Audio Emotion Recognition using Machine Learning to support Sound Design

Stuart Cunningham
Manchester Metropolitan University
Manchester, UK
s.cunningham@mmu.ac.uk

Harrison Ridley
Manchester Metropolitan University
Manchester, UK
HARRISON.RIDLEY@stu.mmu.ac.uk

Jonathan Weinel
Coventry University
Coventry, UK
jonweinel@gmail.com

Richard Picking
Wrexham Glyndŵr University
Wrexham, UK
r.picking@glyndwr.ac.uk

ABSTRACT

In recent years, the field of Music Emotion Recognition has become established. 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. Audio features are extracted and used as the input to two machine-learning approaches: regression modelling and artificial neural networks, in order to predict the emotional dimensions of arousal and valence.

It is found that shallow neural networks perform better than a range of regression models. Consistent with existing research in emotion recognition, prediction of arousal is more reliable than that of valence. Several extensions of this research are discussed, including work related to improving data sets as well as the modelling processes.

CCS CONCEPTS

• **Applied computing → Sound and music computing**; Media arts; • **Human-centered computing** → Interaction design theory, concepts and paradigms; Empirical studies in interaction design.

AM '19, September 18–20, 2019, Nottingham, UK
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/3356590.3356609

KEYWORDS

Affect, arousal, audio emotion recognition, audio features, emotion, IADS, regression, neural networks, valence

ACM Reference Format:
Stuart Cunningham, Harrison Ridley, Jonathan Weinel, and Richard Picking. 2019. Audio Emotion Recognition using Machine Learning to support Sound Design. In *Audio Mostly 2019: A Journey in Sound, September 18–20, 2019, Nottingham, UK.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3356590.3356609

1 INTRODUCTION

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.

Affective Computing and Audio

Affective computing is a growing and interdisciplinary research field concerned with the emotional interaction between technology and humans [30]. 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 [8, 18, 28, 33, 35]. In this article, however, we turn our focus to the area of Audio Emotion Recognition (AER), which deals with affect in non-musical sound. This 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 [7]. For the purposes of this article, we define sound effects as including the many layers of audio, other than music, that are found in media, such as ambience, Foley, dialogue, and human utterances.

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, whilst dimensional models use quantitative values on one or more dimensions. An example of a categorical model can be seen in the work of Ekman [13] or Panksepp [29], whilst dimensional models may be seen in those of Thayer [42] or Russell [34].

The research documented in this article adopts the use of the latter of these: Russell's circumplex model of affect, which is a two-dimensional Cartesian emotion space consisting of axes relating to arousal (vertical) and valence (horizontal) [34]. This approach is typical in the field of emotion recognition. Whilst 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 [18, 33, 35].

Models for the prediction of emotion in media make use of the coefficient of determination R^2 as a performance metric. In the field of MER, the upper range of R^2 values for arousal are approximately 80% to 85% and approximately 60% to 70% for valence [12, 18, 24]. An aim of our work is to determine if similar levels of performance can be achieved in AER.

The IADS Data Set

The International Affective Digitized Sound (IADS) system [4] is a corpus of validated, emotionally annotated sounds. This data 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 [2]. Ratings are presented for each dimension using a 9-point scale and each sound is rated by a minimum of 100 participants. The mean duration of the 167 sound samples in the IADS set is 6.014 seconds (σ= 0.017 seconds). The overall distribution of the IADS ratings in arousal and valence space is illustrated in Figure 1. By extracting data relating to the arousal and valence dimensions therein, we make use of the IADS in our attempt to create computational models of emotional response to sound.

Affective Audio

While Picard (1997) [30] 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 such as games, audio-visual art and film, which increasingly are embedded within computer systems. For instance, Weinel's work [43] 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* [34]). For instance, Gabrielsson and Lindstrom's [14] meta-study of music and emotion reveals musical features such as rhythm, melody, pitch and tonality may often be associated with specific affective responses. Weinel argues that these features are primarily involved in the production of affect, while 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.

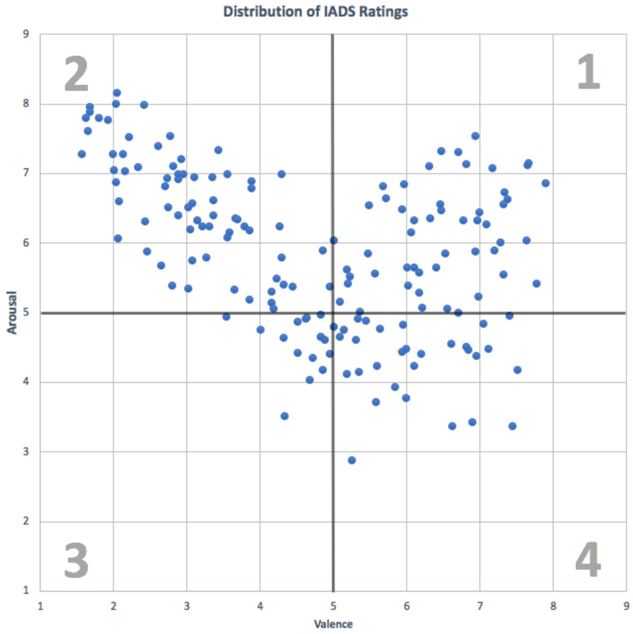

Figure 1: IADS Mean Values in Arousal and Valence Space

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 (such as through soundscapes), 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 [37] discussion, we may consider how sounds such as alarms or dogs barking have representational or symbolic meaning, but also give rise to emotional responses. These emotional responses may include aspects that are culturally shared, and those that are highly individualised and subjective. For example, sound such as 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. Such sounds can be understood as cross-cultural, relating to shared cultural knowledge and *semantic memory* (in terms of Schacter and Tulving's [36] theory of semantic and episodic memory). Yet alarms can also trigger highly individualised responses, such as for persons with Post-Traumatic Stress Syndrome (PTSD), for whom the sound may trigger traumatic autobiographical *episodic memories.*

Affective Sound Design for Film and Other Media

It has been suggested [1] that sound design can *"actively shape how we perceive the image"*. This shaping of how an audience's experience through sound can be performed through the elicitation of emotions, among 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"* [9], whilst also describing the fictional world, giving it a *"particular toning"* [20]. 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 [1, 5].

It is hypothesised that by using affective audio in cinema, a filmmaker may be able to ensure their audience feels a particular emotion at specific points throughout the film. Some research into how this can be implemented has already been undertaken. Most notable is the work of Hillman and Pauletto [16, 17], 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"*.

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 aren't accustomed to the sounds [19, 37]. Further, the notion of *"Acoustic Violence"* coined by Miyara [27] 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 [37] — 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 techniques such as this may be used as plot devices.

Successful sound design does not only access the representational aspects of sound, but 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* [40], 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* [25], 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 five minutes 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* [23]. 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 [22]. All of this adds to the party atmosphere of the film, and may subtlety elicit euphoric and sometimes nostalgic feelings in the audience, as it never over emphasises anything.

In video games, sound design serves similar purposes in an interactive context. The diegetic soundscape serves to make spatial environments convincing navigable spaces, leading to presence and immersion through what Cajella [6] 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.

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 more emotionally impactful work 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.

Audio Affect Identification

In noting that the presence of film when studying affective responses makes it difficult to isolate the sonic aspects, work by Bradley and Lang [3] 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 whilst listening to the sounds based on arousal, valence and dominance on 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.* [32] 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 while Spaniards rated sounds in a very similar way to Americans, there was some — if only minor — differences. It found that Americans tend to rate sounds with more positive valance than Spaniards, whilst less activating 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 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 [39] 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 whilst the results obtained were not entirely useful for predicting responses, the categorisation of sounds it produced may be beneficial to future research.

Audio Emotion Recognition

Sundaram and Schleicher [41] conducted experiments modelling the affective response of listeners to a range of sound recordings. What makes their work novel is 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 is based upon the difficulties associated with using categorical approaches for sounds with multiple acoustic sources, but is 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 [2]. 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.* [10] 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 Figure 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. This finding must be considered in the context of a theoretical 25% allocation by chance, further skewed by the distribution of the sounds, as evidenced in Figure 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.* [11] 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 is 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.* [38] 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, among others. 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 then annotated by a small group of four participants, which is 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.

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.

Analysis Method

All 167 sounds from IADS were employed within the analysis. The sounds were peak-normalised to control loudness. This replicates the conditions reported by the originators of IADS in their participant study [4]. Audio features were extracted using the *Matlab 2018a* software and the *Matlab Audio Analysis Library* [15], using the settings of a 50 ms window with a 50% overlap. These values have been shown effective in other works relating to emotion analysis [8]. The set of 35 features from the *Audio Analysis Library* were extracted, which comprise of: zero-crossing rate; energy; energy entropy; spectral centroid (mean); spectral centroid (spread); spectral entropy; spectral flux; spectral roll-off; 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 [11]. These following features were obtained by making use of the *MIRToolbox* [21] 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. Variations were performed using five and ten fold Cross-Validation (CV) with, and without, dimension reduction via Principal Component Analysis (PCA), which explained 95% of the variance.

A shallow, two-layer feed-forward ANN was configured, with one hidden layer. Some brief trial and error gave indication that a network with eight neurons provided fair outcomes in terms of computational performance and RMSE and R^2 metrics. One network was created for the output of emotional arousal and another for valence. Whilst 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. This also allows the performance of the ANN to be examined easily in terms of each of the aforementioned dimensions. Extensive training and tuning of parameters was not undertaken at this time since the intention was to investigate the general feasibility of the ANN approach and not to find optimal values, which can be a time-consuming process. Instead, a brief period of trial-and-error training took place manually, consisting of no more than ten or fifteen training iterations.

The IADS data were divided 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 [26].

Results: Regression Fitting

The results for prediction of arousal are provided in Table 1. By inspecting the returned RMSE and R^2 values, it can be seen that the 5 fold CV Squared Exponential Gaussian Process Regression [31] 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 Figure 2.

Table 1: 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

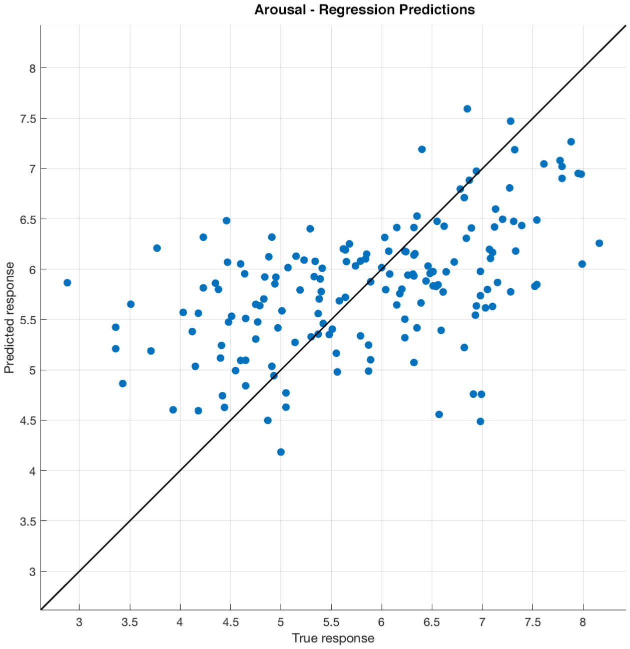

Figure 2: 5 Fold Squared Exponential GPR - Arousal

The results for prediction of valence are provided in Table 2. 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 Figure 3.

The best performing models were variations on GPR, 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 output range, a trend exemplified in the case of valence (Figure 3), where it can be seen that the bulk of predictions sit between a value of 4 and 6. The effect is less pronounced for arousal, commensurate with its better performance in terms of the RMSE and R^2 metrics.

Table 2: 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

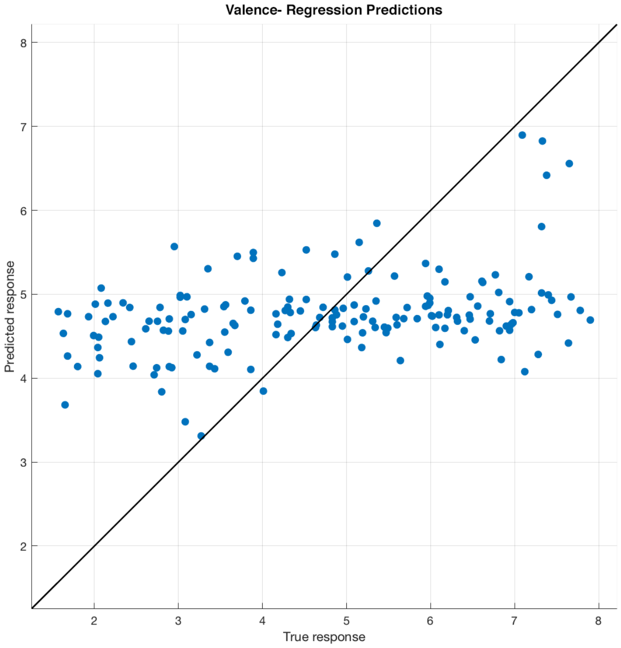

Figure 3: 5 Fold Rational Quadratic GPR - Valence

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, completing within seconds. The best obtained values for the metrics of RMSE and R^2 for the test data set are reported in Table 3 with respect to the dimensions of arousal and valence. Graphs representing the performance of the ANN are shown in Figure 4 for arousal and in Figure 5 for valence.

Consistent with the literature on AER and MER, prediction of arousal is better than that of valence using the ANN. Although the number of samples used in the test data set represents 15% (25 sounds) from the IADS ratings, there is less clustering of predictions in the middle of the range of output variables, as observed when using regression.

Table 3: Neural Network Performance - Test Data

Dimension	RMSE	R^2
Arousal	0.987	0.345
Valence	0.514	0.269

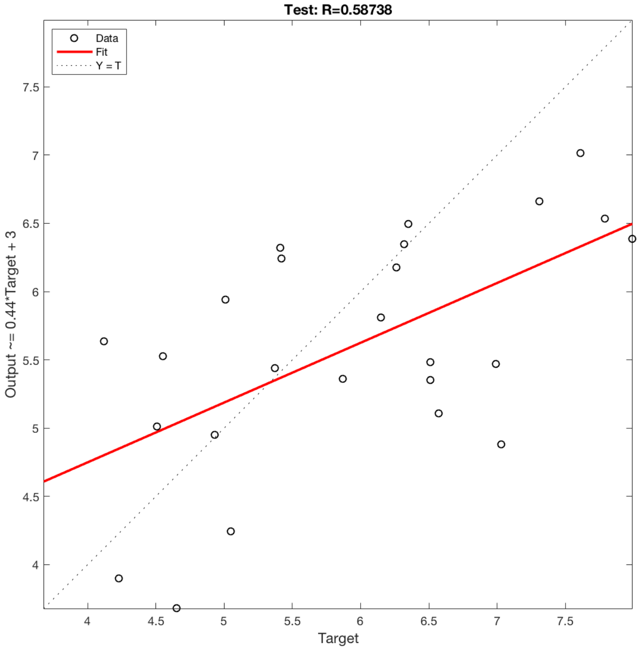

Figure 4: ANN - Arousal - Regression Plot for Test Data

4 CONCLUSIONS AND FUTURE WORK

Our results provide an interesting contrast to the values reported in the findings of MER research, suggesting 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 better than regression. The ANN models accounted for 34.5% of the variance in the prediction of arousal and for 26.9% in valence. Only a short amount of time was spent experimenting with the parameters of the ANN, which is a limitation of its performance. The better performance of the ANN, coupled with the best-fit regression models using GPR, indicates that emotion prediction, using these features, is non-linear.

There is only one other work in AER to which these results can be directly compared, where prediction of arousal accounted for 37.2% of the variance and prediction of valence achieved 24.0% [38]. These values support the generalisation that arousal is easier to model than valence. Given the larger number of annotators and sounds samples in IADS it may be reasonable to postulate that differences between our results and those of Schuller *et al.* represent statistical noise. These findings show that AER does not yet match the levels of performance found in MER. However, the results presented are limited by a lack of data sets in AER.

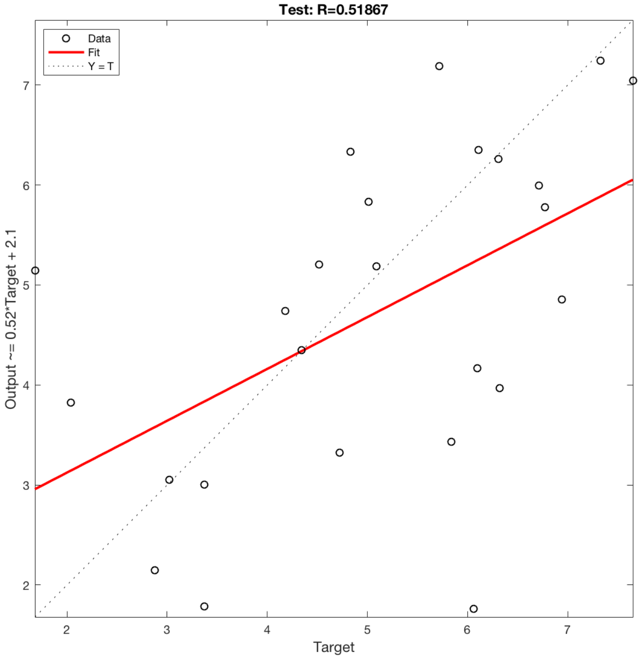

Figure 5: ANN - Valence - Regression Plot for Test Data

As shown in Figure 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 to MER data sets, which can range from 30 to over 100,000 songs [28, 35]. 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 musical domains and thus may not 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.

REFERENCES

[1] David Bordwell and Kristin Thompson. 1985. Fundamental aesthetics of sound in the cinema. *Film sound: Theory and practice* (1985), 181–199.

[2] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[3] Margaret M Bradley and Peter J Lang. 2000. Affective reactions to acoustic stimuli. *Psychophysiology* 37, 2 (2000), 204–215. https://doi.org/10.1111/1469-8986.3720204

[4] Margaret M Bradley and Peter J Lang. 2007. The International Affective Digitized Sounds (; IADS-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3* (2007).

[5] Noël Burch. 1985. On the structural use of sound. *Film sound: theory and practice* (1985), 200–09.

[6] Gordon Calleja. 2011. *In-game: From immersion to incorporation.* MIT Press.

[7] Michel Chion. 2019. *Audio-vision: sound on screen.* Columbia University Press.

[8] Stuart Cunningham, Jonathan Weinel, and Richard Picking. 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, 37. https://doi.org/10.1145/3243274.3243313

[9] Lucy Fife Donaldson. 2017. Feeling and filmmaking: the design and affect of film sound. *The New Soundtrack* 7, 1 (2017), 31–46. https://doi.org/10.3366/sound.2017.0095

[10] Konstantinos Drossos, Andreas Floros, and Nikolaos-Grigorios Kanellopoulos. 2012. Affective acoustic ecology: Towards emotionally enhanced sound events. In *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound.* ACM, 109–116. https://doi.org/10.1145/2371456.2371474

[11] Konstantinos Drossos, Rigas Kotsakis, George Kalliris, and Andreas Floros. 2013. Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal. In *IISA 2013.* IEEE, 1–6. https://doi.org/10.1109/IISA.2013.6623709

[12] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. 2009. Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *Ismir.* 621–626.

[13] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200. https://doi.org/10.1080/02699939208411068

[14] Alf Gabrielsson and Erik Lindström. 2010. The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications* 367400 (2010).

[15] Theodoros Giannakopoulos and Aggelos Pikrakis. 2014. *Introduction to audio analysis: a MATLAB® approach.* Academic Press.

[16] Neil Hillman and Sandra Pauletto. 2014. The Craftsman: The use of sound design to elicit emotions. *The Soundtrack* 7, 1 (2014), 5–23. https://doi.org/10.1386/st.7.1.5_1

[17] Neil Hillman and Sandra Pauletto. 2016. Audio Imagineering: Utilising the Four Sound Areas Framework for Emotive Sound Design within Contemporary Audio Post-production. *The New Soundtrack* 6, 1 (2016), 77–107. https://doi.org/10.3366/sound.2016.0084

[18] Xiao Hu and Yi-Hsuan Yang. 2017. Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop songs. *IEEE Transactions on Affective Computing* 8, 2 (2017), 228–240.

[19] Brandon LaBelle. 2010. *Acoustic territories: Sound culture and everyday life.* Bloomsbury Publishing USA.

[20] Birger Langkjær. 2009. Making fictions sound real-On film sound, perceptual realism and genre. *MedieKultur: Journal of media and communication research* 26, 48 (2009), 13–p. https://doi.org/10.7146/mediekultur.v26i48.2115

[21] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. 2008. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications.* Springer, 261–268.

[22] Vincent LoBrutto. 1994. *Sound-on-film: Interviews with creators of film sound.* Greenwood Publishing Group.

[23] George Lucas. 1973. *American Graffiti.* Universal Pictures.

[24] Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. 2016. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing* 9, 2 (2016), 240–254. https://doi.org/10.1109/TAFFC.2016.2598569

[25] Michael Mann. 2006. *Miami Vice.* Universal Pictures.

[26] Donald W Marquardt. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* 11, 2 (1963), 431–441. https://doi.org/10.1137/0111030

[27] Federico Miyara. 1999. Acoustic Violence: A New Name for an Old Social Pain. *Hearing Rehabilitation Quarterly* 24, 1 (1999), 18–21.

[28] Shasha Mo and Jianwei Niu. 2017. A Novel Method based on OMPGW Method for Feature Extraction in Automatic Music Mood Classification. *IEEE Transactions on Affective Computing* (2017). https://doi.org/10.1109/TAFFC.2016.2523503

[29] Jaak Panksepp. 1992. A critical role for" affective neuroscience" in resolving what is basic about basic emotions. (1992). https://doi.org/10.1037/0033-295X.99.3.554

[30] Rosalind W Picard. 2000. *Affective computing.* MIT press.

[31] Carl Edward Rasmussen. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning.* Springer, 63–71.

[32] Jaime Redondo, Isabel Fraga, Isabel Padrón, and Ana Piñeiro. 2008. Affective ratings of sound stimuli. *Behavior Research Methods* 40, 3 (2008), 784–790. https://doi.org/10.3758/BRM.40.3.784

[33] Antonio Rodà, Sergio Canazza, and Giovanni De Poli. 2014. Clustering affective qualities of classical music: Beyond the valence-arousal plane. *IEEE Transactions on Affective Computing* 5, 4 (2014), 364–376. https://doi.org/10.1109/TAFFC.2014.2343222

[34] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161. https://doi.org/10.1037/h0077714

[35] Pasi Saari, György Fazekas, Tuomas Eerola, Mathieu Barthet, Olivier Lartillot, and Mark Sandler. 2015. Genre-adaptive semantic computing and audio-based modelling for music mood annotation. *IEEE Transactions on Affective Computing* 7, 2 (2015), 122–135. https://doi.org/10.1109/TAFFC.2015.2462841

[36] D. Schacter and E. Tulving. 1994. Whater Are the Memory Systems of 1994. In *Memory Systems.* MIT Press, 341–380.

[37] R Murray Schafer. 1993. *The soundscape: Our sonic environment and the tuning of the world.* Simon and Schuster.

[38] Björn Schuller, Simone Hantke, Felix Weninger, Wenjing Han, Zixing Zhang, and Shrikanth Narayanan. 2012. Automatic recognition of emotion evoked by general sound events. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 341–344. https://doi.org/10.1109/ICASSP.2012.6287886

[39] Ryan A Stevenson and Thomas W James. 2008. Affective auditory stimuli: Characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories. *Behavior research methods* 40, 1 (2008), 315–321. https://doi.org/10.3758/BRM.40.1.315

[40] Oliver Stone. 1987. *Wall Street.* Twentieth Century Fox.

[41] Shiva Sundaram and Robert Schleicher. 2010. Towards evaluation of example-based audio retrieval system using affective dimensions. In *2010 IEEE International Conference on Multimedia and Expo.* IEEE, 573–577. https://doi.org/10.1109/ICME.2010.5583001

[42] Robert E Thayer. 1990. *The biopsychology of mood and arousal.* Oxford University Press.

[43] Jonathan Weinel. 2018. *Inner Sound: Altered States of Consciousness in Electronic Music and Audio-visual Media.* Oxford University Press.