Modelling Affect for Horror Soundscapes

Phil Lopes, Antonios Liapis *Member, IEEE* and Georgios N. Yannakakis *Senior Member, IEEE*

Abstract—The feeling of horror within movies or games relies on the audience's perception of a tense atmosphere — often achieved through sound accompanied by the on-screen drama — guiding its emotional experience throughout the scene or game-play sequence. These progressions are often crafted through an a priori knowledge of how a scene or game-play sequence will playout, and the intended emotional patterns a game director wants to transmit. The appropriate design of sound becomes even more challenging once the scenery and the general context is autonomously generated by an algorithm. Towards realizing sound-based affective interaction in games this paper explores the creation of computational models capable of ranking short audio pieces based on crowdsourced annotations of tension, arousal and valence. Affect models are trained via preference learning on over a thousand annotations with the use of support vector machines, whose inputs are low-level features extracted from the audio assets of a comprehensive sound library. The models constructed in this work are able to predict the tension, arousal and valence elicited by sound, respectively, with an accuracy of approximately 65%, 66% and 72%.

Index Terms—Horror, Sonification, Tension, Crowdsourcing, Preference Learning, Rank annotations

1 Introduction

Audio is often associated with classical or contemporary musical pieces. The reality however is that audio can be more than just "music", but a meticulous crafted sonority that complements visual and interactive experiences, often described as audiovisual metaphors [1]. Sound design is an important part of both film [1], [2] and digital games [3], [4], [5], where sound designers fine tune the intended emotional experience, through expert knowledge, to the exact imagery on-screen. In digital games this process is harder, as sounds must accommodate player interactivity, and virtual environments that vary between different visual styles along the course of an entire game [3], [6]. The task of sound design can become even more challenging when games procedurally generate these virtual environments as the layouts — and potentially even the visuals — are generated in real-time. This paper investigates the construction of several data-driven models capable of ranking the perceived emotion of horror sounds across three affective dimensions: tension, arousal and valence. Such models may offer an additional layer of sound autonomy for procedural content generation systems, allowing them to more closely recreate the emotional progression that audio designers construct. Motivated by the lack of such a model for game sound design this paper introduces a crowdsourcing methodology for deriving the computational mapping between sounds within the horror genre and their perceived affect.

Models as the ones constructed in this paper can also be applicable for tools that aid the development process. Due to the increasing complexity of developing contemporary digital games, several development tools such as *Unity* (Unity Technologies, 2005) and the *Unreal Engine* (Epic Games, 1998) have been used to aid the creation of content and reduce development costs. Although academic work in game technology tends to focus on level design [7], these tools are not exclusive to level designers, and include several features for sound designers, 3D modellers, animators and writers. Complementary, AI-assisted game design systems such as *Sentient Sketchbook* explore how tools can pro-actively help designers, by offering suggestions and detailed information of several important level design parameters [8]. We argue that a system similar to *Sentient Sketchbook* can be realised for sound design, through the development of automated systems capable of suggesting specific audio assets based on the developers' thematic intent of a level's soundscape. This study offers the first operational step towards achieving such a goal.

The field of music emotion recognition has often concentrated on the detection of emotions within contemporary and classical musical pieces [9], [10]. This paper argues that these models can also be used on sounds with the intent of accompanying audiovisual experiences, especially when considering that the objective of audio within both the film and digital game media is to purposefully stimulate certain player emotions [11]. We construct several preference learned models, using the rank support vector machine algorithm [12], to predict the global rank of horror sounds across the three affective dimensions of the Schimmack and Grob model [13]: tension, arousal and valence. Although previous work has explored the creation of preference models that rank emotion in audio [14], to the authors' best knowledge such a model has never been constructed for sound intended to accompany audiovisual horror experiences.

Human preference annotations were gathered using a crowdsourcing platform, allowing participants to rank pairs of sounds on the perceived tension, arousal and valence. We assume there is an underlying function between low-level descriptors extracted from each sound in the audio library and the perceived emotions annotated by human participants that a preference learning mechanism may derive. This paper presents several models capable of predicting a global rank of elicited tension, arousal and valence, respectively, with a 65% and 66%, 72% average accuracy via 5-fold cross-validation.

All authors are with the Institute of Digital Games, University of Malta (e-mail: {louis.p.lopes; antonios.liapis; georgios.yannakakis}@um.edu.mt

This paper also introduces the first attempt of deriving a mapping between sound effects and perceived affect directly giving insight to the important step of *studio manipulation* [5] for sound design. Studio manipulation is a way of creating novel audio pieces by combining different audio signals or altering the audio through signal processing effects. Multiple types of audio effects exist such as reverb or echo, that given certain parameter values can significantly alter the perceived sound of the original audio file[1]. We, thus, argue that a computational mapping between sound effects and emotional manifestations would provide AI-assisted and affect-driven systems the ability to suggest certain effects that the designers might want to use, or even allow automated systems to extend their internal audio library through the usage of audio manipulation effects. This study presents models capable of ranking the impact of an audio effect on a sound in terms of elicited tension, arousal and valence, with a 5-fold cross-validation accuracy of 72%, 70% and 65%, respectively.

The paper is structured as follows. An overview of related work is presented in Section 2, followed by a detailed description of the experimental methodology in Section 3. The performance obtained from the different models trained is presented in Section 4, followed by a detailed discussion in Section 5. The paper concludes with Section 6.

2 Background

This section gives a brief introduction to the notion of audiovisual metaphors, and a review of the related work in both modelling sound-elicited affect and preference learning.

2.1 Affect and Audiovisual Metaphors

Beyond music, audio has often been used as an accompaniment of the on-screen imagery of film and digital games. Described as audiovisual metaphors [1], this technique is often used to emphasize certain emotions of characters or scenes towards the audience. This work specifically explores the creation of a model capable of ranking the perceived affect of audio, intended for the creation of audiovisual metaphors.

Fahlenbrach [1] describes audiovisual metaphors as shared emotional and physical characteristics of the on-screen pictures and sounds, that once effectively merged are capable of conveying powerful emotions within the audience. Perceived meaning of audiovisual metaphors relate to an individual's personal emotional experience. Personal factors include cultural and social background (e.g. symbolism and its meaning both in terms of audio and imagery), personal association towards the on-screen drama (i.e. associative emotion such as sorrow or fear), and even stimulus-response-patterns derived from both sound and imagery. Fahlenbrach exemplifies how audio is effectively used in the Stanley Kubrick film "The Shining", in the popular staircase scene, where the conjunction of the careful editing of the on-screen imagery and the chaotic dissonance of the sound convey a sense of dread and tension. This is a popular approach of treating sounds within the horror genre (whether that is a movie or a video game), where both the absence of sound and the use of short uncomfortable audio cues are consistently interwoven for the creation of tense and frightening experiences [11].

This paper explores the creation of a system capable of ranking short musical pieces based on how tense, arousing and pleasurable participants perceive them. Such a system may provide recommendations to sound designers for their personal sound libraries — e.g. by suggesting different audio files depending on the game context. It can also offer automated systems an approach for sonifying virtual game worlds, which can follow designer defined emotional patterns [15].

2.2 Modelling the Affect of Sound

Modeling affect in the domain of music and sound has traditionally divided studies with respect to their annotation approach. While several researchers often study emotion representation through discrete models [16], [17], alternatively others have argued that dimensional approaches to emotion representation are superior [18], [19], [20].

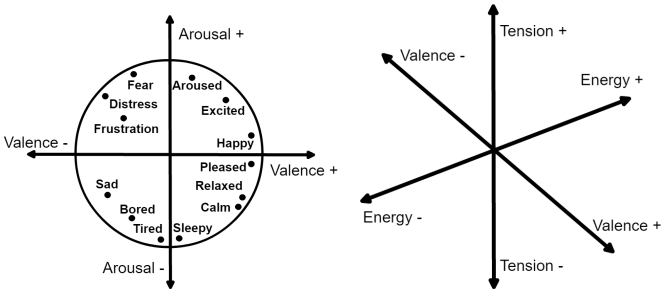
According to discrete models, all emotions can be derived from a limited set of universal emotions, such as fear, anger, disgust, sadness and happiness [16], [20], where each emotional state is considered independent from any other. Within the context of music, discrete models have been altered to better represent emotions expressed by music, such as disgust which rarely is perceived musically and thus has been replaced with tenderness [21], [22]. The Geneva Emotion Music Scale (GEMS) has been used as an alternative discrete model for representing affect in music; the model classifies emotion into nine categories [17]: wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness. According to [18], however, there is evidence for the superiority of dimensional models over discrete models for affect modelling in music.

Emotion is often represented across dimensions in a continuous space. Arguably the most popular model of that type is the Russell circumplex model [23], where emotions are represented as two dimensional planes (see Fig. 1a): arousal (activation-deactivation) and valence (pleasure-displeasure). Alternatively, the Thayer model [24] proposed a variant to the Russell circumplex, and argues that both dimensions are actually "tense arousal" (Arousal) and "energetic tension" (Valence). Schimmack and Grob [13] present an alternative study based on a 3 dimensional model of affect containing two dimensions for valence and arousal, with an additional dimension for tension (see Fig. 1).

Due to the importance of tension within the horror genre and our emphasis on tension-based game adaptation, we study sounds based on annotations across the three dimensions of the Schimmack and Grob model [13]. This allows for each audio asset to be annotated on the dimension of tension, while still leaving the possibility open to study the valence and arousal dimensions.

Emotion recognition in audio is an active field of research [9], [10], [25]; however, the focus of these studies is usually on musical audio pieces and not on audio that is intended for audiovisual accompaniment. Although previous work has used film soundscapes as a way of comparing

1. Examples of audio effects: https://goo.gl/kfHP7Y



(a) Russell's Circumplex Model (b) Schimmack and Grob Model

Fig. 1: Russell's circumplex (Fig. 1a) is a two dimensional model consisting of valence and arousal, each ranging from a negative to a positive value of affect. Alternatively, the model of Schimmack and Grob (Fig. 1b) is a three dimensional model, consisting of valence, tension and energy (or arousal), which also range from negative to positive values of affect. For example in the Schimmack and Grob model fear can be considered a high energy, high tension and low valence emotion; while excitement a high energy, low tension and high valence emotion.

emotional models [18] or investigating the variations of affect across multiple genres [26], it has rarely been a main focus within literature. It is also worth mentioning that most work within music emotion recognition tends to focus on the Russell model specifically [10], [14]. This work, instead, offers a new perspective by both exploring the affective space of the sound domain and by investigating an additional dimension (tension) as described in the Schimmack and Grob model.

2.3 Ranking-based Annotation and Crowdsourcing

A number of studies in the fields of affective computing and human computer interaction already suggest that rank-based surveys is a far more accurate representation of an annotator's subjective assessment [27], [28], [29], when compared to rating-based (e.g. Likert scale [30]) questionnaires. Instead of quantifying individual items based on a scale of variable length, rank-based annotation asks participants to compare between a set of different items and rank them according to a variable of a studied phenomenon. Ranking eliminates the amount of subjectivity and variant interpersonal biases caused by a number of factors such as arbitrary scale perception effects, order effects, scale inconsistency effects, and social and cultural preconceptions that emerge from the use of ratings [27], [29]. Crowdsourcing is a powerful tool for acquiring significant amounts of user annotated data which has been used in a number of research domains for soliciting subjective notions such as the appeal of a narrative [31] or the annotation of a subjective experience such as game aesthetics [32].

This work employs a rank-based crowdsourcing approach with the aim of soliciting human pairwise ranks between sound samples within the sound library [15]. Annotations acquired from crowdsourcing will train data-driven computational models capable of predicting global ranks of tension, arousal and valence specifically for the horror genre.

2.4 Preference Learning for Affect Modelling

Preference Learning (PL) is a supervised learning methodology, where the goal is to derive a global ranking function from a set of annotated ranks [33]. PL for affective modelling was introduced by Yannakakis [34] and has since then been used extensively within the domain of affective interaction, for e.g. personalizing game levels [32] and for affect-driven camera control [35]. Rank Support Vector Machines (RankSVM), a variant of SVMs, was introduced by Joachims [12] as a way of ranking webpages based on their click rate. A RankSVM consists of projecting pairwise data onto a feature space combined with ranked annotations, adjusting a weight vector ($\vec{w}$) so that all points in the training dataset are ordered by their projection onto $\vec{w}$. Although RankSVMs started as a way of optimizing webpage queries, it has been applied to several other domains quite successfully such as for the detection of emotion in speech [36] and musical pieces [37].

Specifically in audio, Yang et al. [14] used preference learning for music emotion recognition. RankSVMs were used to rank different musical pieces — represented with Russell's circumplex model of affect [23] —- based on low-level audio descriptors commonly extracted in music information retrieval. Inspired by the success of RankSVM affect models in music, in this study we train a number of RankSVM models and test their capacity to predict a global order of audio assets, with and without audio processing effects, using pairwise rank annotations obtained from crowdsourcing. We build upon the methodology presented by Yang et al [14] and we extend it in the domain of sound (within games and beyond) through a crowdsourcing approach. Beyond arousal and valence, we put an emphasis and further model the affective dimension of tension. We also focus on sound designed specifically for the horror genre. Finally, we also study how audio signal modification techniques, such as reverb, can alter the perception of emotion in the original sound.

3 Methodology

This section describes how audio assets were selected to form a sound corpus, the annotation methodology followed via crowdsourcing and the subsequent preprocessing of the data. Finally we detail the algorithms which were used to construct computational models for tension, arousal and valence.

3.1 System Overview

Figure 2 shows the system overview utilized within this paper. A sound library of horror soundscapes is used and subsequently annotated by individual participants. Participant responses are obtained via a crowdsourcing methodology and subsequently stored into a database. Each sound in the library is represented by a feature vector, consisting of the low-level features extracted from each sound. A relationship between these low-level features and the participant annotations are then learned through a supervised learning method. In the context of this work a rank support vector machine was used to create a predictive global ordering of

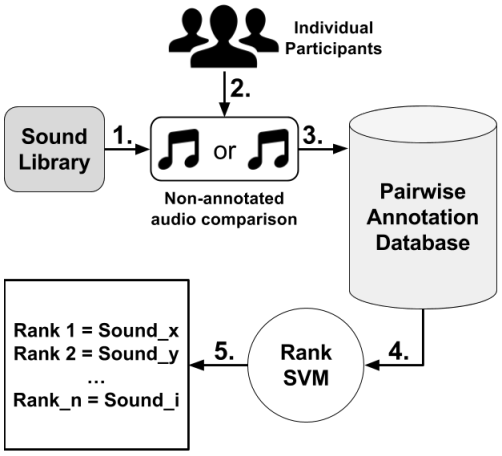


Fig. 2: The system pipeline presented in this paper: 1) The sound library provides a pair of sounds; 2) Participants compare sound pairs based on the perceived tension, arousal and valence; 3) Participant annotations are kept in an annotation database; 4) Annotations are used to train RankSVM models; 5) Trained RankSVMs predict a global ordering of unseen sounds.

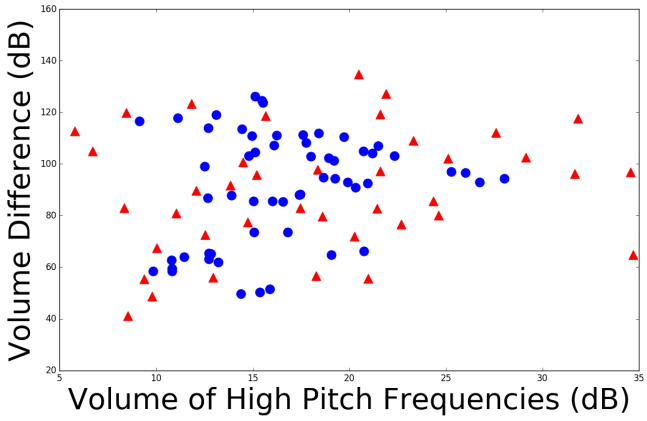


Fig. 3: Scatter plot of the entire *Sonancia* library. Triangles and circles are the selected and unselected audio assets, respectively. Volume difference is the peak-to-peak volume difference, while the volume of high pitch frequencies is the average power of all frequencies above 5k.

sounds according to the perceived affect of tension, arousal and valence.

3.2 The Sound Library

All audio assets were chosen from the existing database of 97 sound assets featured in the horror game generation system *Sonancia* [15]. Audio files consist of short audio loops between 5 and 10 seconds long. Each audio asset was recorded and produced by a horror sound expert using the *FM8* (Native Instruments) tool and the *Reaper* (Cockos) digital audio workstation. Due to the overwhelmingly high number of possible audio pair combinations out of 97 assets, 40 assets were carefully chosen by analysing their signal according to their *pitch* and *loudness*. To obtain pitch and loudness, we transformed each audio asset into a Hanning windowed spectrum with a linear frequency distribution, using the *Audacity* (Audacity Team) software. A spectrum is the power density (measured in decibels, dB), which measures the intensity and consequently "the loudness" of each frequency band, and in turn affects the overall pitch of sound.

According to Garner, et al. [38] loud and high pitch sounds tend to have a higher impact in eliciting fearful emotions. For this reason it was decided to plot each audio asset according to the peak-to-peak difference of volume, representing loudness, and the average power of frequencies above 5k, representing high pitch (see Fig. 3). To obtain a high degree of audio variability, the average Euclidean distance between all sounds in the loudness-pitch space was calculated. The 40 sounds with the highest distance were picked for the crowdsourcing experiment (see Fig 3).

Audio Effect Library

Audio signal processing effects, which will henceforth be referred to as *audio effects*, are processes that modify the original audio signal. In sound production, effects are widely used for multiple applications such as cancelling unwanted frequencies (i.e. low-pass and high-pass filters), add emphasis to certain recordings in the master recording (e.g. add an echo to the solo instrument) or even correct/change the pitch of a signal (i.e. automatic tuning). In both films and digital games, effects are regularly used for the same purposes mentioned, and additionally for simulating virtual environments [5], for the creation of novel sounds (e.g. the roar of a dinosaur) or for adding more emphasis to the base sound to convey more power than the original recording (e.g. the sound of a gun). An example of a specific audio snippet influenced by several different audio effects is available online[2] for the interested reader.

In this study we want to explore how effects can influence the perceived emotion in comparison to the original audio signal, and if a data-driven approach can potentially derive this relation between a sound effect, the audio piece and the perceived emotion. In this way, effects could potentially be used to alter the perceived emotion of an audio asset to accommodate the needs of a sound designer. Each effect is unique in altering the audio signal, but can be combined in a sequence, for designers who want to achieve a specific outcome. Different effects tend to differ on the number and type of adjustable parameters, which can affect the original audio signal to various ways and degrees. To accomplish this we decided to constrain the effect types to Reverb, Echo, Chorus, Flanger, Low Pass Filter, High Pass Filter and Pitch Shift. For each effect all the parameters were empirically predefined. Using the built-in *Unity* (Unity Technologies, 2005) effects library, we were able to modify the audio signal of the chosen audio assets and record them accordingly.

3.3 Audio signal preprocessing and feature extraction

Low-level descriptors (LLD) consist of, or more accurately "represent" information extracted from an audio signal itself. Usually there are three levels of extraction granularity

2. https://goo.gl/kfHP7Y

which are associated to 1) selecting arbitrary points in the signal; 2) defining sequential regions (i.e. frames) and 3) using pre-segmented regions. Depending on this granularity, several statistical values can be derived.

For this work the openSMILE audio feature extraction tool was used [39], and each value was normalized by its distance from the population mean (z-score). OpenSMILE is an open-access audio feature extraction tool that has been widely used for speech emotion recognition [40], [41], [42]. Features extracted followed the 'INTERSPEECH 2009 Emotion Challenge' feature set [43], comprising 384 statistical features. Each LLD is extracted through a sequential frame window of 25ms at a frequency of 10ms. In total 32 different types of LLDs are extracted: the root-mean-square signal frame energy (R_{Egy}), twelve mel-frequency cepstral coefficients ($MFCC_1$ to $MFCC_{12}$), zero-crossing rate of time signal (ZCR), the probability of voicing ($VProb$) and the fundamental frequency computed by the Cepstrum ($F0$). Each LLD is smoothed by an average filter according to the previous, current and following window. The additional 16 LLDs consist of the first order delta (Δ) of all the previous LLDs smoothed by the average filter. In total, 12 statistical features are derived from each LLD, resulting to a combined feature set of 384 features. The statistical features consist of:

- the maximum value of the contour (Max);
- the minimum value of the contour (Min);
- the difference between the maximum and minimum values (Rg);
- the absolute position of the maximum value (in frames) (F_{Max});
- the absolute position of the minimum value (in frames) (F_{Min});
- the arithmetic mean of the contour (μ);
- the standard deviation (σ);
- the skewness (λ);
- the kurtosis (kt);
- the slope of a linear approximation of the contour (apr_s);
- the offset of a linear approximation of the contour (apr_o);
- the difference between the linear approximation and the actual contour (quadratic error) (apr_e).

All these features were used to create two different datasets. The first dataset contains the statistical features obtained from audio pieces without any signal modification effects applied (the base audio dataset). The second dataset contains the statistical features of both base audio and each audio piece affected by every signal effect (the effect audio dataset). Furthermore the effect audio dataset contains 3 additional features, consisting of 3 binary values representing the specific effect that the audio is being affected by, out of the possible 7 different effect types. In particular "000" represents no effect, whereas any other 3-bit combination represents a particular effect.

3.4 Feature Selection

To reduce the feature dimensionality of the datasets, several feature selection methods were used. Due to the success of Mel-Frequency cepstral coefficients (MFCCs) in voice emotion recognition [44], two variants of both base and effect audio datasets were created, consisting of only the MFCC statistical features. Sequential feature selection (SFS) using both linear and radial basis function (RBF) support vector machines were also used to further reduce the dimensionality space of the datasets. SFS consists of sequentially selecting features that are best capable of improving the prediction accuracy, until the accuracy ceases to improve. A set of different parametrizations were used across all of the datasets, in order to experiment on how different SFS parameters could effectively be used in training RankSVM models. Additionally it is important to note that feature selection is exclusively run on the training data. Once training completes, all features selected through SFS are subsequently used with unseen data for validation.

3.5 Crowdsourcing Sound Annotations

To effectively obtain the ground truth of sound-elicited emotion, a large quantity of human annotated data was necessary for all the different combinations of audio samples and effects. Obtaining large corpora of training data through crowdsourcing has proven to be effective in several domains that involve annotations of subjective notions [31], [32]. For that purpose, a website[3] was developed allowing users to easily rank two different sounds based on the tension, valence and arousal affective dimensions. The start-up screen presents a detailed description of the experiment and each emotional definition (i.e. what is tension, arousal and valence). These descriptions are also shown in an unobstructed position during the experiment, by simply resting the mouse cursor on the question mark icon, in case a reminder is necessary. Each user is also asked to fill in a demographics survey consisting of age, gender, musical knowledge and how the user feels towards the horror genre. The system will log these details for each annotation, in case users decide to quit the experiment before all allocated sounds are annotated.

For annotating sounds we adopt a rank-based approach due to its evidenced effectiveness for highly subjective notions such as affect and emotion [27], [29], [45]. In the context of this work, sound annotation consists of reporting the emotional preference of the user between a pair of different audio assets (e.g. Sound A and Sound B) according to tension, valence and arousal using a 4-alternative forced choice (4-AFC) questionnaire. In particular, users must listen to each sound, and pick one of 4 different alternatives, for each affect dimension:

- Sound A is preferred over Sound B;
- Sound B is preferred over Sound A;
- Both are preferred equally;
- Neither is preferred.

For each participant the system can present either two different sound assets to annotate (base sound annotation experiment), or an audio asset and the same asset influenced by an effect (sound effect annotation experiment). Both experiments appear seamlessly to participants when using the crowdsourcing online framework, without specific

3. http://sonancia.institutedigitalgames.com

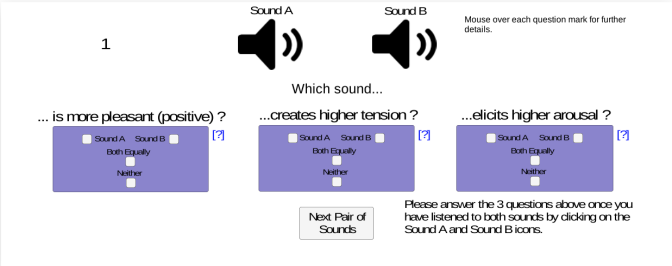


Fig. 4: The crowdsourcing annotation tool for sounds. The top two icons allow users to select and play one specific sound of the selected pair; only one sound can play at a time to avoid cacophony. The 4-AFC questions below ask the participant to rank valence, tension and arousal, respectively. Once participants have answered all questions, the user may press the "Next Pair of Sounds" button below, allowing the system to log and confirm their choices.

information about which effect is being used and which sounds are being played.

Each user is assigned two different audio samples from a general list of all existing sounds in the library. This list was randomly ordered a priori, making sure that users obtain the least amount of repeated sounds during their annotation task, and that the entire library is equally distributed to different users. Each user must listen to both sounds (in any order) and rank them, and they may hear them again any number of times. The system ensures that users have listened to both sounds at least once, and ranked them before moving to another pair of sounds. Figure 4 shows the user interface of the sound pair ranking annotation. To further validate and remove outliers that may derive from participants or a system failure, the crowdsourcing framework also logs the following data for each pair of sounds:

- The reported ranking (preference);
- The total time spent completing the task;
- The total amount of clicks;
- The time spent listening to each sound sample;
- The number of times the user listened to each sound;
- The number of times the user changed his responses and all previous values (if any) before committing to an answer.

Participants are asked to annotate a minimum of 6 sound pairs (3 pairs for the base audio dataset and 3 pairs for the effect audio dataset). After 6 pairs have been annotated, participants are encouraged to keep annotating more pairs but they may quit the experiment at any time they wish. To avoid losing information from annotators who disconnect early, each annotation is logged on to the server immediately after the user commits and confirms his answer.

The total number of pair combinations for the base sound annotation experiment is determined by the permutation of n ($n = 40$), being the total number of sound assets in the library and the combination size r ($r = 2$ being a pair): $P_2^{40} = 1560$. The total number of sound asset pairs required for the sound effect experiment is 1280 which is the product of 40 sounds times 32 effects per sound.

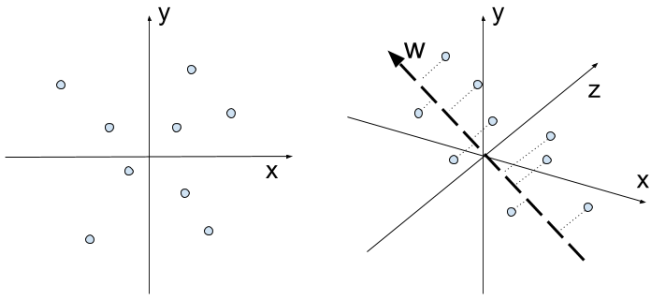


Fig. 5: Each transformed data point $\phi(q)$ is projected onto $\vec{w}$. The ordering of each projection according to the direction of $\vec{w}$ dictates the global order.

3.6 Preference Learning

All computational models constructed in this paper are trained using the Preference Learning Toolbox (PLT) [46]. PLT is an open-source accessible software featuring a variety of pre-processing, feature selection and preference learning algorithms such as evolving artificial neural networks (ANNs), ANNs trained with backpropagation, and Rank Support Vector Machines (RankSVM). Due to the reported efficiency of RankSVMs in numerous studies (e.g. see [14]) and its comparative computational benefits over ANNs we opted to employ RankSVM for the task of model construction based on annotated preferences.

RankSVMs are modified versions of regular support vector machines which were first introduced by Joachims [12]. This specific type of SVM attempts to maximize the Kendall's τ [47] between the expected ranking r^* and the proposed $r_{f(q)}$, where the feature space consists of a mapping $(\Phi(q, d))$ between a sound q and its ranking label d. The algorithm optimizes a boundary $\vec{w}$ (classifier) so that it accurately determines the ranking order of the feature space. Specifically all points in the feature space are generated by the training data and are labelled by their ranking information, which is subsequently used to find the boundary $(\vec{w})$ capable of describing their rank order (see Figure 5). RankSVMs also allow the application of different kernel types such as Radial Basis Function (RBF), which was also used as an alternative, to the linear SVM for comparative purposes. Within this work support vectors consist of pairwise preferences where the difference between their feature vectors represents the specific preference, similar to the methodology used in [48]. A global order can then be derived through the prediction of preferences from each pair in the dataset.

4 Analysis and Results

This section discusses the core descriptive statistics of the obtained crowdsourced data collected and the key results obtained. In particular, in Section 4.1 we analyse the data collected from the crowdsourcing experiments; then in Section 4.2 we derive a global order of affective rank responses and finally, in Section 4.3 and Section 4.4 we present the core results from the affect modelling experiments on the crowdsourced data via preference learning.

TABLE 1: The average time (in seconds) and the respective standard error in parenthesis of (from left to right): total time required for both experiments; total time for base sound experiment; total time for sound effect experiment; total time listening to sound A for both experiments; total time listening to sound B for both experiments.

Both Exp	Sound Exp.	Effect Exp.	Sound A	Sound B
47.23 (4.1)	41.88 (2.8)	51.59 (7.01)	16.8 (0.47)	15.54 (0.46)

4.1 Data collected

The crowdsourcing platform was heavily disseminated over social media platforms, including Twitter, Facebook and Reddit; scientific conferences and within the University campus. At the time of writing 1009 annotations are collected in total: 453 of these annotations consist of comparisons between two different sounds, while the remaining 556 are comparisons between a sound and one of its effects. Annotators were 31.2% female, 67% male and 1.1% did not specify. The majority of annotations came from the age group between 25 to 34 years of age (52%), while the second highest was between 18 and 24 years (23.2%). Further, 73% of the annotators were non-musicians (never played an instrument), while the remaining were non-professional (21%) and professional musicians (5%). Interestingly the majority of annotations came from people who enjoy the horror genre (56.5%); 13% of these stated it was their favourite genre. Approximately a fourth of annotators (26%) claimed they do not enjoy this genre, while the remainder 16% did not have an opinion on this specific question. Table 1 shows the average times taken to complete tasks during the crowdsourcing experiment.

To combat bias and ambiguity within the data annotations, a random order was applied to the dataset. Additionally several annotations were pruned due to ambiguous answers by participants, defective annotations and the lack of sufficient annotations for specific sounds. The pruning methods used are further described below in each section. The dataset used for the following experiments are publicly available here[4].

4.1.1 Sound Ranking Experiment

The sound ranking experiment amassed a total of 453 annotations. The distribution among the four available preference options is shown in Table 2. For tension and arousal participants were more forthright in preferring one of the two sounds, although a slight skew is noticeable towards sound B. Valence on the contrary presented very balanced responses between A and B; however a high number of participants stated that neither sounds were pleasurable, which is not surprising considering the audio library used was specifically designed for the horror genre.

In order to apply supervised preference learning, ambiguous annotations were discarded (i.e. Both Equally and Neither) [28]. Following pruning the total resulting base sound annotations amounted to 403, 389 and 334 for tension, arousal and valence, respectively.

4. http://www.autogamedesign.eu/sonancia

TABLE 2: The preference distribution of the crowdsourced sound ranking experiment.

Affect	A	B	Equally	Neither	Total	**Baseline Accuracy**
Tension	187	216	34	16	453	**53.6%**
Arousal	170	219	29	35	453	**56.3%**
Valence	168	166	10	109	453	**50.1%**

For comparison purposes a baseline value was derived, and consists of the maximum accuracy obtained by exclusively picking either sound A or B (i.e. the most dominant preference of the two). Based on Table 2, for tension and arousal the baseline always picks sound B, and for valence always picks sound A. The baseline accuracy is computed as the highest number of A or B chosen (e.g. B in tension, 216 times) divided by the number of times A or B was chosen (e.g. 403 times for tension). We can observe that there is no clear primacy or recency effects and that baseline accuracy is very close to chance levels for all three affective dimensions examined, meaning no clear favouritism was visualized between either sound A or B.

Some insight might be gleaned from the relationship between global ranks of valence, arousal and tension. Although there is a positive rank correlation between tension and arousal (0.25) and a negative correlation between valence and arousal (-0.13), respectively, this effect is not substantial. There is however a substantial negative rank correlation between tension and valence (-0.45). This is not surprising, as it is due to both the inherent nature of the audio assets themselves, and also the opposite nature of these two dimensions; being tense is rarely pleasurable.

4.1.2 Sound & Effect Ranking Experiment

For this experiment both the audio signal effect annotations were combined with the previous sound ranking annotations. This allows for the creation of a more generalized model, able to predict a rank between two diverging sounds and between an audio piece with or without an effect. It also increases the amount of training data to a total of 1009 annotations. For the sake of simplicity a sound that is not influenced by an effect will be referred to as a "base sound".

Table 3 shows the preference distribution of both experiments. An initial analysis of data reveals that the majority of users (79%) annotated sounds that are influenced by effects as less tense and arousing than the base sounds. Interestingly, a slight majority stated that sounds influenced by effects were more pleasurable than the base sounds (63.6%). We assume that this was due to the capacity of some effects to lower substantially the volume of the original sound, which potentially correlates to how users relate to arousal and tension. Further analysis of the preference distribution also shows a significant skew towards the sound B option across all affect annotations. This skew is most likely caused due to the current annotation dataset which associated effected sounds to sound A, and eventually influenced the participants' reported preference.

TABLE 3: The preference distribution of post-pruned crowdsourced annotations, for the sound and effect ranking experiment.

Affect	A	B	Equally	Neither	Total	**Baseline Accuracy**
Tension	177	377	221	82	857	**68%**
Arousal	162	367	216	122	867	**69%**
Valence	244	181	141	336	902	**57%**

Noticeably there is also a higher number of ambiguous answers, suggesting that certain effects did not influence the base sound in such a way that was noticeable to the participants. These results also show a particular challenge with the effect parametrization, which we did not anticipate. For the purposes of this experiment a global set of parameters were defined for each of the effect types beforehand. However, some sounds were unaffected by these parameters (e.g. sounds without a frequency filtered by an effect). For example a sound which consists of low frequencies will be rarely affected by a high pass filter, as this effect may merely remove high frequencies.

Ambiguous rankings (both equally or neither) were discarded from the datasets for each affective dimension. Four entries were also removed from the dataset due to a failure with the logging system. Several sound and effect pairs were also removed from the dataset, due to audio clipping issues providing unreliable low-level features of those sounds. In total 554 (306 sound and 245 effects), 529 (295 sound and 234 effects) and 425 (267 sound and 158 effects) data points were kept for tension, arousal and valence, respectively.

The baseline accuracy was computed based on the most preferred sound between A and B, as described in Section 4.1.1. The baseline accuracy for all affects increased substantially compared to the previous experiment, as shown in Table 3. For tension and arousal, users picked sound B twice as often as sound A. The observed skewness of the baseline is likely due to the lack of a complete annotation corpus, as previously described, and due to the fact that participants often preferred the base sounds instead of the ones with effects. Effected sound was always sound A, which users often did not consider as tense or arousing as sound B.

Similarly to the previous results both valence-arousal (-0.15) and the tension-valence (-0.42) rankings are negatively correlated, although with slightly differing results. However, the correlation between tension-arousal (0.46) increased. This is most likely due to the influence of some effects on the volume of the base sound, which potentially made the effected sounds quieter. Louder sounds tended to be perceived as both more tense and more arousing in comparison to those with a lower volume.

4.2 The Global Order of Sound Annotations

The 40 sounds are ranked based on the human-annotated tension preferences. The *global order* is derived through the pairwise preference test statistic [28] which is calculated as $P_i = (\sum_i^N z_i)/N$, where P_i is the preference score of sound i, z is $+1$ if the sound i is preferred or -1 if the sound is not preferred in a pair of sounds, and N is the number of samples for sound i. The obtained preference scores P define the global order (rank) of each sound with respect to tension, arousal and valence.

Figure 6 shows the obtained preference scores P_i for each affective dimension and sound asset, ordered by the global ranking of the tension dimension. By observing the figure we can see that both tension and valence tend to oppose each other quite frequently. Surprisingly the arousal and tension dimensions did present some diverging results, which were not expected, such as situations where participants annotated a specific sound as being tense, but not arousing, e.g. sound 9; or very arousing but not particularly tense, e.g. sound 8. Interestingly the sound ranked highest in both the valence and arousal dimensions was the same, but, that sound is only ranked 32nd out of 40 in the tension global order (see Fig. 6). A general observation, however, is that highly tense sounds are annotated as arousing with rather low valence, whereas, less tense sounds are usually characterised by higher valence and lower arousal values. This observation naturally follows the rank correlations between the affective dimensions.

For the interested reader, the 5 top and bottom ranked sounds in the tension dimension can be listened to online[5]. When listening to all the aforementioned sounds, the first 4 consist mainly of high pitch sounds, while sound 5 is a constant low pitch sound. Although the first 4 sounds are in-line with the studies of Garner et al. [38], we hypothesize that sound 5 obtained such a high rank due to how uncomfortable it is to listen in a constant loop. Interestingly, the sound that ranked first is a higher pitch version of sound 38 (one octave lower) and 40 (two octaves lower) which is also in-line with Garner et al.'s findings. However a notable exception is present with sound 36, which consists of a high pitch sound compared to any of the top 5 tense sounds.

For comparison purposes the top and bottom 5 ranked sounds for the arousal dimension can be listened to online[6]. Most top ranked sounds consist of lower pitches when compared to the previous tension global rank, with the exception of sound 4, which is the same sound that was ranked third for tension. However, most users considered sounds with a lower pitch as more arousing than higher pitch sounds. This is evident with sound 2 and 38 which consist of the same sound in a lower and higher octave, respectively. High ranked sounds also consist of a mix between audio with small rhythmic patterns, present in sound 1 and 2, while sounds 3 and 5 consist of audio with no specific rhythm.

As with tension and arousal, the top and bottom five ranked audio assets for the valence dimension can be heard online[7]. Most highly ranked sounds consist of audio where the majority of frequencies were in the moderate octave range; on the other hand, higher pitched sounds were ranked lower.

To study the relationship between high pitch or high volume, which are indications of tense sounds [38], and the obtained global ranks, the kendall's τ correlation coefficient was calculated [47]. Table 4 shows the correlation and p-

5. https://goo.gl/Z2ihfo
6. https://goo.gl/IbY0gf
7. https://goo.gl/E7VIu0

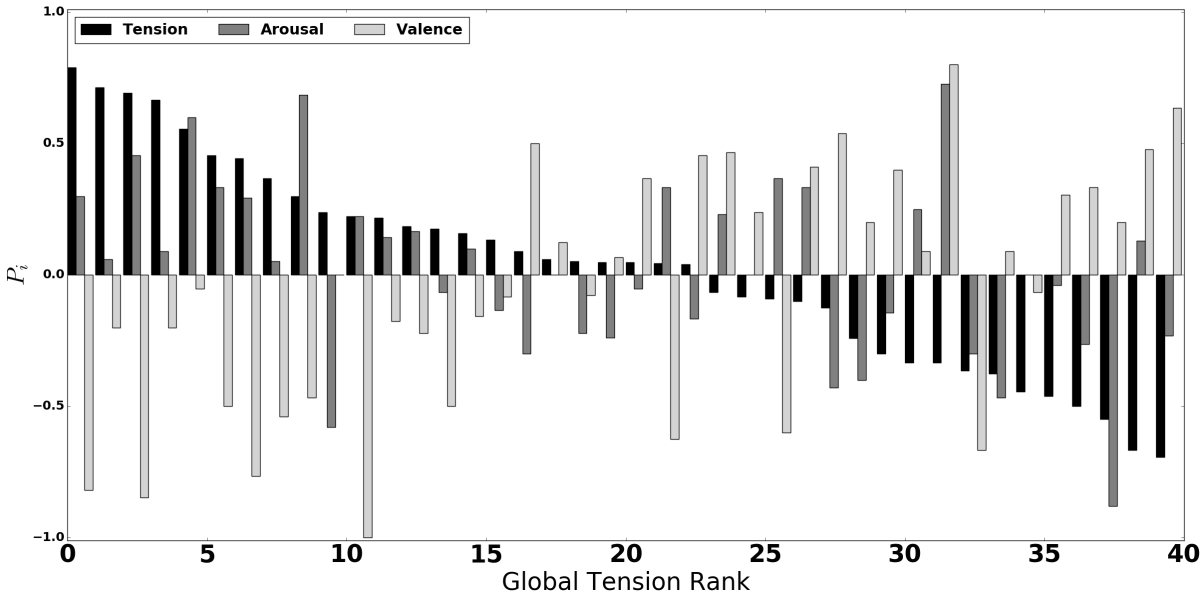


Fig. 6: The global order and distribution of the annotated sounds in each affective dimension: tension (black), arousal (grey) and valence (white). The *y-axis* consists of the preference score value (P_i) and the *x-axis* consists of the sound rank according to the tension dimension, ordered by the most to less tense sounds.

TABLE 4: Kendall's τ correlation and p-value (in parenthesis) between the global order of each affect and the rank of both the volume difference and high pitch frequencies.

	High Pitch Frequency	Volume Difference
Tension	0.04 (0.67)	0.05 (0.61)
Arousal	0.20 (0.06)	-0.04 (0.67)
Valence	0.19 (0.07)	0.04 (0.69)

values, between the global order of each affective dimension and the rank of both the volume difference and the high pitch frequencies. Our analysis strongly suggests that perceived emotion in audio has a deeper complexity, and that a linear relationship between low-level features and a perceived effect might not be sufficient. It thus suggests that a more complex relationship, possibly supported by additional features can potentially improve the task of audio affect modelling.

4.3 Learning to Rank Sound

The creation of a model that is capable of ranking "unseen" sound assets can be beneficial to automated sonification systems that may evaluate the affective impact of a new sound and place it within a particular context in any form of human computer interaction: for instance, in a particular room of a new game level. This can, in turn, allow the system to create specific emotional progressions based on how each sound asset is ranked by the model. This section discusses the results obtained from training different models capable of ranking sounds based on tension, arousal and valence. Please note that for the remainder of this paper we present the best average accuracy obtained for each affective dimension but we also provide the accuracy of the best fold in parentheses.

Figure 7 shows the average 5-fold cross-validation accuracy of the two different RankSVM kernels employed (linear and RBF). For tension the best average obtained was 65% (68%), using SFS on the MFCC LLDs and a RBF kernel set to a gamma value of 0.2. The linear kernel performed worse in comparison to RBF, but was still able to improve upon the baseline. SFS proved to be advantageous for the tension dimension, as it consistently improved accuracy regardless of the kernel used.

Interestingly arousal was the most difficult to predict of the three affective dimensions, which was surprising considering that literature states otherwise [14]. Without the application of SFS the accuracy of the models rarely achieves the baseline independently of the kernel parameters or the dataset used. Analysing Fig. 7 we can see that most models are capable of achieving higher accuracies in comparison to the baseline, where the main exception is the linear models trained exclusively with MFCC. The best obtained accuracy is 66% (69), 10% over the baseline, by applying SFS with all the LLDs and training with the RBF kernel. Surprisingly the MFCC trained models obtained much higher accuracies through the RBF kernel. There was also not much difference between both All and MFCC trained model types. Considering that arousal is often closely associated to rhythm [49], it is surprising that it achieved similar accuracies as these types of features are absent in the MFCC dataset. A potential reason why the other affective dimensions outperformed arousal significantly, is due to it being an uncommon affective description to an untrained annotator (crowd), compared to the other affective dimensions of tension and pleasure (valence).

Contrary to arousal, valence was easier to predict and corresponding models yield the best accuracies compared to the other two affective dimensions (see Fig. 7). The best average accuracy of 72% (79%) was obtained using an RBF

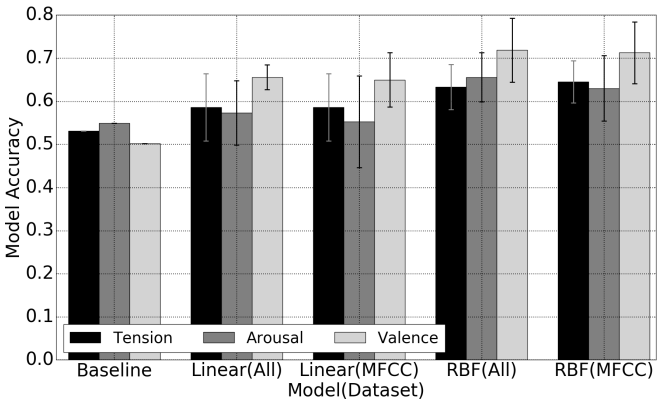


Fig. 7: Learning to rank sound: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models, employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). Sequential feature selection is applied in all experiments reported. Presented accuracies for RBF consist of the best accuracy obtained via extensive parametrization testing.

kernel on the "All'' dataset, whose features were selected through SFS. This specific model was able to improve upon the baseline by 22%. Despite a few exceptions, models trained without SFS still managed to obtain values above 60%, while models that did apply SFS obtained a substantial increase in both datasets.

In conclusion both Tension and Arousal were indeed harder to train in comparison to the Valence affect. We hypothesize that this was due to the specific sound library used, which focused specifically on sounds in horror. It is easier to learn the relationship between pleasurable sounds, when a low number of these potentially exist within the library. On the other hand for Tension and Arousal a greater "competition'' between high-tense and high-arousal sounds exist, making it harder to learn these relationships due to potentially unclear distinctions, and possibly diverging user opinions within the annotations.

4.3.1 Selected Features

For brevity Table 5 shows the selected features obtained through SFS, of the most accurate fold of a 5-fold cross-validation experiment with the highest average accuracy across all folds. This is necessary as each fold is trained independently with feature selection and then subsequently tested on unseen validation data, meaning that each fold will select substantially different features. For tension, the majority of features selected were MFCC statistics, suggesting that out of all features available MFCC descriptors were more capable of finding a relation to tension than the other descriptors. Interestingly, the fold presented in Table 5 was the only fold to utilise one feature, and achieved an impressive testing accuracy.

Alternatively both the RBF(All) and RBF(MFCC) arousal models achieved similar average accuracies, despite using a diverging number and set of features. While RBF(MFCC) obviously focused on MFCC features exclusively, it relied on a lot less features than the models trained with RBF(All).

TABLE 5: The selected features of the most accurate fold with the best obtained average accuracy model parameters of each affect.

	Tension	Arousal	Valence
Model	RBF(MFCC)	RBF(All)	RBF(All)
Selected Features	$Rg(MFCC_3)$	$kt\Delta MFCC_1$ $\mu\Delta MFCC_9)$ $Min(\Delta MFCC_3)$ $apr_o(MFCC_4)$ σR_{Egy}	$F_{Max}(\Delta MFCC_3)$ $\mu\Delta MFCC_{10}$ $\mu MFCC_6$ $Max(MFCC_1)$

Using SFS with RBF(All) consistently chose RMSEnergy features, which then influenced the remaining chosen set as the algorithm attempted to find the best combination as to optimize accuracy. This particular example shows one the main weaknesses of SFS. Being a greedy algorithm, SFS chose the best feature that maximizes model accuracy sequentially. However, this does not guarantee that the set of features in conjunction outperforms another feature set, as different combinations might result in better predictions even though the first selected feature was performing worse. Therefore, feature pruning can still be beneficial when using an SFS algorithm. Alternatively a genetic feature selection algorithm might prove more useful in future studies, even though it is computationally more intensive. A Sequential Backward Selection can also potentially help, as it starts from the entire combination of features and removes each feature that does not significantly decrease model accuracy.

Similarly to tension, the valence models also abundantly chose MFCC statistical features. It does suggest that both tension and valence have a closer relationship to tonic and harmonic features.

4.4 Learning to Rank Sound & Effects

This section presents the predictive accuracy obtained from training various SVM models that rank both base sounds and how their perceived affect is influenced by different effects, and between the base sounds.

Figure 8 shows the average tension, arousal and valence accuracy over 5 folds for different RankSVMs. Unfortunately no significant improvements were obtained from the baseline, suggesting that certain types of sound effects were more detrimental than helpful to the overall prediction of perceived affects.

For tension the Linear(MFCC) model consistently obtained averages between 62% and 68%. The linear RankSVM performed better with the All dataset, but compared to the RBF kernel it performed worse. For tension the highest average accuracy obtained was 72% (78%).

Arousal models using the MFCC features performed slightly worse than the entire LLD feature dataset. Additionally with the exception of the RBF(All) models, arousal rarely achieved average accuracies surpassing the baseline, even though the performance increased in comparison to the previous experiment. We assume that this jump in accuracy was due to how certain effects altered the base sound's volume, which has often been closely associated

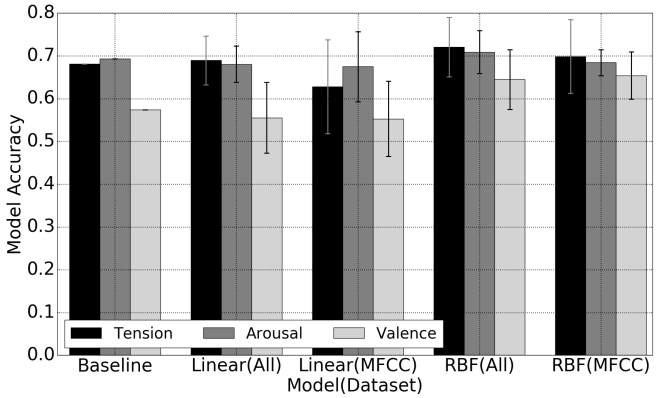


Fig. 8: Learning to rank sound and sound effects: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). Sequential feature selection is applied in all experiments reported. Presented accuracies for RBF consist of the best accuracy obtained via extensive parametrization testing.

to loudness [49]. Arousal consistently underperformed obtaining values below the baseline, with the exception of RBF(All), which was able to slightly surpass the baseline with an accuracy of 71% (76%).

Compared to the other affective dimensions, valence performed worse. We believe that this is due to the significant amount of ambiguous rank data obtained in this affect dimension in comparison to the others. However, RBF models with SFS were consistently able to achieve accuracies above the baseline. Valence RankSVMs did not manage to achieve average accuracies above 65%, despite parameter tweaking. Also unlike all the other dimensions, valence models failed to hit the 70% average accuracy bar. Applying SFS was crucial for improving performance of valence models: initial testing showed that these models rarely achieved average accuracies above the 60% mark, without SFS. The best average accuracy obtained was 65% (71%), with the RBF(MFCC) model.

4.4.1 Selected Features

Similarly to Section 4.3.1 this section will detail the selected features chosen by the SFS algorithm using the same annotations for simplicity. Additionally the effect input parameter is represented as $Effect_x$, where x is the effect's index.

In this particular experiment the SFS was less biased towards MFCC statistics, even though they are still quite substantially present. Interestingly the effect input binaries did not prove to be particularly helpful for affect prediction, with only the tension model taking one into account. Additionally in the majority of RBF(All) models presented statistical features related to R_{Egy} more consistently than the previous experiment. We hypothesize that this was due to how sound effects substantially change the volume and/or pitch in comparison to the base sound. These alterations can particularly influence how tension, arousal and valence are perceived in comparison to the base sound. High volume can influence tension and arousal [38], while a too high or a too low pitch can cause a sense of discomfort impacting the valence state.

TABLE 6: The selected features of the most accurate fold with the best obtained average accuracy model parameters of each affect.

	Tension	Arousal	Valence
Model	RBF(All)	RBF(All)	RBF(MFCC)
Selected Features	$M_{in}(R_{Egy})$ $apr_s(MFCC_9)$ $Effect_2$ $F_{Max}(MFCC_6)$ $F_{Min}(MFCC_{11})$ $\sigma F0$ $M_{in}(MFCC_2)$ $Rg(MFCC_4)$ $\sigma MFCC_3$ $\mu VProb$ $Max(MFCC_6)$ $\sigma\Delta MFCC_6$	$M_{in}(R_{Egy})$ $\mu MFCC_7$ $apr_e(MFCC_1)$ $\mu\Delta MFCC_4$ $apr_e(ZCR)$	$kt\Delta MFCC_2$ $kt\Delta MFCC_3$ $apr_e(MFCC_{11})$ $apr_o(\Delta MFCC_{12})$ $F_{Max}(MFCC_6)$ $Rg(MFCC_4)$ $Rg(MFCC_8)$

TABLE 7: Comparison of the rankings between the base sound and 4 different effects in the predictive global ranking of the most accurate fold of the tension, arousal and valence affects. For brevity the highest ranked effect or base sound is chosen for analysis.

	Tension	Arousal	Valence
Effects	Rk 1 (Echo) Rk 42 (Chorus) Rk 63 (Reverb) Rk 93 (Reverb)	Rk 1 (Reverb) Rk 2 (Reberb) Rk 5 (Reverb) Rk 21 (Chorus)	Rk 1 (Reverb) Rk 15 (Reverb) Rk 196 (Reverb) Rk 225 (Flange)
Base Sound	Rk 8	Rk 4	Rk 1069

4.5 Rank Comparison of Sound and Effects

To study the impact of effects on each affective dimension, the predicted global rank obtained from the most accurate fold for tension, arousal and valence of Table 6 is analysed. Table 7 shows the rankings of a base sound and its 4 highest ranked effects within the predicted global rank. For the interested reader all sounds presented in Table 7 can be listened to here[8].

Valence showed the most surprising results, where effects greatly influenced the enjoyability of high pitch base sounds. Particularly reverb often influenced both pitch and volume substantially improving the enjoyability of the sound in comparison to the base sound. Tension in particular showed more varied effect influences, where certain effects had a higher consistency of improving the perceived tension (e.g. Echo and Reverb), while others often deteriorated (e.g. Flange) in comparison to the base sound. Alternatively, arousal was not influenced by effects. The base sound is often within the general rank vicinity of its effects, having a minor impact on the base sound.

5 Discussion

Music-elicited emotion recognition is a complex task due to the ambiguous nature of human emotions and the

8. http://www.goo.gl/Qmp019

subjectivity of sound perception. In this work we attempted to construct models capable of learning the relationship between low-level statistical descriptors of audio, and their perceived emotion. The best models constructed for tension obtained average accuracies between 65% and 72%. Results obtained from crowdsourced user annotations suggest that a divergence exists between tension and the affective dimensions of arousal and valance, which validates, in part, the viability of the Schimmack and Grob [13] model. However, due to the context of this work within the horror genre, a more general approach might be required to attest these findings. It is also worth noting that the tension affective models obtained similar or higher predictive accuracies, when compared to models of arousal and valence.

For the base sound comparison experiments, the most successful affect consisted of the valence models, which achieved a cross-validation accuracy of 72%. Surprisingly, arousal performed much worse, achieving only 66% prediction accuracy. We hypothesize that this is due to the LLDs being too specific to the voice emotion recognition problem, which tends to concentrate on harmony and timbre (e.g. Mel-Frequencies) rather than rhythm. This can be observed through the selected features of both the valence and tension models, where there was a substantial favouritism towards MFCC statistical features. Although arousal did outperform tension, it did so by a relatively small margin ($\sim 1\%$). We hypothesize that the similarity of sounds within the audio library also contributed to the higher training difficulty, as more conflicting annotations might be present within these affective states. This could potentially be rectified with more user annotations, solidifying the relationship between the different sound pairs. Unfortunately the retention of large quantities of user annotated data is still a difficult task to achieve, even when utilising a crowdsourcing solution.

Interestingly valence models performed better than what we initially expected, assuming that the majority of models obtained would have a high degree of ambiguity. This expectation is consistent with the participant annotations, given that valence resulted in the most ambiguous answers compared to the other two affective dimensions. We hypothesize that due to this ambiguity, sounds that were in fact annotated as pleasurable presented clearer distinguishable features facilitating training. In general, due to the context of this work within the framing of horror, there is little we can clearly state about valence, as sounds were specifically designed to be unpleasant.

For our second set of experiments, models were trained on the combined annotations of base sound and sound-effect pairs, which improved the accuracies of both the tension and arousal models. The best model obtained for tension and arousal achieved an average accuracy of 72% and 71%, respectively. We believe that this improvement is due to the dominant preference of sounds uninfluenced by effects in these two affective dimensions, which consequently facilitated learning. Effects that had little influence on the base sounds, were also heavily filtered by participants who could not distinguish any difference between them. This allowed us to retain sound features of the more influential effects to train our models. Valence however was slightly harder to train compared to arousal and tension (the best model achieved an average accuracy of 65%) due to the reasons stated earlier. Unlike tension and arousal there was no clear valence preference between base sound and sounds with effects.

Although there was a large user participation in the crowdsourcing annotation experiment, we were unable to obtain annotations for all possible pair of sounds or base sound-effect pairs. This was apparent for the sound effect-pair experiment, where we were unable to get more than half of the required annotations (1009 out of 2840), while also discarding ambiguous user answers. This caused the data to be particularly skewed towards sounds without effects, which was evident in our effect experiment baselines.

Crowdsourcing data suggest that effects did not produce the variation intended between the base and effected sounds. This is likely why the majority of effect annotations were ambiguous and subsequently discarded. This ambiguity stems from the constant parameters that were set for each sound in the library. Even the application of certain effects to specific sounds may not be appropriate; for instance, applying low pass filters to sounds whose signal is mostly of low frequency may result in complete silence. This limitation can potentially be eliminated by ad-hoc selecting each effect parameters that best alter each sound within the library. Another potential solution would be to automate this process, allowing a machine learning model to set effect parameters that best alter a specific sound.

While our feature extraction is already thorough, more sound features need to be investigated in future studies. Preliminary results, however, suggest that the models' accuracy does not improve with the addition of certain features, as shown by the small accuracy variation between the "MFCC'' and "All'' datasets.

As a final step towards realising affective interaction via sounds in horror games we intend to use our models in already developed tools that can be used by sound designers directly. *Sonancia* [15] provides an appropriate platform for future experiments with the affect models introduced in this paper. Sonancia procedurally generates game levels and corresponding sounds based on a designer defined progression of tension. Our models can be used to autonomously select sounds from the library, apply particular sound effects and subsequently place the resulting audio asset within the virtual world to match the defined progression. Other potential application domains include experience-driven generated games [50] in which the obtained models would allow designers or automated processes to specify intended experiences for players. This can be achieved for diagnostic or therapeutic purposes [51], for realising effective game-based learning [52], [53] or alternatively for enabling an AI-assisted game design approach [54] that can suggest soundscapes which are expected to elicit particular emotive patterns.

6 Conclusion

This paper studied how sound, specifically designed for horror, can influence the emotional state of human users in the tension, arousal and valence affective dimensions. We also investigated how sounds passed through a digital signal processing effect could potentially alter the emotional state perceived. User preferences of each sound and effect

pairing from our library were annotated for each dimension via crowdsourcing. A global rank of each sound in the library was constructed from the preferences obtained for each affective dimension. Our findings suggest that highly ranked sounds in the tension dimension are often ranked lower in the valence dimension, revealing a negative correlation between the two. Participants also tend to prefer sounds without effects in terms of tension and arousal, while no clear preference was derived for valence.

Further to the descriptive statistical analysis, this paper proposed several data-driven models capable of predicting the global rank of horror sounds within the same affective dimensions. Low-level descriptors for each sound and sound effect were extracted with the openSMILE sound feature extraction tool. The features were divided into two different datasets: one containing all of the extracted features, and another containing only the MFCC features. RankSVM models, using both the linear and RBF kernels, were trained to predict the annotated user preferences on both datasets in conjunction with sequential forward feature selection. In general, results obtained suggest that tension and arousal had a similar degree of training difficulty. Valence proved to be less difficult to predict in the sound comparison experiment, which is consistent with many other studies in the domain of music-based affect modelling. However, once the effect dataset was included, valence models performed worse in comparison to the other two affective dimensions. Nevertheless, it might be dangerous to derive a general conclusion about the viability of these models outside the horror genre, as the sound library used has a potential bias towards unpleasant sounds.

The key findings of the paper suggest that a model of tension could potentially be constructed. Even though tension models did not substantially outperform other affect dimensions, they did consistently obtain similar performances showcasing robustness across learning tasks.

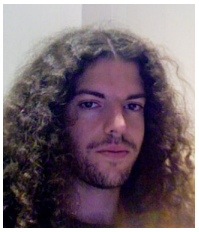
ACKNOWLEDGMENT

This work was supported, in part, by the FP7 Marie Curie CIG project AutoGameDesign (project no: 630665) and the Horizon 2020 project CrossCult (project no: 693150).

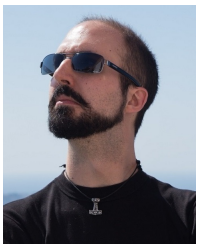
REFERENCES

[1] K. Fahlenbrach, "Emotions in sound: Audiovisual metaphors in the sound design of narrative films," *Projections: The Journal for Movies and Mind*, vol. 2, no. 2, pp. 85–103, 2008.
[2] D. Sonnenschein, *Sound design*. Michael Wiese Productions, 2001.
[3] K. Collins, *Playing with sound: a theory of interacting with sound and music in video games*. MIT Press, 2013.
[4] H.-P. Gasselseder, "Re-scoring the games score: Dynamic music and immersion in the ludonarrative." in *Proceedings of the Intelligent Human Computer Interaction Conference*, 2014, pp. 1–8.
[5] R. Stevens and D. Raybould, *The Game Audio Tutorial: A Practical Guide to Creating and Implementing Sound and Music for Interactive Games*. Taylor & Francis, 2013.
[6] S. Serafin and G. Serafin, "Sound design to enhance presence in photorealistic virtual reality." in *Proceedings of the International Conference on Auditory Display*, 2004.
[7] J. Togelius, G. N. Yannakakis, K. O. Stanley, and C. Browne, "Search-based procedural content generation: A taxonomy and survey," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 172–186, 2011.
[8] A. Liapis, G. N. Yannakakis, and J. Togelius, "Sentient sketchbook: Computer-aided game level authoring." in *Proceedings of the 8th Conference on the Foundations of Digital Games*, 2013, pp. 213–220.
[9] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.
[10] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2010, pp. 255–266.
[11] I. Ekman and P. Lankoski, "Hair-raising entertainment: Emotions, sound, and structure in silent hill 2 and fatal frame," *Horror video games: Essays on the fusion of fear and play*, pp. 181–199, 2009.
[12] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.
[13] U. Schimmack and A. Grob, "Dimensional models of core affect: A quantitative comparison by means of structural equation modeling," *European Journal of Personality*, vol. 14, no. 4, pp. 325–345, 2000.
[14] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.
[15] P. Lopes, A. Liapis, and G. N. Yannakakis, "Targeting horror via level and soundscape generation," in *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.
[16] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[17] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, p. 494, 2008.
[18] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, 2010.
[19] I. Daly, E. B. Roesch, J. Weaver, and S. J. Nasuto, "Machine learning to identify neural correlates of music and emotions," in *Guide to Brain-Computer Music Interfacing*. Springer, 2014, pp. 89–103.
[20] K. Trochidis and E. Bigand, "Emotional response during music listening," in *Guide to Brain-Computer Music Interfacing*. Springer, 2014.
[21] L.-L. Balkwill and W. F. Thompson, "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," *Music Perception: An Interdisciplinary Journal*, vol. 17, no. 1, pp. 43–64, 1999.
[22] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of music*, vol. 24, no. 1, pp. 68–91, 1996.
[23] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[24] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1989.
[25] P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot, and M. Sandler, "Genre-adaptive semantic computing and audio-based modelling for music mood annotation," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 122–165, 2016.
[26] T. Eerola, "Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research*, vol. 40, no. 4, pp. 349–366, 2011.
[27] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Dont classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.
[28] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: a comparative study of self-reporting," in *Proceedings of the Affective Computing and Intelligent Interaction conference*. Springer, 2011, pp. 437–446.
[29] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.
[30] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, vol. 22, no. 140, 1932.
[31] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl, "Story generation with crowdsourced plot graphs." in *The 9th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.
[32] N. Shaker, G. N. Yannakakis, and J. Togelius, "Crowd-sourcing the aesthetics of platform games," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, no. 3, 2013.

[33] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer, 2011.
[34] G. N. Yannakakis, "Preference learning for affective modeling," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2009, pp. 1–6.
[35] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 313–340, 2010.
[36] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
[37] Y.-H. Yang and H. H. Chen, *Music emotion recognition*. CRC Press, 2011.
[38] T. Garner, M. Grimshaw, and D. A. Nabi, "A preliminary experiment to assess the fear value of preselected sound parameters in a survival horror game," in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*. ACM, 2010, p. 10.
[39] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.
[40] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.
[41] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.
[42] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2013.
[43] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2009.
[44] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[45] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proceedings of the Affective Computing and Intelligent Interaction conference*. IEEE, 2015, pp. 574–580.
[46] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709*, 2015.
[47] D. Wilkie, "Pictorial representation of kendall's rank correlation coefficient," *Teaching Statistics*, vol. 2, no. 3, pp. 76–78, 1980.
[48] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 1999.
[49] E. R. Miranda and J. Castet, *Guide to Brain-Computer Music Interfacing*. Springer, 2014.
[50] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.
[51] C. Holmgård, G. N. Yannakakis, H. P. Martínez, K.-I. Karstoft, and H. S. Andersen, "Multimodal PTSD characterization via the startlemart game," *Journal on Multimodal User Interfaces*, vol. 9, no. 1, pp. 3–15, 2015.
[52] P. Lopes, A. Liapis, and G. N. Yannakakis, "The C2Create authoring tool: Fostering creativity via game asset creation," in *Proceedings of the Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–2.
[53] R. Khaled and G. N. Yannakakis, "Village voices: An adaptive game for conflict resolution." in *Proceedings of the 8th Conference on the Foundations of Digital Games*, 2013, pp. 425–426.
[54] G. N. Yannakakis, A. Liapis, and C. Alexopoulos, "Mixed-initiative co-creativity," in *Proceedings of the 9th Conference on the Foundations of Digital Games*, 2014.



Phil Lopes is currently a PhD Student at the Institute of Digital Games of the University of Malta. He completed his M.Sc. in Computer Science at the University of Lisbon, Portugal. During his master thesis, he developed a mixed-initiative musical tool called the Traveling Percussionist. Additionally at the GAIPS INESC-ID laboratory at the Instituto Superior Técnico he developed the Geometry Friends competition framework. Currently his research focus consists of developing new tools and methodologies for procedurally generating content at the interplay of audio, level design and visuals within the horror digital game genre; while also exploring new ways of automating and adapting sound to 3D virtual environments. Currently his most developed system is the Sonancia generator, a multi-faceted generator for horror.



Antonios Liapis is a Lecturer at the Institute of Digital Games, University of Malta (UoM). He received his 5-year Diploma (2007) in Electrical and Computer Engineering from the National Technical University of Athens and the M.Sc. (2011) and Ph.D. (2014) in Information Technology from the IT University of Copenhagen. He does research on the crossroads of game design, artificial intelligence and computational creativity. More specifically, he explores the limits of computational input to the human-driven design process in computer-aided design tools. Beyond AI-assisted game design, his research pursuits revolve around procedural content generation, digital aesthetics, evolutionary computation, neuroevolution and constrained optimization. He has published over 50 international journal and conference papers in the aforementioned fields, and has won several awards. Moreover, he has led or participated in the design and development of several games of varying scope and for different target audiences, including two FP7 ICT projects.

Georgios N. Yannakakis is an Associate Professor at the Institute of Digital Games, University of Malta. He has received the Ph.D. degree in Informatics from the University of Edinburgh in 2006. Prior to joining the Institute of Digital Games, UoM, in 2012 he was an Associate Professor at the Center for Computer Games Research at the IT University of Copenhagen. He does research at the crossroads of artificial intelligence, computational creativity, affective computing, advanced game technology, and human-computer interaction. He pursues research concepts such as user experience modelling and procedural content generation for the design of personalized interactive systems for entertainment, education, training and health. He has published over 200 journal and conference papers in the aforementioned fields. His research has been supported by numerous national and European grants and has appeared in *Science Magazine* and *New Scientist* among other venues. He is currently an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES. He has been the General Chair of key conferences in the area of game artificial intelligence (IEEE CIG 2010) and games research (FDG 2013).