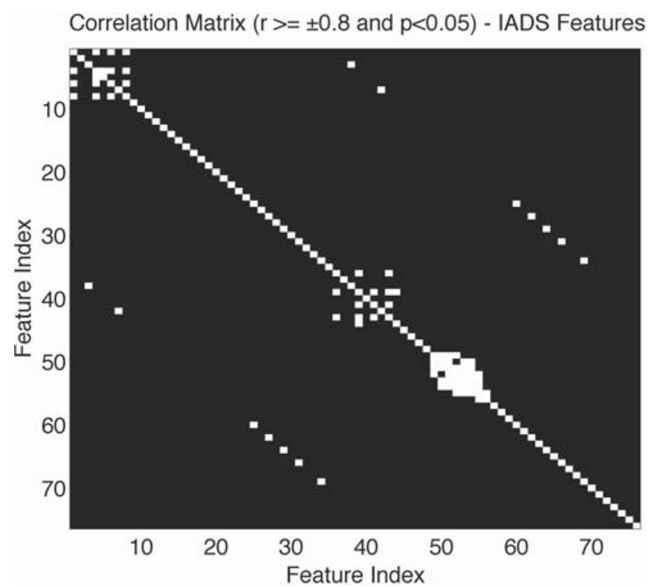
**Fig. 3** Correlation matrix - IADS features

of the feature set, as well as the weighting of individual features, was established using multiple regression analysis.

A correlation matrix was produced for each paired combination of the 76 features extracted from the IADS set, where the metric employed was Pearson's  $r$  product-moment correlation coefficient. The absolute value of Pearson's correlation coefficient  $|r|$  was calculated and used to produce a correlation matrix for the 76 features, as shown in Fig. 3.

Once the inclusion of its identity matrix was accounted for and the duplicate pairs in the upper-half of the matrix are removed, Fig. 3 reveals that a number of features in the IADS set may be correlated and can be visually identified by the light patches. To formally and objectively identify such features, further use can be made of the correlation statistics, where a threshold can be defined. Correlation between features that exceeds this threshold can then be considered as being statistically similar. To achieve this, the  $r$  value and confidence interval  $p$  were interrogated for each pair of features. A value of  $r \geq 0.8$  is accepted as representing a strong or very strong correlation [1] and was chosen as the threshold for the correlation coefficient and a threshold of  $p < 0.05$  was selected for the confidence interval. The same criteria were then applied to the features in the original correlation matrix, resulting in a binary correlation matrix that identifies which features meet the threshold test. A revised correlation matrix is presented in Fig. 4.

Once again, after the identity matrix is discounted and the duplicate matches from the upper-half of the matrix are removed, the significantly correlated features can be identified. A total of 36 unique matching pairs of features were identified representing 47.37% of the feature set.



**Fig. 4** Correlation matrix - strongest ( $r \geq \pm 0.8$  and  $p < 0.05$ ) IADS features

Notably, the features with the the largest number of strong and significant correlations, specifically 3 or more, were

- standard deviation of MFCCs 6 to 12, which correlated with a mixture of one another;
- standard deviation of the spectral rolloff, which correlated with zero-crossing rate standard deviation, spectral centroid standard deviation and spectral entropy standard deviation;
- standard deviation of the spectral centroid, which was correlated with zero-crossing rate standard deviation, spectral entropy standard deviation, spectral rolloff standard deviation, and MFCC 1 standard deviation;
- mean spectral rolloff, which was correlated with zero-crossing rate mean, spectral centroid mean and spectral entropy mean;
- mean spectral entropy, which was correlated with spectral centroid mean, and spectral rolloff mean;
- mean spectral centroid, which was correlated with zero-crossing rate mean, spectral centroid spread mean, spectral entropy mean and spectral rolloff mean;
- mean zero-crossing rate, which was correlated with spectral centroid mean, spectral entropy mean and spectral rolloff mean.

It was notable that those features listed, with the exception of zero-crossing rate are predominantly in the frequency domain and thus relate to spectral components of the audio samples. These findings suggest that it could be possible to reduce the size of the feature set and that, subsequently, techniques for dimension reduction may produce favourable outcomes. To further investigate the role



**Table 2** Regression performance - arousal

| CV      | PCA | Model            | RMSE  | $R^2$ |
|---------|-----|------------------|-------|-------|
| 5-fold  | Yes | Linear           | 1.157 | 0.01  |
| 5-fold  | No  | Squared exp. GPR | 0.989 | 0.28  |
| 10-fold | Yes | Exp. GPR         | 1.179 | -0.02 |
| 10-fold | No  | Squared exp. GPR | 0.998 | 0.27  |

5-fold CV squared exponential Gaussian Process Regression achieves best performance for arousal

that each of the 76 features plays in predicting arousal and valence values, multiple regression analysis was conducted.

For the prediction of arousal, the multiple correlation coefficient showed a strong level of prediction by the features  $R = 0.824$ . Upon analysing the model using an analysis of variance (ANOVA), the IADS feature set statistically significantly predicted arousal,  $F(76, 90) = 2.505$ ,  $p < 0.0005$  outcomes. From examining the coefficients of the regression model, the statistically significant features adding to the prediction of arousal were the mean and standard deviation of MFCC 1, with  $p < 0.05$ . Although not quite statistically significant, the coefficient weights for the mean and standard deviation spectral centroid were notably large.

For the prediction of valence, the multiple correlation coefficient showed a good level of prediction by the features  $R = 0.759$ . Upon analysing the model using an ANOVA, it was found that the IADS feature set statistically significantly predicts valence,  $F(76, 90) =$

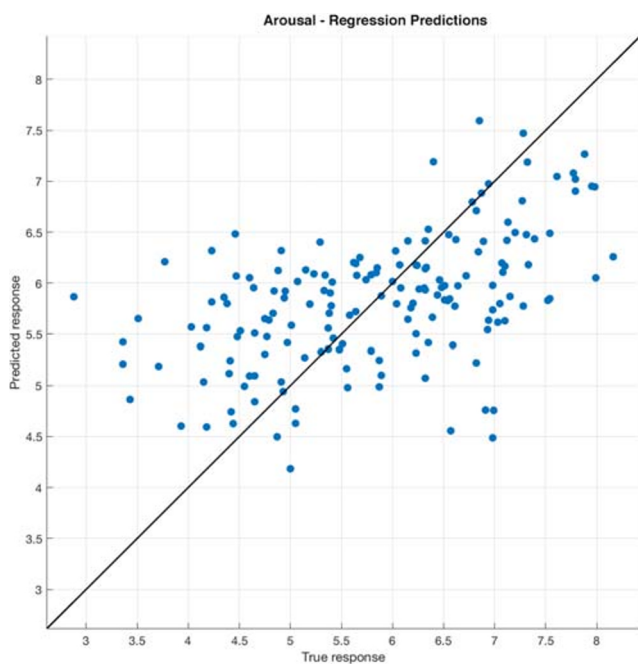
**Table 3** Regression Performance - Valence

| CV      | PCA | Model                  | RMSE  | $R^2$ |
|---------|-----|------------------------|-------|-------|
| 5-fold  | Yes | Stepwise linear        | 1.756 | 0.00  |
| 5-fold  | No  | Rational quadratic GPR | 1.645 | 0.12  |
| 10-fold | Yes | Squared exp. GPR       | 1.746 | 0.02  |
| 10-fold | No  | Matérn 5/2 GPR         | 1.656 | 0.12  |

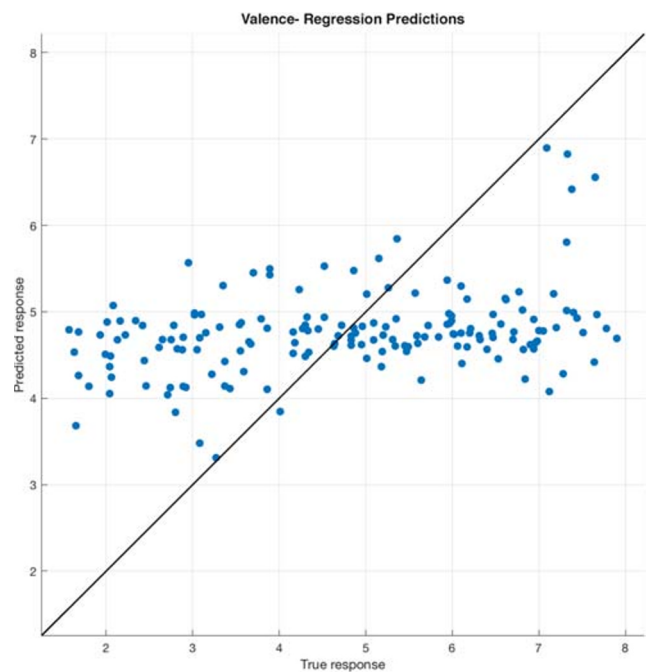
5-fold CV Rational Quadratic GPR achieves best performance for valence

1.613,  $p < 0.02$  outcomes. From examining the coefficients of the regression model, the statistically significant features adding to the prediction of valence were as follows: mean spectral rolloff; mean chroma vectors 6 and 10; standard deviation of MFCCs 1, 2, 8 and 10; and the standard deviation of chroma vector 10,  $p < 0.05$ .

Whilst the correlation matrix revealed that it may have been possible to reduce the number of features slightly, the multiple regression analysis suggests that there are few single features that contribute significantly to the prediction of arousal values in the IADS set. In the case of valence, a limited quality prediction might be possible by reducing the number of features and dimensions of the features. As such, a decision was taken at this stage to use the original 76 features in the machine learning stage and dimension reduction was attempted, but with the expectation that it would be unlikely to yield any major benefit, at least in the case of prediction of arousal.



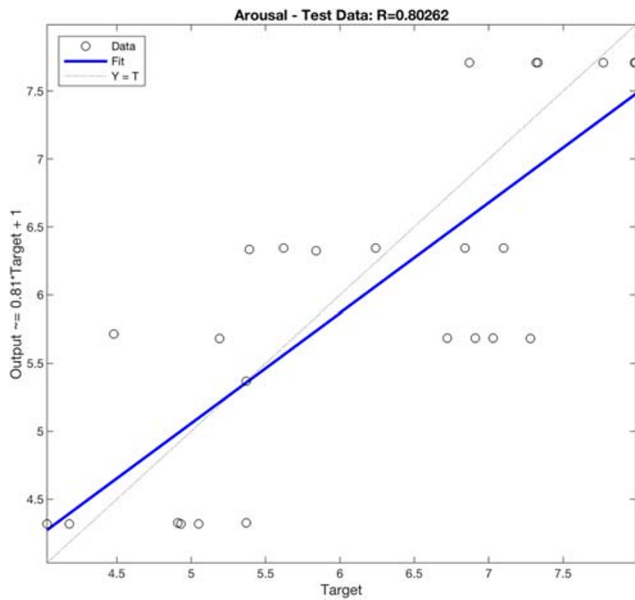
**Fig. 5** 5-fold squared exponential GPR - arousal



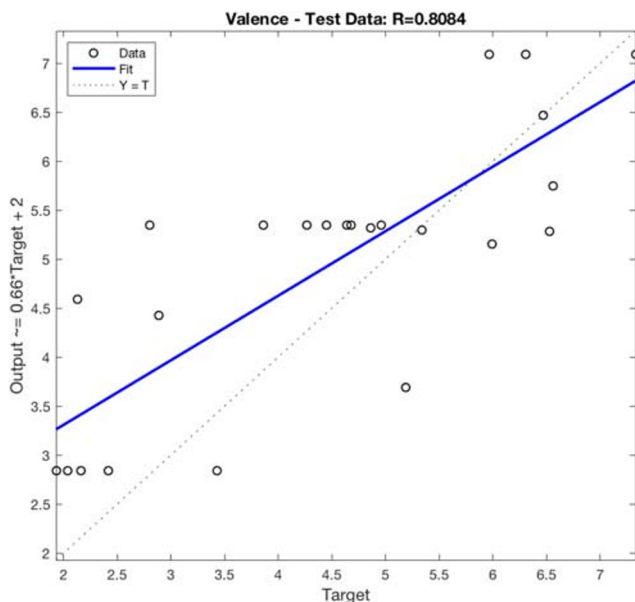
**Fig. 6** 5-fold rational quadratic GPR - valence

**Table 4** Neural network performance - test data

| Dimension | RMSE  | $R^2$ | Neurons |
|-----------|-------|-------|---------|
| Arousal   | 0.748 | 64.42 | 2       |
| Valence   | 0.839 | 65.35 | 2       |



**Fig. 7** ANN - arousal - regression plot for test data



**Fig. 8** ANN - valence - regression plot for test data

### 3.3 Results: regression fitting

The results for prediction of arousal are provided in Table 2. By inspecting the returned RMSE and  $R^2$  values, it can be seen that the 5-fold CV squared exponential Gaussian Process Regression [49] method performs best (RMSE = 0.989,  $R^2 = 0.28$ ), closely followed by the 10-fold CV squared exponential GPR (RMSE = 0.998,  $R^2 = 0.27$ ). The best fit regression model for arousal is shown in Fig. 5.

**Table 5** ANN performance (Arousal) - best outcomes over all neurons and training iterations

| Neurons | Training iterations |       |       |       |
|---------|---------------------|-------|-------|-------|
|         | 1                   | 10    | 100   | 1000  |
| 1       | 0.057               | 0.254 | 0.634 | 0.548 |
| 2       | 0.154               | 0.333 | 0.465 | 0.644 |
| 3       | 0.269               | 0.380 | 0.597 | 0.630 |
| 4       | 0.302               | 0.285 | 0.523 | 0.547 |
| 5       | 0.188               | 0.458 | 0.531 | 0.624 |
| 6       | 0.300               | 0.419 | 0.418 | 0.566 |
| 7       | 0.396               | 0.452 | 0.476 | 0.576 |
| 8       | 0.114               | 0.294 | 0.481 | 0.495 |
| 9       | 0.157               | 0.237 | 0.451 | 0.496 |
| 10      | 0.111               | 0.316 | 0.364 | 0.527 |
| 11      | 0.111               | 0.343 | 0.374 | 0.507 |
| 12      | 0.236               | 0.276 | 0.393 | 0.502 |
| 13      | 0.272               | 0.305 | 0.493 | 0.485 |
| 14      | 0.236               | 0.353 | 0.432 | 0.476 |
| 15      | 0.229               | 0.343 | 0.399 | 0.495 |
| 16      | 0.118               | 0.285 | 0.448 | 0.481 |
| 17      | 0.084               | 0.321 | 0.349 | 0.502 |
| 18      | 0.246               | 0.342 | 0.451 | 0.601 |
| 19      | 0.122               | 0.270 | 0.444 | 0.514 |
| 20      | 0.103               | 0.282 | 0.397 | 0.448 |
| 21      | 0.240               | 0.397 | 0.434 | 0.567 |
| 22      | 0.162               | 0.290 | 0.497 | 0.532 |
| 23      | 0.205               | 0.311 | 0.335 | 0.531 |
| 24      | 0.300               | 0.262 | 0.444 | 0.636 |
| 25      | 0.148               | 0.398 | 0.476 | 0.504 |
| 26      | 0.173               | 0.375 | 0.429 | 0.543 |
| 27      | 0.251               | 0.326 | 0.398 | 0.494 |
| 28      | 0.298               | 0.322 | 0.448 | 0.546 |
| 29      | 0.149               | 0.326 | 0.377 | 0.523 |
| 30      | 0.144               | 0.516 | 0.417 | 0.518 |
| Mean    | 0.196               | 0.336 | 0.446 | 0.535 |
| S.D.    | 0.080               | 0.064 | 0.066 | 0.050 |

Best arousal outcomes achieved at: 1 training iteration and 7 neurons; 10 training iterations and 30 neurons; 100 training iterations and 1 neuron; and 1000 training iterations and 2 neurons

The results for prediction of valence are provided in Table 3. The best performing model was the 5-fold CV Rational Quadratic GPR (RMSE = 1.645,  $R^2 = 0.12$ ), followed by the 10-fold CV Matérn 5/2 GPR (RMSE = 1.656,  $R^2 = 0.12$ ). The best fit regression model for valence is shown in Fig. 6.

Overall, the best performing models were variations on the GPR approach, suggesting that the modelling of arousal and valence using this set of features does not follow a clearly predictable trend. Both arousal and valence regression models tended to predict in the middle of the possible data range, a trend that was best exemplified in the case of valence (Fig. 6), where it can be seen that the bulk of predictions sit between a valence level of 4 and 6. The effect was less pronounced in the case of arousal, which is commensurate with the improved performance in terms of the RMSE and  $R^2$  metrics.

### 3.4 Results: neural network fitting

Due to the use of a small number of neurons in the hidden layer, the training, validation and test processes were fast, each cycle completing within seconds. However, due to the exhaustive search process between 1 and 30 neurons and varying training iterations exponentially in a set of 1, 10, 100 and 1000, the complete experiment duration was in the region of 16 h because of its complexity being polynomial, approximately  $O(n^2)$ . The best obtained values for the metrics of RMSE and  $R^2$  for the test data set are reported in Table 4 with respect to the dimensions of arousal and valence. Graphs representing the performance of the ANN are shown in Fig. 7 for arousal and in Fig. 8 for valence.

To provide a complete picture of the entire set of results obtained with each variation of neuron size and training iterations, Tables 5 and 6 show the combinations therein and the  $R^2$  value obtained by the best performing network at each point. Doing so was useful to obtain a more general view of the performance achieved using the ANN approach and to confirm that the overall best performing results were not numerical outliers obtained by chance during the experiment. Such additional context is supported by Fig. 9, which compares the performance of the best networks produced with the mean values of the experiment.

Performance of arousal and valence prediction, in terms of the best outcomes from the experiment, was found to be extremely similar. Unusually, prediction of valence was slightly more accurate than that of arousal, although analysis of the mean results obtained across all ANN iterations shows that prediction of arousal was generally more accurate than that of valence, which is consistent with the literature on AER and MER. Generally, although the number of samples used in the test data set represents only 15% (25 sounds) of the IADS ratings, there was not

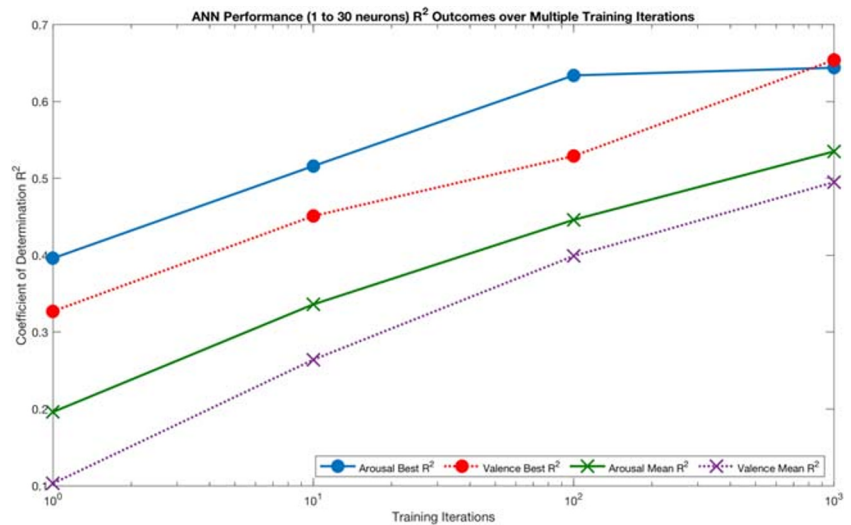
**Table 6** ANN performance (valence) - best outcomes over all neurons and training iterations

| Neurons | Training iterations |       |       |       |
|---------|---------------------|-------|-------|-------|
|         | 1                   | 10    | 100   | 1000  |
| 1       | 0.000               | 0.302 | 0.447 | 0.497 |
| 2       | 0.108               | 0.327 | 0.371 | 0.654 |
| 3       | 0.155               | 0.192 | 0.491 | 0.637 |
| 4       | 0.137               | 0.267 | 0.498 | 0.553 |
| 5       | 0.095               | 0.382 | 0.364 | 0.539 |
| 6       | 0.048               | 0.397 | 0.419 | 0.508 |
| 7       | 0.020               | 0.153 | 0.445 | 0.487 |
| 8       | 0.116               | 0.451 | 0.422 | 0.557 |
| 9       | 0.076               | 0.325 | 0.390 | 0.514 |
| 10      | 0.032               | 0.243 | 0.509 | 0.459 |
| 11      | 0.327               | 0.202 | 0.429 | 0.490 |
| 12      | 0.035               | 0.205 | 0.436 | 0.465 |
| 13      | 0.035               | 0.219 | 0.379 | 0.428 |
| 14      | 0.071               | 0.206 | 0.341 | 0.449 |
| 15      | 0.058               | 0.358 | 0.290 | 0.503 |
| 16      | 0.052               | 0.226 | 0.350 | 0.457 |
| 17      | 0.159               | 0.222 | 0.464 | 0.493 |
| 18      | 0.212               | 0.277 | 0.381 | 0.473 |
| 19      | 0.001               | 0.188 | 0.360 | 0.410 |
| 20      | 0.201               | 0.249 | 0.354 | 0.441 |
| 21      | 0.143               | 0.235 | 0.311 | 0.426 |
| 22      | 0.130               | 0.287 | 0.375 | 0.455 |
| 23      | 0.121               | 0.280 | 0.367 | 0.541 |
| 24      | 0.264               | 0.189 | 0.529 | 0.475 |
| 25      | 0.099               | 0.195 | 0.351 | 0.514 |
| 26      | 0.098               | 0.298 | 0.426 | 0.515 |
| 27      | 0.050               | 0.121 | 0.396 | 0.467 |
| 28      | 0.143               | 0.240 | 0.388 | 0.510 |
| 29      | 0.020               | 0.252 | 0.362 | 0.453 |
| 30      | 0.080               | 0.429 | 0.337 | 0.485 |
| Mean    | 0.103               | 0.264 | 0.399 | 0.495 |
| S.D.    | 0.076               | 0.079 | 0.058 | 0.054 |

Best valence outcomes achieved at: 1 training iteration and 11 neurons; 10 training iterations and 8 neurons; 100 training iterations and 24 neurons; and 1000 training iterations and 2 neurons

the same amount of clustering of predictions in the middle of the range of output variables, as was evidenced when using the regression approach. However, both of the best models exhibit a behaviour of outputting tightly clustered values at certain input ranges. It is expected that this was the result of there being two neurons in both of the best performing models and a consequence of finding a model that was mathematically able to minimise the performance metrics during training and validation, but which does not provide an output that a human would expect to see. This is

**Fig. 9** ANN performance metrics over all neurons and training iterations



a direct consequence of the nature of ANNs and nature of the training and weight allocation process. Nevertheless, it was observed that many other models produced during the experiment produced regression plots that exhibited a more natural distribution for only marginal reductions in their  $R^2$  values.

#### 4 Conclusions and future work

Our results provide an interesting contrast to the values reported in the findings of MER research. The regression model outcomes were poor, compared with MER rates, and largely unsatisfactory for any type of AER task. Regression models have been shown to be effective in the MER field and can compete with other techniques, namely ANNs and other methods [20, 39, 52], but seem to be ineffective in AER. The findings presented here suggest that the recognition of affect in non-musical sounds may require different approaches and alternate or new audio features.

ANN approaches to emotion prediction performed much better in our experiments than regression, showing greater potential, with predictions comparable with work in MER. The ANN models accounted for 64.4% of the variance in the prediction of arousal and for 65.4% in valence. It was found that varying the number of neurons in the networks created did little to improve the  $R^2$  values obtained but that increasing the number of training iterations, thereby reducing the potential of being trapped in a sub-optimal local minima. Granting that the exhaustive process of trying out combinations of neurons and training iterations was time consuming, it is a process that needs only be carried out once before the best model(s) can be stored and reused later, each with a trivial complexity at run-time. The better performance of the ANN, coupled with the relative best fit

regression models using GPR, provides a strong indication that emotion prediction, using this set of features, follows a non-linear model.

Analysis of the features extracted from the IADS set resulted in a mixed set of findings. Despite there being a number of features that were strongly and significantly correlated, predominantly those in the frequency domain, they were not universally applicable in the prediction of affect. In the case of arousal, only the first MFCC showed a statistically significant role in the prediction of arousal, whilst multiple spectral, chroma and MFCC features added significantly to the prediction of valence. However, the subsequent application of PCA prior to regression modelling did not yield better outcomes, suggesting that other methods of feature and dimension reduction could be investigated to improve the efficiency of future AER systems.

In the course of the research presented in this paper, only one other work was found in the literature on AER to which these results can be directly compared, that of Schuller and colleagues [57]. In their research, the prediction of arousal accounted for 37.2% of the variance and prediction of valence achieved 24.0%. These values support the generalisation that arousal is easier to model than valence. The results presented in our work show increased performance over those obtained by Schuller et al. and provides strong evidence for the use of ANN approaches to be employed in subsequent AER research. More broadly, these findings provide evidence that AER does not yet match the levels of performance found in MER, but is beginning to come close. Although our findings represent a significant and incremental improvement and contribution to knowledge in AER, the results presented are limited by a general lack of data sets in the AER field and must be taken in this context.

As shown in Fig. 1, the IADS ratings distribution is not uniform. The majority of arousal ratings, 121 of them, lie in the top two quadrants. As such, modelling and training processes will be biased. The situation is less extreme in the case of valence, with 81 sounds located in quadrants 1 and 4. Nevertheless, at 167 samples, the IADS is small compared with MER, where data sets range from around 30 to over 100,000 songs [43, 54]. Recognition of these limitations of the IADS might be dealt with by creation of a larger set of validated samples with a more uniform distribution.

An avenue for future work would be to take a rigorous and extensive approach in finding optimal parameters that can be used to enhance the predictions made by the neural network. The audio features used are also an area to explore. It may be the case that the best set of features has not yet been considered by any research in the field. Typical audio features are oriented towards signal processing or music information retrieval (MIR) domains and thus may not adequately account for the salient aspects in AER. As an extension to this, another way to train an ANN would be to use the time-series audio sample data as the input.

## Compliance with ethical standards

**Conflict of interests** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Akoglu H (2018) User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* 18(3):91–93
- Atakulreka A, Sutivong D (2007) Avoiding local minima in feedforward neural networks by simultaneous learning. In: Australasian joint conference on artificial intelligence. Springer, pp 100–109
- Austin ML (2016) Chapter 2 - safe and sound: Using audio to communicate comfort, safety, and familiarity in digital media. In: Tettegah SY, Noble SU (eds) *Emotions, technology, and design, emotions and technology*. Academic Press, San Diego, pp 19–35. <https://doi.org/10.1016/B978-0-12-801872-9.00002-8>
- Beauchamp R (2012) *Designing sound for animation*. Routledge, Evanston
- Bordwell D, Thompson K (1985) *Fundamental aesthetics of sound in the cinema. Film sound: theory and practice*, pp 181–199
- Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25(1):49–59
- Bradley MM, Lang PJ (2000) Affective reactions to acoustic stimuli. *Psychophysiology* 37(2):204–215
- Bradley MM, Lang PJ (2007) *The international affective digitized sounds (; iads-2): affective ratings of sounds and instruction manual*. University of Florida, Gainesville, FL, Tech. Rep. B-3
- Burch N (1985) On the structural use of sound. *Film sound: theory and practice*, pp 200–09
- Calleja G (2011) *In-game: from immersion to incorporation*. MIT Press, Cambridge
- Chion M (2019) *Audio-vision: sound on screen*. Columbia University Press, New York
- Choi Y, Lee S, Jung S, Choi IM, Park YK, Kim C (2015) Development of an auditory emotion recognition function using psychoacoustic parameters based on the international affective digitized sounds. *Behavior Research Methods* 47(4):1076–1084
- Choromanska A, Henaff M, Mathieu M, Arous GB, LeCun Y (2015) The loss surfaces of multilayer networks. In: *Artificial intelligence and statistics*, pp 192–204
- Cid F, Manso LJ, Núñez P (2015) A novel multimodal emotion recognition approach for affective human robot interaction. *Proceedings of FinE*, pp 1–9
- Cunningham S, Ridley H, Weinel J, Picking R (2019) Audio emotion recognition using machine learning to support sound design. In: *Proceedings of the 14th international audio mostly conference: a journey in sound on ZZZ, AM'19*. ACM, New York, pp 116–123. <https://doi.org/10.1145/3356590.3356609>
- Cunningham S, Weinel J, Picking R (2018) High-level analysis of audio features for identifying emotional valence in human singing. In: *Proceedings of the audio mostly 2018 on sound in immersion and emotion*. ACM, p 37
- Donaldson LF (2017) Feeling and filmmaking: the design and affect of film sound. *The New Soundtrack* 7(1):31–46
- Drossos K, Floros A, Kanellopoulos NG (2012) Affective acoustic ecology: towards emotionally enhanced sound events. In: *Proceedings of the 7th audio mostly conference: a conference on interaction with sound*. ACM, pp 109–116
- Drossos K, Kotsakis R, Kalliris G, Floros A (2013) Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal. In: *IISA 2013*. IEEE, pp 1–6
- Eerola T, Lartillot O, Toiviainen P (2009) Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: *Ismir*, pp 621–626
- Ehlers A, Clark DM (2000) A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy* 38(4):319–345
- Ekman P (1992) An argument for basic emotions. *Cognition & Emotion* 6(3-4):169–200
- Fadil C, Alvarez R, Martinez C, Goddard J, Rufiner H (2015) Multimodal emotion recognition using deep networks. In: *VI Latin American congress on biomedical engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*. Springer, pp 813–816
- Gabrielsson A, Lindström E (2010) The role of structure in the musical expression of emotions. *Handbook of music and emotion: theory, research applications* 367400
- Gerhard D (2003) *Audio signal classification: history and current techniques*. Citeseer
- Giannakopoulos T, Pikrakis A (2014) *Introduction to audio analysis: a MATLAB® approach*. Academic Press, New York
- Grodal T (2009) *Embodied visions: evolution, emotion, culture, and film*. Oxford University Press, Oxford

28. Haq S, Jackson PJ, Edge J (2008) Audio-visual feature selection and reduction for emotion classification. In: Proc Int Conf on Auditory-Visual Speech Processing (AVSP'08), Tungalooma, Australia
29. Hillman N, Pauletto S (2014) The craftsman: the use of sound design to elicit emotions. *The Soundtrack* 7(1):5–23
30. Hillman N, Pauletto S (2016) Audio imagineering: utilising the four sound areas framework for emotive sound design within contemporary audio post-production. *The New Soundtrack* 6(1):77–107
31. Horowitz S, Looney SR (2014) *The essential guide to game audio: the theory and practice of sound for games*. Routledge, Evanston
32. Hu X, Yang YH (2017) Cross-dataset and cross-cultural music mood prediction: a case on western and chinese pop songs. *IEEE Trans Affect Comput* 8(2):228–240
33. LaBelle B (2010) *Acoustic territories: sound culture and everyday life*. Bloomsbury Publishing, USA
34. Langkjær B (2009) Making fictions sound real-on film sound, perceptual realism and genre. *MedieKultur: Journal of Media and Communication Research* 26(48):13–p
35. Lartillot O, Toiviainen P, Eerola T (2008) A matlab toolbox for music information retrieval. In: *Data analysis, machine learning and applications*. Springer, pp 261–268
36. Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics* 2(2):164–168
37. LoBrutto V (1994) *Sound-on-film: interviews with creators of film sound*. Greenwood Publishing Group
38. Lucas G (1973) *American Graffiti*. Universal Pictures
39. Malheiro R, Panda R, Gomes P, Paiva RP (2016) Emotionally-relevant features for classification and regression of music lyrics. *IEEE Trans Affect Comput* 9(2):240–254
40. Mann M (2006) *Miami Vice*. Universal pictures
41. Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11(2):431–441
42. Miyara F (1999) Acoustic violence: A new name for an old social pain. *Hearing Rehabilitation Quarterly* 24(1):18–21
43. Mo S, Niu J (2017) A novel method based on ompgw method for feature extraction in automatic music mood classification. *IEEE Transactions on Affective Computing*
44. Nardelli M, Valenza G, Greco A, Lanata A, Scilingo EP (2015) Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Trans Affect Comput* 6(4):385–394
45. Noroozi F, Marjanovic M, Njegus A, Escalera S, Anbarjafari G (2017) Audio-visual emotion recognition in video clips. *IEEE Trans Affect Comput* 10(1):60–75
46. Paleari M, Huet B, Chellali R (2010) Towards multimodal emotion recognition: a new approach. In: *Proceedings of the ACM international conference on image and video retrieval*. ACM, pp 174–181
47. Panksepp J (1992) A critical role for “affective neuroscience” in resolving what is basic about basic emotions. *Psychological Review* 99(3)
48. Picard RW (2000) *Affective computing*. MIT Press, Cambridge
49. Rasmussen CE (2003) Gaussian processes in machine learning. In: *Summer school on machine learning*. Springer, pp 63–71
50. Redondo J, Fraga I, Padrón I, Piñeiro A (2008) Affective ratings of sound stimuli. *Behav Res Methods* 40(3):784–790
51. Reynolds S (2013) *Energy flash: a journey through rave music and dance culture*. Faber & Faber
52. Rodà A, Canazza S, De Poli G (2014) Clustering affective qualities of classical music: Beyond the valence-arousal plane. *IEEE Trans Affect Comput* 5(4):364–376
53. Russell JA (1980) A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161
54. Saari P, Fazekas G, Eerola T, Barthet M, Lartillot O, Sandler M (2015) Genre-adaptive semantic computing and audio-based modelling for music mood annotation. *IEEE Trans Affect Comput* 7(2):122–135
55. Schacter D, Tulving E (1994) What are the memory systems of 1994. In: *Memory systems*. MIT Press, pp 341–380
56. Schafer RM (1993) *The soundscape: Our sonic environment and the tuning of the world*. Simon and Schuster, New York
57. Schuller B, Hantke S, Wening F, Han W, Zhang Z, Narayanan S (2012) Automatic recognition of emotion evoked by general sound events. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp 341–344
58. Seng KP, Ang LM, Ooi CS (2016) A combined rule-based & machine learning audio-visual emotion recognition approach. *IEEE Trans Affect Comput* 9(1):3–13
59. Smith J (2015) *THE AUTEUR RENAISSANCE, 1968-1980*. Rutgers University Press, pp 83–106. <http://www.jstor.org/stable/j.ctt16t8zf9.7>
60. Stevenson RA, James TW (2008) Affective auditory stimuli: characterization of the international affective digitized sounds (iads) by discrete emotional categories. *Behavior Research Methods* 40(1):315–321
61. Stone O (1987) *Wall street*. Twentieth Century Fox
62. Sundaram S, Schleicher R (2010) Towards evaluation of example-based audio retrieval system using affective dimensions. In: *2010 IEEE international conference on multimedia and expo*. IEEE, pp 573–577
63. Thayer RE (1990) *The biopsychology of mood and arousal*. Oxford University Press, Oxford
64. Weinel J (2018) *Inner sound: altered states of consciousness in electronic music and audio-visual media*. Oxford University Press, Oxford
65. Wening F, Eyben F, Schuller BW, Mortillaro M, Scherer KR (2013) On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in Psychology* 4:292
66. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 16(2):448–457

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.