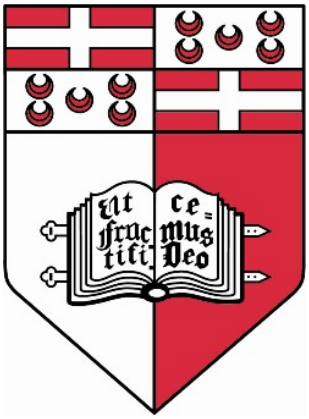
Generating Multifaceted Content in Games: A study on Levels and Sound



Phil Lopes
Institute of Digital Games
University of Malta

A thesis submitted for the degree of
Doctor of Philosophy

May, 2017

Abstract

Audio is an often overlooked aspect of digital games as it consistently remains unnoticeable to players entwining perfectly with the on-screen experience. Good sound design is considered to enhance the interactive experience, never overshadowing either the narrative or visuals of the game. Audio can be used to convey information about the world nearby (e.g. footsteps, voices), simulate real-world soundscapes (e.g. rain or wind), and enhance the experience of the on-screen drama and gameplay sequences. Narrative-heavy games, such as *Amnesia: The Dark Descent* (Frictional Games, 2010), rely heavily on emotional progressions that unveil during play. Players are tasked to push through situations that evoke fear, relief or even confusion, perfectly conveyed by the synchronization of both visual and audio facets; for instance, sombre music slowly fades in as players gradually step into an unlit room.

Procedural content generation is a popular field within digital game AI research, however most work tends to focus specifically on the creation of virtual spaces. This thesis argues that this type of generation can be quite limiting, especially for certain types of genres. By interweaving different facets in the content creation process, as is done in actual game development, can potentially provide a deeper and a more exciting experience for the players. This thesis introduces a system called *Sonancia*, an autonomous content generator capable of constructing horror levels and their respective soundscapes. A number of AI techniques have been used and tested for both the construction of levels that adapt to a user (or a machine) defined progression of emotion, and the creation of soundscapes that adapt to these emotional progressions. *Sonancia* is evaluated thoroughly via extensive user studies.

Acknowledgements

Alone this thesis would not have been possible, and for this I will be eternally grateful to everyone who helped me in this journey. I would like to first extend my gratitude to my supervisor Georgios N. Yannakakis and my co-supervisor Antonios Liapis, who were key in guiding my growth as a scientist and an independent researcher. Without their input and insight I believe that all of this would have been impossible.

I would also like to deeply thank my colleagues and my close friends Amy K. Hoover, Costantino Oliva and Hector Martinez, who were not only great friends who supported me during my PhD journey, but also motivated me and gave fundamental input for my research.

A big thank you goes to my friends Márcio Camisinha and Noelia Diaz Pita for offering me shelter during my early days in Malta, and being the best friends I could ask for during my stay. I would also like to extend my gratitude to Davide Nunes, Geraldo Nascimento and João Costa who not only listened to my ramblings, but also provided extensive input through thought provoking scientific and technological discussions.

A big thank you also goes to everyone at the Institute of Digital Games, particularly Ashley Davies, Daniele Gravina, Daniel Karavolos and Mirjam Eladhari for all the fun times, gameplay sessions and great discussions.

A deep thank you also goes to my parents and brother, including his beautiful family who have always supported my choices and without them I would not be where I am today.

Finally, I would like to deeply thank Joana Paramés who has been my support and source of motivation through this entire process. Supporting and motivating me through this difficulty journey, and having the patience to listen and giving me the strength and persistence to continue until the end.

Statement of Originality

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. (Louis Philippe Simões Castelo Branco Lopes)

To Joana, Lina, Sergio and Mike

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most impressive characteristics of digital games is its ability to transport players into a myriad of different fantastical worlds. Unlike other mediums, digital games provides a mechanism for effective interaction between the player and the virtual world, allowing for a wide range of playful spaces capable of sparking the imagination of players. Player agency provides for unique emotional opportunities due to their direct influence on the virtual world, its dangers and the adventures experienced within. In fact these experiences are meticulously crafted by experts from a wide range of different artistic domains, from writers, musicians, designers, to illustrators, where all of their artistic renditions are orchestrated into a seamless emergent gameplay experience. This is especially true within the genre of horror, where the fusion of sound and visuals provide an intense fearful atmosphere (Ekman and Lankoski, 2009). The creaking of footsteps on wood, a screeching wail in the distance or a simple animation of a door forcibly being shut behind the player, are all content orchestrations that impact the players emotional state. Several studies have suggested the development of computer based algorithms capable of autonomously generating such content, however these systems are often limited to one particular aspect of games (Togelius et al., Togelius et al. (2011); Shaker et al., 2015). This thesis argues that both the ability to generate and orchestrate different types of artistic content can provide a more meaningful player experience. In order to investigate our proposal, this thesis will explore the combination of two specific gaming artefacts: Audio and Level Architecture.

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 (Fahlenbrach, 2008). Sound design is an important part of both film (Sonnenschein, 2001; Fahlenbrach, 2008) and digital games (Collins, 2013; Gasselseder, 2014; Stevens and Raybould, 2013), 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 (Collins, 2013; Serafin and Serafin, 2004).

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. Procedural Content Generation (PCG) is an extensive area of game research, and is often used as an effective method to reduce content creation costs and increase game longevity (Togelius et al., 2011). Unfortunately, the research of procedural

content pertaining to both audio and other artistic domains of digital games is uncommon, likely due to the additional demands that sound often requires (Collins, 2013). Considering that digital games are multi-faceted creative domains (Liapis et al., 2014), where facets such as audio, visuals, levels and game mechanics work in conjunction to create interactive digital experiences, this thesis attempts to bring these domains closer by studying possible orchestrations between level architecture and audio within the context of horror.

The survival horror genre is unique in its heavy reliance on both sound and visuals, in order to convey powerful negative emotions (Ekman and Lankoski, 2009). It also focuses on exploration and hiding as players have limited combat ability, e.g. no weapons or limited ammunition (Perron, 2009). These complex characteristics of player affect raise important challenges for the generation of levels and soundscapes, where the objective is evoking these types of emotions within players. For instance, can a level generator anticipate and influence the affective state of a player, while consistently balancing feelings such as stress and relief; or how players navigate through a level under the effects of stress caused by previously encountered monsters. Framing this thesis within the domain of horror allows for a gameplay context where visual and audio fidelity are both influential characteristics of the genre.

1.1 Motivation

The development of Procedural Content Generation (PCG) algorithms is an active field of research in academia, where several domains of digital games have been explored (Togelius et al., Togelius et al.; Shaker et al., 2015). However, a common characteristic within the field is to generate singular faceted content, where level architecture is by far the most popular type (Shaker et al., 2012; Cardamone et al., 2011; Togelius et al., 2007; Liapis et al., 2013a). Even though digital games are multi-faceted experiences as suggested by Liapis et al. (2014), this is rarely taken into account in the PCG research community. Approaches such as the Angelina system by Cook et al. (2012) prove that the diverging facets of games are just as important, where the interplay between visual, audio, gameplay and narrative are key to the experience. In fact, this thesis argues that a procedural generator can be an orchestrator of game content, orchestrating human authored assets in order to produce an interesting interplay between the diverging artistic domains, while also allowing the system itself to contribute with it's own creations. Figure 1.1 showcases the proposed method of this thesis for multi-faceted procedural content generation. To effectively combine diverging types of artefacts we propose that an intent, or more precisely an objective, is used as an input for the generator, a formalization of what type of experience, or theme is desired. Thus we propose a solution inspired by the work of Colton et al. (2011); Colton (2012); Cook et al. (2012), where a framing device guides the generation process, formalizing the desired intentions of a designer (or another system), effectively informing the generator on: which artefact types are required; how to construct the machine authored content; and how to combine the diverging human and machine authored artistic creations. In order to accomplish this, such a system would require a degree of autonomy and decision making, so that content can be effectively merged based on the intent defined. This thesis proposes a fusion between search-based approaches and machine learning techniques in order to solve this problem, where machine authored content and orchestration is provided by the first methodology, while machine learning provides context based information respectively to each human created artefact. This way the search-based methodology can utilise the

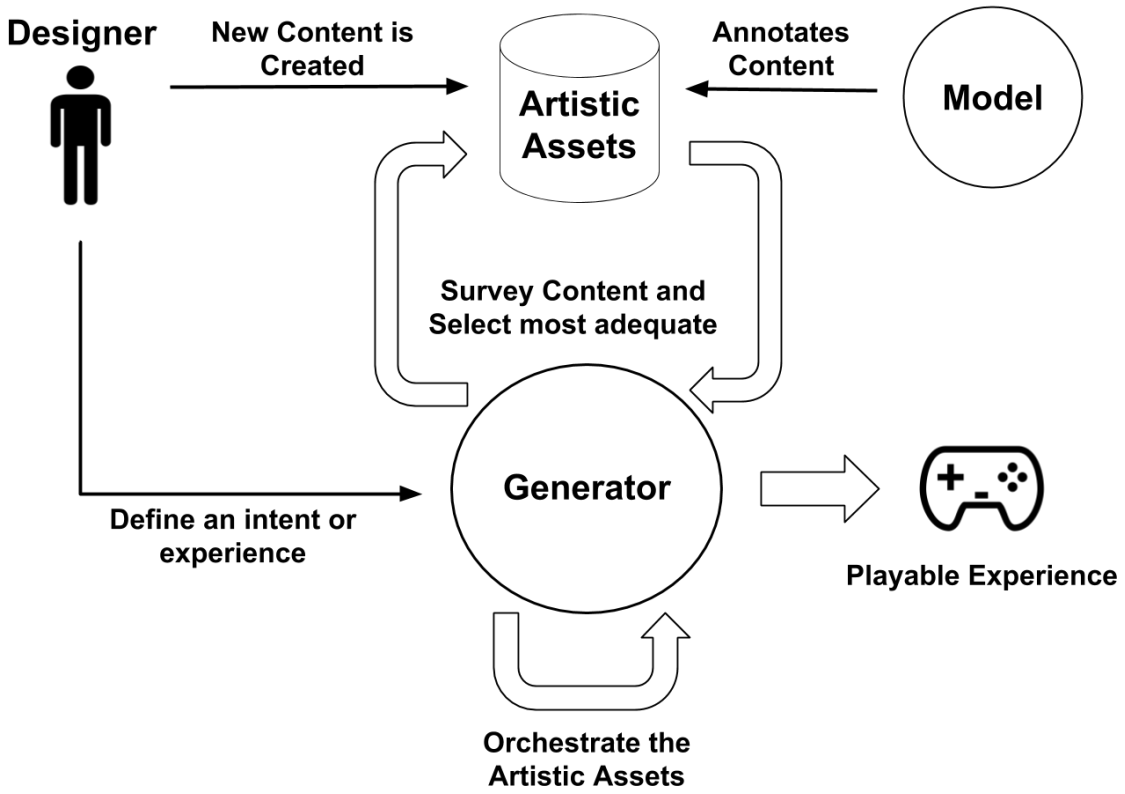


Figure 1.1: A Multi-Faceted Content Generator as suggested within this thesis. The designer intent is the "*blueprint*" that allows the generator to orchestrate the different domain artistic artefacts. Artefacts are surveyed from a global repository, where each artefact is annotated by machine learned predictors. The system then orchestrates and generates the remaining necessary content from the selected artistic artefacts. The output consists of a playable experience based on both the designer intent and the asset annotations.

information provided by machine learned predictors to more effectively make judgements on how to orchestrate human authored content.

A system such as the one proposed is a highly ambitious endeavour, going beyond the resources available for the realisation of this thesis. Thus, we propose a case study exemplifying the viability of such a system. The system proposed in this thesis, called *Sonancia*, is a multi-faceted generator, where both level architecture and audio facets are explored. Constraining the system to two facets, both simplifies the problem and at the same time allows to effectively study the interplay between level generation and audio, a concept rarely explored within PCG literature (Shaker et al., 2015). Furthermore, such a system also satisfies the idea of combining machine authored levels with human developed audio assets.

Given the focus on both audio and level architecture, the developed system was specifically framed towards the construction of content within the horror genre. Due to the genre's heavy reliance on visuals and audio (Perron, 2009; Perron, 2004; Ekman and Kajastila, 2009; Ekman and Lankoski, 2009), where the combination of level and audio interplay, in addition to the specific emotional progression the genre is known for, provides an interesting challenge for a multi-faceted generator. Thus we propose that as an important first step

towards the realization of multi-faceted procedural content generation, this work will explore if a generative system can actually construct a multi-faceted level where its emotional progression can match the actual emotional experience of a human player.

1.2 Solving the Problem

To accomplish the task suggested several challenges must be addressed. First a method for formalizing the designer intent is necessary, where the system must also be capable of interpreting it and subsequently construct the content based on what was described. Thus, it is also necessary to define how this intent will be interpreted by the procedural generator. Emotional progressions are a fundamental characteristic of narratives, which has often been exploited in previous studies (Cheong and Young, 2008). This is also true for the genre of horror. Events are presented in a specific ordering so as to provide the audience with an emotional "roller-coaster" throughout the experience. Particularly in horror, a common technique in the genre is to slowly build tension until reaching a climactic point (Perron, 2009). Distilling the horror genre into an emotional progression effectively simplifies the usability of the system, where a designer can simply trace the emotional progression intended. This subsequently provides the system with a "blueprint" of intent with enough abstraction, allowing the generator to have a degree of flexibility to achieve the experience defined, while also providing a common link between diverging facets.

The second component necessary is how the intent is then translated into a fully playable experience. Considering the focus of this thesis is specifically the level architecture and audio facets, some game-play concessions were made, specifically the controls, style and the player's perspective (i.e. First-Person). The characteristics of a level such as its length, the placement of enemies and light sources are the principal components that are influenced and adapted towards the emotional intent. Audio is then subsequently placed in the architecture based on two parameters: 1) the intended tension intensity of an area within the level; and 2) the perceived affect of the audio asset. The former is derived by analysing the generated level's progression from start to finish, where tension rises and decays based on how many monsters and light sources are present. The latter is obtained through machine learned predictors capable of ranking the low-level descriptors of audio based on their perceived tension, arousal and valence.

Thus we reach the third necessary component. Modelling the affect of audio for the efficient orchestration between level and audio assets, and the capability to annotate new and un-seen audio pieces, allowing further audio assets to be added and used immediately by the generative system. To achieve this, a crowdsourcing solution was constructed in order to obtain human annotations of affect of diverging audio pieces. This data was then used to train our predictors, in order to find a relationship between features extracted from audio and the annotation data. This way the system is capable of autonomously making decisions based on two parameters, how tense an audio piece is perceived and how tense should the level be at that point in time.

The final component consists of validating the system with actual human players, which will serve as a baseline and an initial overview of the system's viability once all components are working in conjunction. To actually measure the emotional impact of each participant, psychophysiology theory is applied where each player's physiology is measured during play. Furthermore, this thesis will explore an innovative method for annotating affect in digital games, where players review their gameplay and annotate their perceived affect in real-time.

1.3 Contribution

Although the work presented in this thesis is framed towards the area of Procedural Content Generation, several challenges addressed within this work can be put forth towards other domains of research. In particular contributions to modelling the affect of audio, experience-driven systems and affective computing.

- **Methods for Multi-Faceted Generation**: a general approach for the construction of multi-facet generation is presented. A set of methodologies are put forth and subsequently evaluated within this thesis, in order to accomplish multi-facet generation. Furthermore, to the best of our knowledge, this is the first attempt at developing a multi-faceted system that combines search-based methodologies with machine learned predictors, for the construction of digital game content.
- **Methods for Formalising and Adapting Emotional Intent towards Procedural Content**: this thesis presents a set of methods and suggestions for the re-contextualization of designer intent into playable multi-faceted levels. Furthermore, the set of tools and the implemented methodologies used within this thesis are publicly available[1] for the purposes of re-usability and re-adaptation.
- **Methods for Constructing Audio Affect Models of Tension**: although previous studies within the music emotion recognition field provided extensive research on models capable of ranking the affect of audio. This thesis provides an additional contribution by constructing models capable of ranking tension, which to the authors' best knowledge has not been explored in previous research.
- **Methods for Construction of Audio Affect Models as Tools**: this thesis 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 offer an additional layer of sound autonomy for procedural content generation systems. Motivated by the lack of such a model for game sound design this thesis introduces a crowdsourcing methodology for deriving the computational mapping between sounds within the horror genre and their perceived affect. Models such as the ones constructed in this thesis 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.
- **Methods for Constructing Audio Affect Models of Soundscapes**: the field of music emotion recognition has often concentrated on the detection of emotions within contemporary and classical musical pieces. This thesis argues that these models can also be used on sounds with the intent of accompanying audiovisual experiences. Several preference learned models are built to predict the global rank of horror sounds across three affective dimensions. Although previous work has explored the construction of preference models that rank emotion in audio, to the authors' best knowledge such a model has never been constructed for sound intended to accompany audiovisual horror experiences.

[1] `https://goo.gl/3P06kJ`

- **Rank-Based Soundscape Audio Annotation Dataset**: for the construction of data-driven models presented within this thesis, a crowdsourcing system was built for the collection of rank-based human annotations of audio. A total of 1009 pairwise ranking annotations for tension, arousal and valence were collected and parsed. The dataset is currently publicly available[2] for future research and development of alternative solutions than those presented in this work.
- **Methods for Real-Time Annotation of Affect in Players**: annotating the perceived affect of participant gameplay is not a trivial task, where the length of a game session and memory decay can contribute to added noise influencing annotation. Thus, the digital game research community has turned towards real-time annotating solutions. This thesis presents an innovative approach for the real-time annotation of participant gameplay, and investigates how close this methodology corresponds to the ground-truth (i.e. participant physiology). Furthermore, like the remaining work developed in this thesis, the tool is also publicly available[3].

1.4 Publications

The work conducted within this thesis resulted in several peer-reviewed publications in journals and conference proceedings:

Journals

- Lopes P., Liapis A., and Yannakakis G. N.: *"Modelling Affect for Horror Soundscapes"*, The IEEE Transactions on Affective Computing, 2017.
- Liapis A., Yannakakis G. N., Alexopoulos C., and Lopes P.: *"Can Computers Foster Human Users' Creativity? Theory and Praxis of Mixed-Initiative Co-Creativity"*, Digital Culture & Education (DCE), 8 (2). 2016.

Conference Proceedings and Workshop Papers

- Lopes P., Yannakakis G. N., and Liapis A.: *"RankTrace: Relative Annotation Tells You More About Your Ground Truth"* (under review).
- Lopes P., Liapis A., and Yannakakis G. N.: *"Framing Tension for Game Generation"*. Proceedings of the International Conference on Computational Creativity. 2016.
- Lopes P., Liapis A., and Yannakakis G. N.: *"Sonancia: A Multi-Faceted Generator for Horror"*. Proceedings of the IEEE Computational Intelligence in Games Conference. 2016.
- Lopes P., Liapis A., and Yannakakis G. N.: *"Targeting Horror via Level and Soundscape Generation"*. Proceedings of the AAAI Artificial Intelligence and Interactive Digital Entertainment Conference. 2015.
- Lopes P., Liapis A., and Yannakakis, G. N.: *"Sonancia: Sonification of procedurally generated game levels"*. Proceedings of the 1st Computational Creativity and Games Workshop. 2015.

[2] `http://www.autogamedesign.eu/sonancia`
[3] `https://goo.gl/3P06kJ`

- Lopes P., and Yannakakis, G. N.: *"Investigating collaborative creativity via machine-mediated game blending"*. Proceedings of the AAAI Artificial Intelligence and Interactive Digital Entertainment Conference. 2014.
- Eladhari M. P., Lopes P. L., and Yannakakis G. N.: *"Interweaving Story Coherence and Player Creativity through Story-Making Games"*. Interactive Storytelling, pp. 73-80. Springer International Publishing, 2014.
- Lopes P., Liapis A., and Yannakakis G. N.: *"The C2 create authoring tool: Fostering creativity via game asset creation"*. Proceedings of the IEEE Computational Intelligence in Games Conference. 2014.

1.5 Summary of the Thesis

This thesis is organized as follows:

- **Chapter 2** surveys the relevant literature and the state-of-the-art of the diverging research fields this dissertation touches upon. Specifically related work on musicology theory, procedural content generation, computational creativity and audio affect modelling are presented.
- **Chapter 3** presents the details of all the algorithms used for the realisation of this dissertation. In-depth and detailed descriptions of preference learning, automated feature selection and evolutionary computation algorithms are provided.
- **Chapter 4** introduces the *Sonancia* system and describes the entire pipeline and its components. Details on the formalisation of designer intent, level generation and sonification are presented.
- **Chapter 5** analyses the parametric sensitivity of the proposed level generator for the *Sonancia* system. Several types of designer intent are extensively tested with diverging parametrizations. Furthermore, a fully autonomous system is also investigated, capable of also generating intent and extensively testing the flexibility of level generation.
- **Chapter 6** investigates the development of several audio affect models capable of ranking audio based on the perceived affective state, and how features were extracted from raw audio signals. An in-depth statistical analysis is provided from the crowdsourcing annotations, followed by a detailed analysis of the results obtained from different trained models.
- **Chapter 7** studies the viability of the proposed system through an extensive user study. The data collection process and experimental protocol are presented. The feature extraction methodology for both physiological signals and real-time annotations are subsequently detailed. A statistical analysis of the experimental results are then provided and discussed.
- **Chapter 8** summarizes the main findings of this dissertation and provides both a reflection of the methodologies proposed and a post-mortem of the presented work. This chapter also concludes the dissertation and offering some final remarks on the work provided.

1.6 Summary

This chapter put forth the core questions and problems that motivated the work described in this dissertation. Particularly the focus on multi-faceted content generation is presented due to a lack of such research within the PCG literature. Furthermore, the combination of diverging facets provides an interesting problem where such generators require orchestrating capabilities, in order to accurately combine the multi-domain assets for play. The final question resides on how this content is orchestrated, or more precisely what is the common thread that combines the diverging facets. Potential solutions to the aforementioned problems are then put forth through the suggestion of a hybrid procedural generator that combines search-based methodologies with machine learning predictors. This combination allows content to be generated and effectively orchestrated by the search-based algorithms, which use artefact information provided by machine learned models. In order to guide the orchestration process this thesis proposes a framing device of designer defined intended experience, allowing the generator to adhere content towards this specific goal. In order to test this theory a simplified version of such a system, restricted to both the level architecture and audio facets is proposed. An in-depth description of this proof-of-concept system is subsequently presented. A list detailing the contributions of this dissertation are then provided, suggesting how the different aspects of this work can contribute to research and tools within the domains of PCG, audio affect modelling and affective computing. Concluding the chapter a brief summary of resulting publications, derived from the work realised in this dissertation, and the different chapters of the thesis are given.

Chapter 2

Related Work

Digital games are interactive and multi-faceted experiences, where the combination of diverging content types interplay between each other, in order to construct challenging and engaging experiences. These facets can range from creative artefacts such as audio recordings, 2- or 3-dimensional art, narrative texts or virtual environments; to rule-based systems that dictate the possible player interactions and how the virtual world responds to players. This thesis investigates an initial study towards the autonomous construction of multi-faceted content, by focusing specifically on two diverging facets: audio and the virtual environment. More specifically the autonomous construction of levels where audio is subsequently orchestrated, also autonomously, within these generated levels. Furthermore, in order to contextualize the multi-faceted generator this study is framed within the genre of horror, allowing the system to mould the content it generates towards a designer defined progression of tension. Thus, this thesis takes into account the literature from several diverging areas of research ranging from musicology theory, procedural content generation systems, computational creativity and affective computing. This section surveys the relevant literature in the context of this thesis within each of these fields of scientific research.

Given this focus on audio, section 2.1 offers a survey of musicological literature. The importance of audio within film and digital games is first explored, with a particular focus within the genre of horror. This survey gives insight to the importance of audio within both the genre of audio and in the medium explored in this thesis. It also describes the importance of abstract sonority, which enhances the emotional impact of certain scenes and narrative elements.

Autonomous creative systems has been a core concept within the computational creativity community, since its inception. It is a widely debated concept, where the main ideology is the simulation of creativity, or more specifically, the act of creative expression from a computational system. The community has recently turned its attention towards digital games, due to this mediums reliance on several creative artefacts. *Sonancia*, the work presented in this thesis, is built specifically around the idea of an autonomous creator capable of interpreting the intent of a designer, or a literary style, for the construction of virtual environments and accompanying soundscapes. Section 2.3 will offer a closer look on the relevant literature within the computational creativity community, and the concepts utilised for the creation of the system presented in this thesis.

Procedural content generation (PCG), has often revolved around the procedural construction of virtual environments, ranging from racing tracks to 2-dimensional levels for the game *Super Mario World* (Nintendo, 1990). Section 2.2 offers an overview of several

methodologies previously explored in academia and the digital game industry, from constructive systems to search-based methodologies, or more recently the personalization of content towards a player's affective state. In order to apply autonomous level construction this thesis explores key concepts and methodologies previously investigated, by utilising a search-based algorithm for the adaptation of content towards defined progressions of tension.

Emotions such as anxiety or stress are commonly preyed upon within the genre of horror. This thesis in particular explores how a multi-faceted level generator can utilise abstract representations of emotional progressions for both the construction of a level and its subsequent audio orchestration. For the latter this thesis employed several methodologies commonly investigated within the affective computing field, in particular within the area of music emotion recognition. In the same vein of other affective computing work, music emotion recognition consists of investigating the relationship between low-level features of raw audio signals and their perceived emotional impact. Particularly, in the context of this work, such systems can be particularly helpful for audio orchestration, where sounds are chosen based on a statistical model of perceived emotional impact. This allows the system to be flexible and dynamically tag each audio piece based on their perceived affect. Section 2.4 overviews the related literature within the field of music emotion recognition and to a lesser extent affective computing. By utilising such methods we hypothesize that a statistical model could potentially represent the designers intent more accurately.

2.1 Ludomusicology: Audio in Digital Games

Audio is commonly dismissed within the digital game medium, often overlooked in favour of other more appealing facets such as visual fidelity. It is often said that good audio design is the one that often goes unnoticed, where it works in tandem with the player interactivity and the on-screen visuals in order to enhance their emotional impact. Although this is simply the nature of how audio is good at seamlessly blending into the overall experience, it does have several important and diverging applications within digital games (Collins,2013). For the purposes of clarity when referring to audio within the context of this thesis, we refer specifically to *pieces of recorded sound that can be reproduced digitally.*

The majority of digital games already display a form of "procedural sound", where player actions determine what sounds or music the game should play (e.g. players firing guns in the background of a multi-player shooter) or through the simple actions of non-player characters that inhabit that virtual world (Garner and Grimshaw, 2014). This thesis however argues for a new approach of coupling sound with game levels which better orchestrates the two with the specific aim of enhancing the player experience and immersion. Collins (2013) argues that the visual aspects in digital games tend to be the player's principal focus during play, while sound tends to work at a more subconscious level. Several studies also suggest that picking the appropriate sound (i.e. what to play) and placing it at the appropriate moment (i.e. when to play) can significantly enrich the player experience as diegetic and non-diegetic sounds have a tendency of helping players immerse themselves into the virtual world (Gasselseder, 2014). This is a core reason why more audio solutions need to be explored within procedural content generation.

For the purposes of this thesis, which deals specifically with the concept of audio within the digital game space, this section commences by providing a brief survey of several audio methodologies utilised for player interaction, narrative and world building. In order to explore the concept of multi-faceted procedural content generation, we chose to frame the

study specifically within the horror genre. This section will delve into several concepts of musicology theory within the horror genre, detailing the impact and importance of audio as a method of creating tension for the genre. Furthermore the section explores the concept of audio as more than "music", but as pieces of sonority with the objective of adding emphasis to the on-screen drama.

2.1.1 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 (Fahlenbrach, 2008), this technique is often used to emphasize certain emotions of characters or scenes towards the audience.

Fahlenbrach (2008) 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 (Ekman and Lankoski, 2009).

This thesis explores the construction 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.

2.1.2 The Sound of Horror and the Genre

Popular gothic novelist Howard P. Lovecraft once defined fear as the oldest and strongest possible human experience. The self-subjugation of fearful emotions is a common tendency among humans, where the genre acts as a stimulant for playful and a controlled exploitation of these primal emotions. Perron (2009) suggests that by exteriorizing these doubts and anguishes of death, the supernatural, and other transgressions, helps humans understand and express these feelings better. From fearful tales passed along generations, to written novels by popular authors such as Lovecraft, Edgar Allan Poe and Stephen King, the genre of horror has consistently transcended into more popular mediums, digital games being the most recent. With the advancement of both visual and audio fidelity, contemporary games such as *Amnesia: The Dark Descent* (Frictional Games, 2008), or even the *Resident Evil* series (Capcom, 1996–2017), have popularized the genre into a set of specific gameplay mechanics, where these games are often referred to as *survival horror*. Although different survival horror games have often presented new ideas to the genre, the player character

vulnerability has consistently been the genre's defining characteristic. Specifically, it consists of limiting a player's ability, where the player controls fragile characters who are particularly vulnerable and often offensively limited (e.g. no weapons or finite ammunition). This vulnerability adds to a heighten sense of danger, as subtle changes to the environment tend to have a higher impact on the players affective state. Ekman and Lankoski (2009) suggests that this vulnerability is what makes sound such an important part of this genre, since any significant changes in the environment is immediately paid attention to, even if it is beyond a player's line of sight.

Initial work by Garner et al. (2010); Garner and Grimshaw (2011) empirically investigated the impact of sound within the horror medium, pin-pointing several characteristics of audio that suggest a heightened sense of arousal and fear. Particularly high pitch and loud sounds have been suggested to have a higher impact on the elicitation of fearful emotions. The game *Amnesia: The Dark Descent* in particular often resorts to musical dissonance within a lot of the soundscapes utilised, a tactic often employed by popular film maker Alfred Hitchcock and Stanley Kubrick (Fahlenbrach, 2008). According to Perron (2004) sound works exceptionally well as a catalyst of foreshadowing future events, or in his own words forewarning potential threats in the environment. Famous examples of this concept include games such as *Silent Hill 2* (Konami, 2001) and *Aliens Vs Predator 2* (Monolith Productions, 2001), where an "alien detector" utilised by players is accompanied by blip noises indicating the threat of something nearby. These blips become increasingly more intense as the threat gets closer, forewarning the player that an attack is probably imminent. These simple noises are powerful methods of evoking player emotion, not for the particular noise itself but the context it represents within the game world.

In order to emulate soundscapes within the context of survival horror, a proof-of-concept game was specifically constructed to realize the theories and concepts explored within this thesis. The game dubbed as *Sonancia* (Lopes et al., 2015) consists of procedurally generating the architecture of a level, and subsequently placing audio assets according to the intent of an emotional progression defined by a designer. The game itself attempts to follow the majority of common principals as defined in the book of Perron (2009), in which player vulnerability is a key part of the experience.

2.1.3 The Perception of Audio in Horror

The horror genre is unique in its heavy reliance on sound to convey negative affective states such as shock, disgust, ecstasy, fear and relief (Ekman and Lankoski, 2009). However, one of the most interesting concepts of the horror genre is how it tends to play with the audience's (or player's) perception of sound within the environment (Ekman, 2005; Ekman and Kajastila, 2009; Kromand, 2008). The source of a sound is an important characteristic of how it is perceived within audiovisual media. The sound of gunfire or the screeching of tires, for example, are directly associated to real-world objects, and thus when heard in media such as film or digital games, it is perceived as sound emanating within the world these fictional characters inhabit. It is also common for such media to come accompanied with orchestral soundtracks, which are often composed purposely for the audience's experience, and thus they are not perceived as directly emanating from within the fictional world. Sound of the prior nature is commonly referred to as diegetic sound, while the latter as non-diegetic sound (Kromand, 2008).

According to Kromand (2008), horror games tend to utilise a combination of both diegetic and non-diegetic sounds, often challenging the barrier of sound perception. This

lack of perception can cause players to distrust between what is effectively "real" within the virtual environment, and what is simply non-diegetic sounds masking as diegetic. The confusion that arises from these situations are fully exploited within the genre, and allows for the orchestration of intense ambient filled environments, leaving the player to doubt and question every move ahead leading to more emergent fearful situations in-game. This fact has been intensively studied by Ekman and Kajastila (2009); Ekman (2005), where sound cues that were localized outside of the virtual world, tended to have a higher impact on the perceived "scariness" of the sound than sounds with actual visible sources. Although the work by Ekman and Kajastila (2009) did use a limited number of subjects, results still suggest that the perception, and not just the sound itself, also contributes to the perceived emotion felt by sounds during play. A common technique used within digital games is masking, where a diegetic sound is used to inform the player of a non-diegetic event. Commonly these types of techniques are used in conjunction with triggering mechanisms, which once triggered by players a particular sound emanates (e.g. such as a growl of a monster), thus informing the player of imminent dangers or events.

For construction of soundscapes within the system presented in this thesis, a combination of both diegetic and non-diegetic sounds were taken into consideration. Although, the bulk of the sound manipulated by the system is non-diegetic, enemies roaming the levels will emanate sound through footsteps and different growl types. These types of sounds allows the game to inform the player on enemy position or if they have been spotted by a particular monster. Non-diegetic is used by the system, in order to personalize procedural level generation towards an intended ambience set by the designer. These sounds are placed based on their predicted affect, obtained from machine learned predictors, and the intended tension required for that particular section of a level.

2.2 Procedural Content Generation

Procedural Content Generation (PCG) is a popular technique that has been consistently used in digital games for over 30 years, with games such as *Rogue* (Toy and Wichman, 1980) being the earliest example. Although early interest of PCG was predominantly about constraining the disc size of digital games to a minimum, other approaches exploited the procedural nature of such algorithms by adding stochastic parametrizations in order to consistently generate diverging mazes. More recently PCG has been a focus of academic interest, including the development of alternative PCG approaches (Togelius et al., 2011) and the adaptation of content to specific player experiences (Yannakakis and Togelius, 2011). Level generation is often the focus of the majority of PCG research, with notable examples including the generation of 2D platform levels (Shaker et al., 2012), real-time strategy maps (Liapis et al., 2013a), racing tracks (Togelius et al., 2007), first-person shooter maps (Cardamone et al., 2011), among others (Shaker et al., 2015). The horror genre is no exception to PCG; games such as *Daylight* (Zombie Studios, 2014) procedurally generate levels and enemy positions, while the AI director of *Left 4 Dead* (Valve, 2008) procedurally spawns zombies according to a tension model, based on the difficulty parameter and the players progression within the level. This thesis draws inspiration from the tension model of defined in *Left 4 Dead*, as the level structure and monster placement are generated based on a designer-specified progression of intended tension.

Even though audio within digital games can already be thought as a form of procedural audio, as stated in the work of Garner and Grimshaw (2014), this thesis argues that more

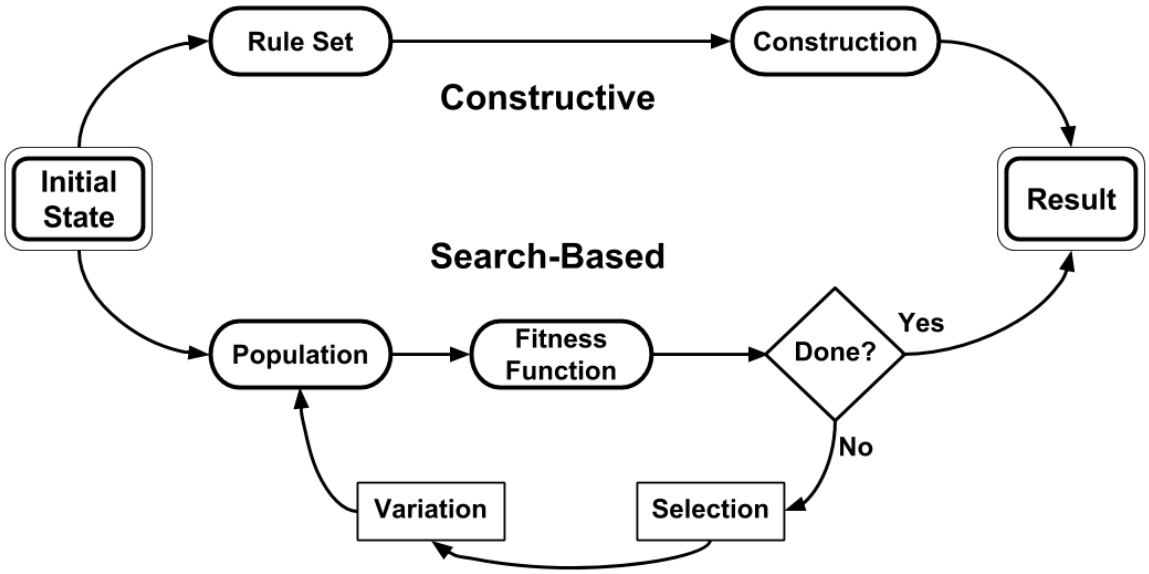


Figure 2.1: Two diverging types of PCG algorithms adapted from the work of Togelius et al. (2011). Constructive PCG is a rule-based algorithm where content is constructed based on a pre-defined set of rules allowing for diverging patterns to emerge. Search-Based PCG consists of defining wanted characteristics of content (i.e. the fitness function), allowing the algorithm to search for level combinations that satisfy these characteristics.

could be accomplished through orchestrating between virtual levels and the sounds played therein. Several professional tools such as the sound middleware of *UDK* (Epic Games, 2004) already some form of procedural sound components, albeit very simple (i.e. variations of notes in a specific scale). This does suggest an increased commercial interest in sound as a procedurally generatable game facet. On the other hand, games such as *Audio Surf* (Fitterer, 2008) and *Vib Ribbon* (Sony Entertainment, 2000) have previously focused on music-driven level generation, where the characteristics of the music influence the level generation. *Proteus* (Key and Kanaga, 2013) explored several ideas on how spatial positioning, visuals and player interaction affected and influenced sounds played in realtime. *AudioInSpace* by Hoover et al. (2015), is another example that combines both gameplay and audio within a side-scrolling space shooter that evolves its shooting mechanics based on the music playing in the background, which is pre-selected by the user or procedurally generated via artificial evolution. The work of Scirea et al. (2014) also investigated how music could be procedurally generated in order to convey narrative foreshadowing in digital games.

This thesis will instead concentrate on developing methodologies capable of generating horror game levels and their corresponding sonification, based on both designer intent and diverging machine learned predictors of tension constructed.

2.2.1 Constructive-Based Procedural Content Generation

It is important to firstly define the difference between constructive-based and search-based PCG. In this thesis we apply the definition of Togelius et al. (2011) for both methodologies. Figure 2.1 showcases the diverging steps necessary for generating content through both constructive and search-based methods. Although the work presented in this thesis specifically

uses a search-based methodology, for the sake of completeness this section will offer a brief survey of constructive based solutions within both industry and academia.

Popular games such as *Spelunky* (Mossmouth, 2008) and *Minecraft* (Mojang, 2011) have been heralded as pioneers of the contemporary resurgence of PCG within the digital game industry. *Spelunky* in particular uses a matrix style representation where each cell represents a possible level component space. Levels are built sequentially in order to guarantee a clear path towards the exit, where each piece is a pre-designed variation of level subsections. The first version of *Minecraft* on the other hand generates maps using a Perlin noise algorithm (Perlin, 1985), which defines the various heights of the map. Based on this height a rule-based system called biomes, defines the climate and subsequently the type of area each section of the map will be, e.g. mountainous region, grass plain or river.

Shaker et al. (2016a) defines several constructive methods that have been heavily used within academic research. The first method described consists of a space partitioning solution, where sections of a hypothetical map are sub-divided using binary space partitioning (BSP). In each partition two random coordinates are chosen in order to define a room structure, and each connections of the BSP tree defines how rooms are subsequently interconnected. The work of Johnson et al. (2010) utilised the simple rules of cellular automata algorithms for the construction of complex "cave-like" level structures. The usage of various techniques popularized within the computer graphics community have also been widely used for the generation of procedural content (Shaker et al., 2016b). Similarly to *Minecraft*, noise or fractal based algorithms are commonly used for the construction of terrain with organic heights and valleys. Indicatively, the work of Dormans (2010); Dormans and Leijnen (2013) has concentrated on exploiting grammar-based rules for the generation of levels. Further, L-Systems have been extensively used within the industry for procedurally generating foliage, such as trees in the popular tool *SpeedTree* (Interactive Data Visualization, Inc., 2002).

2.2.2 Search-based Procedural Content Generation

Although Togelius et al. (2011) attempts to provide a more general definition, in practice search-based PCG has been heavily dominated by evolutionary approaches. Unlike the previous section, where several constructive based algorithms were presented, this survey will review mostly how search-based approaches were used in different digital game domains.

In recent years one of the main focuses of search-based PCG was in the effective generation of 2D platformer levels. The work of Shaker et al. (2012) in particular explored how search-based PCG could be applied for the generation of a variant of *Super Mario World* (Nintendo, 1992) levels. Furthermore the work of Togelius et al. (2007, 2010a,b) explored how search-based methodologies could be applied for the procedural construction of race tracks and real-time strategy maps. More specifically, they explored how a multi-objective search-based approach can be used to generate maps for games, whose balance of space and resources between player opponents are a fundamental aspect of the game. Search-based methodologies have also been used for mixed-initiative tools, aiding designers in the construction of content. Particularly the work of Liapis et al. (2013a,b) focused on aiding designers by providing real-time feedback during the creation process, and allowing designers to generate levels based on a specific set of balancing constraints, which could be further customized by the designer posteriori.

Beyond the generation of levels, search-based PCG has also been used to construct the actual rules of games. Notable examples include the work of Hom and Marks (2007); To-

gelius and Schmidhuber (2008) and Browne (2008). In Hom and Marks (2007) the rules of 2-player games were evolved using a game balancing metric as their fitness function. The search space was constrained to popular 2-player games such as Tic-Tac-Toe, Checkers and Reversi. The Ludi system by Browne (2008) also focused specifically on boardgame generation. Game rules are represented through a complex system of tree type data structures, allowing Ludi to construct a wide variety of game rule combinations. The evolutionary process evaluates the game through a series of game mechanic measures and gameplay simulations, capable of measuring the quality of a game. Outside the spectrum of boardgame rule generation, the work of Togelius and Schmidhuber (2008) specifically explored the concept for digital games, where rules applied to players, items, enemies, allies are generated. This allowed the system to construct several emergent gameplay behaviours with interesting and novel mechanics.

Lastly the work of Hastings et al. (2009) explored the use of a search-based approach for the generation of weapons for the *Galactic Arms Race* space combat game. To evaluate the generated weapons, the fitness function was directly tied into the game. Quality of the content was determined by the amount of usage from the player, more precisely by how many times the weapon was fired. This allowed the system to focus on key weapon characteristics preferred by the player as play progresses, in a way personalizing the weapon generation towards a player's preferred playing style.

For the purposes of this thesis a search-based approach was used for the generation of procedural horror levels. Evolutionary methods were chosen specifically due to their capability to optimize towards defined goals, while still providing a degree of flexibility on how those goals are met. In this thesis a formalization of the intended emotional progression of a level is defined by the designer. The level generation process uses evolutionary methodologies to optimize levels towards this formalization. More in-depth details are provided in both Chapter 4 and Chapter 5.

2.2.3 Experience-Driven Procedural Content Generation

According to Yannakakis and Togelius (2011), experience-driven PCG consists of adapting procedural generated content to the psychophysiology of players, such as measuring heartrate or the amount of skin conductance. The idea behind experience-driven approaches is that games can more accurately adapt content such as levels for example, based on the previous performance of players in addition to their emotional state. Although several types of games have been used in the field of affective computing in order to explore the viability of such systems, the types of research have varied substantially. One example includes using experience-driven systems to balance the difficulty of games such as Pong (Rani et al., 2005) or Tetris (Chanel et al., 2008) based on the player's affective state such as boredom, anxiety and frustration. Other examples include applying experience-driven methodologies to effectively model the players enjoyment while playing a racing game (Tognetti et al., 2010), or their arousal while playing a First-Person Shooter (Drachen et al., 2010). Its application on providing controlled player experiences has also been explored, either to ease the burden of control from the player through the automation of virtual 3D cameras in Third-Person Games (Yannakakis et al., 2010), or simply for therapeutic applications where the simulation must be able to detect emotional states and control the experience adequately, such as the treatment of post-traumatic stress disorder (Holmgard et al., 2013).

Due to the horror genre's focus on the exploitation of human fearful emotions, it has been a notable case study within experience-driven digital game literature. The work of

Nogueira et al. (2014, 2016) for example, explored several methodologies for the construction of player affect models for the horror genre. By re-purposing the game *Vanish* (3DrunkMen, 2013), a procedural generation system was built capable of constructing and placing monster assets at run-time, based on both the affect models and a player's skin conductance. This thesis attempts to build upon the knowledge obtained from Nogueira et al. (2016), by offering additional tools for the improvement of such systems. First by offering audio affect predictors capable of informing procedural generators on an sound asset's perceived affect, allowing for a more accurate orchestration between audio, level structure and AI enemy agents, which is an important component for horror (Ekman and Kajastila, 2009). Secondly by exploring alternative methods for autonomous systems or designers to define intended gameplay experiences, while maintaining a degree of variability between each level, as an attempt to improve gameplay replayability. Finally, we provide an off-line solution in comparison, where affective data is used apriori for the construction of machine learned predictors, which subsequently drive the level generator. Although Nogueira et al. (2016); Yannakakis and Togelius (2011) suggests that the on-line usage of affective information tends to provide the best and most accurate experience, due to its ability to adapt content towards the current affective state of the player, it is unrealistic to expect that affective data will consistently be available. For this reason we argue that alternative solutions are also beneficial to explore. Furthermore, it is also important to keep in mind that the work presented in this thesis is not mutually exclusive to either on-line or off-line solutions, where the re-adaptation of the suggested methodologies is entirely possible for potential future systems with a focus on horror.

Plans and Morelli (2012) also suggested the use of psychophysiology in order to guide how music is generated within digital games. Although research in this particular area is still in its infancy, several projects have shown the viability of such a system. AudioNode (Plans et al., 2015) consists of a proof-of-concept prototype where music is generated based on the anxiety of players, in which the system's main objective is to calm the player using a music generation system. The work by Scirea et al. (2014) on the other hand utilised a pure-data music generator capable of taking affective states based on Russell's circumplex model (Russell, 1980) and generating music accordingly. This particular system was used to foreshadow events in narrative heavy games.

The work of Sorenson et al. (2011) in particular explored the relationship between fun and anxiety, where the latter would consist of increasing the challenge of a particular level, while the former reflects on the positive and negative fun experience of a player, while manipulating this challenging aspect. In that work the authors attempt to model the theory of flow (Csikszentmihalyi, 2014), which implies that a balance of fun lies between emotional boredom (i.e. game is to easy) and frustration (i.e. game is to difficult). Although this thesis deals with the concept of emotional progression, we argue that not all experiences must stem within the concept of difficulty, even though both concepts might be closely related. Within the context of this thesis, emotional progression is seen principally as induced by the different characteristics that populate the level such as enemies, which to a certain extent challenges the player to progress forward; audio and different visuals of levels such as lighting. Thus, given the context of horror, this thesis argues that tension is a more descriptive term than difficulty, as it encompasses to a certain degree the challenging aspect, but also argues that other factors also contribute towards the emotional progression of players.

2.3 Computational Game Creativity

The Computational Creativity (CC) community has delved within a multitude of diverging artistic fields, from the creation of poetry (Veale, 2013; Colton et al., 2012), jokes (Ritchie, 2001), paintings (Colton, 2012) and even music (Eigenfeldt and Pasquier, 2013; Eigenfeldt et al., 2012; Hoover et al., 2012).

Although in recent years digital games have gained traction within the CC community, the potential of this medium is still in its infancy within the community. Comparatively to other creative mediums digital games are a combination of different types of artistic artefacts, that complement and work in conjunction to create an emergent interactive experience. According to Liapis et al. (2014) six diverging types of creativity facets have been identified in games: visuals, audio, narrative, ludus, level architecture and game-play.

The *Sonancia* system built for the realization of this thesis is considered in many parts as an autonomous creator, capable of integrating designer intent for the construction of levels and its soundscapes. Thus, this section will provide an analysis of digital games through the perspective of computational creativity.

2.3.1 Game Facet Blending

Conceptual blending, first proposed by Fauconnier and Turner (2008)), is the human act of metaphorically "blending" various distinct concepts, which then result in a brand new structure with its own emergent properties. More precisely blending can be thought as a mapping between different conceptual spaces resulting in a emergent space, such as the blend of a spoon with a fork, for example resulting in a spork/foon, or of a house and a boat, leading to the word houseboat, which can be thought of as concepts of their own. The notion of blends have been extensively used within computational creativity theory (Veale, 2012; Li et al., 2012) as a way of obtaining further emergent artefacts through the combination of concepts.

Several computational creative systems have stood at the interplay of different multidisciplinary creative domains. It should not come as a surprise, therefore, that several projects in computational creativity tackle the transformation of data from one domain to another, e.g. images to soundscapes (Johnson and Ventura, 2014), news articles to collages (Krzeczkowska et al., 2010), academic papers to songs and their lyrics (Scirea et al., 2015), text descriptions to player abilities (Cook and Colton, 2014), to name a few. Due to the dissimilarities between source and target creative domains, such computational systems must learn to creatively interpret the patterns of the input, and work towards making them apparent in the output while still obeying the constraints and the expressiveness of the target creative domain (e.g. a limited colour palette).

Digital games are a medium combining different creative facets that complement each other to create specific kinds of interactive experiences (Liapis et al., 2014). Beyond the creativity included in designing content for each facet, blending the different facets is a research direction of utmost challenge and promise within computational creativity (Lopes and Yannakakis, 2014; Liapis, 2014). Game generation systems like *Angelina* (Cook et al., 2012) and *Game-o-matic* (Treanor et al., 2012) extensively explore how different facets of games can be combined to create interesting and thought-provoking experiences. Commercial games (designed and fine-tuned by humans) tend to blend either their rules (ludus) or level design (architecture), in the case of e.g. action-RPGs or multiplayer online battle arenas. However, preliminary suggestions for automating such blends creatively have

been put forth by Gow and Corneli (2015). Blends between audio and gameplay have been explored in *AudioInSpace*, where the shooting mechanics of a side-scrolling space shooter change according to the background music, which can be hand-authored (loaded from a music library) or artificially evolved (Hoover et al., 2015).

With this work we intend to further explore the mapping along the multiple facets of creativity existent in games. We hypothesize that the exploration of blended mappings with relation to music, narrative, ludus and level architecture can be highly beneficial towards the construction of automatically generated game systems. Audio in particular has the potential of enhancing the player experience and effectively immersing the player within the virtual world (Collins, 2013). This property is especially important within the genre of horror in which particular audio patterns such as musical foreshadowing, the absence of noise, or even a rise of tempo, volume and pitch can elicit stressful experiences for players (Garner et al., 2010; Ekman and Lankoski, 2009). These audio patterns can be successful in eliciting intense affective responses if they are well interwoven with the design of game levels.

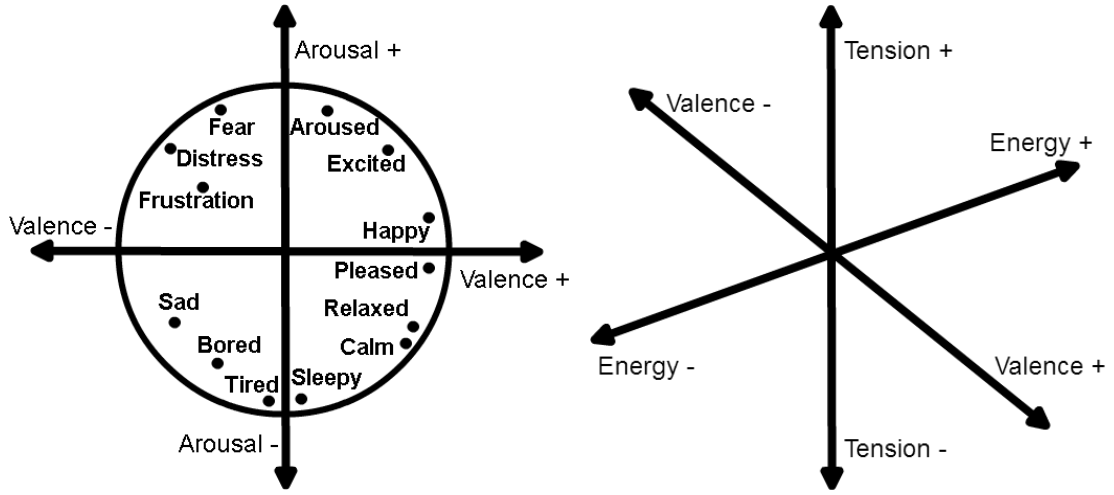
2.4 Modelling the Affect of Audio

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 (Ekman, 1992; Zentner et al., 2008), alternatively others have argued that dimensional approaches to emotion representation are superior (Eerola and Vuoskoski, 2010; Daly et al., 2014; Trochidis and Bigand, 2014).

According to discrete models, all emotions can be derived from a limited set of universal emotions, such as fear, anger, disgust, sadness and happiness (Ekman, 1992; Trochidis and Bigand, 2014), 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 (Balkwill and Thompson, 1999; Gabrielsson and Juslin, 1996). 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 (Zentner et al., 2008): wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness. According to Eerola and Vuoskoski (2010), 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 (1980) circumplex, where emotions are represented as two dimensional planes (see Fig. 2.2a): arousal (activation-deactivation) and valence (pleasure-displeasure). Alternatively, the Thayer (1989) model proposed a variant to the Russell circumplex, and argues that both dimensions are actually "tense arousal" (Arousal) and "energetic tension" (Valence). Schimmack and Grob (2000) 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. 2.2).

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 Schimmack and Grob (2000) model. 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



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

Figure 2.2: Russell's circumplex (Fig. 2.2a) 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. 2.2b) 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.

arousal dimensions.

Emotion recognition in audio is an active field of research (Kim et al., 2010; Aljanaki et al., 2015; Saari et al., 2016); 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 emotional models (Eerola and Vuoskoski, 2010) or investigating the variations of affect across multiple genres (Eerola, 2011), 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 (Kim et al., 2010; Yang and Chen, 2011b). In this thesis, instead, we offer 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 (2000) model.

2.4.1 Learning to Rank the Affect of Audio

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 (Martinez et al., 2014; Yannakakis and Hallam, 2011; Yannakakis and Martínez, 2015), when compared to rating-based (e.g. Likert (1932) scale) 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 (Yannakakis and Martínez, 2015; Martinez et al., 2014). 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 (Li et al., 2013) or the annotation of a subjective experience such as game aesthetics (Shaker et al., 2013). This thesis employs a rank-based crowdsourcing approach with the aim of soliciting human pairwise ranks between sound samples of a horror sound library. 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.

Preference Learning (PL) is a supervised learning methodology, where the goal is to derive a global ranking function from a set of annotated ranks (Fürnkranz and Hüllermeier, 2011). PL for affective modelling was introduced by Yannakakis et al. (2009) and has since then been used extensively within the domain of affective interaction, for e.g. personalizing game levels (Shaker et al., 2013) and for affect-driven camera control (Yannakakis et al., 2010). Rank Support Vector Machines (RankSVM), a variant of SVMs, was introduced by Joachims (2002) 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 (Lotfian and Busso, 2016) and musical pieces (Yang and Chen, 2011a). Within the domain of audio, Yang and Chen (2011b) used preference learning for music emotion recognition. RankSVMs were used to rank different musical pieces — represented with Russell (1980) circumplex model of affect —- based on low-level audio descriptors commonly extracted in music information retrieval. Inspired by the success of RankSVM affect models in music, this thesis trains a number of RankSVM models and tests 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 and Chen (2011b) and extend it in the domain of sound (within games and beyond) through a crowdsourcing approach. Artificial Neural Networks (ANNs) have also been a popular algorithm within the machine (preference) learning community, particularly outside of the domain of music, rank based ANNs have been extensively used to model player behaviours (Martínez et al., 2011; Yannakakis et al., 2010). Considering the efficiency of ANNs within other affective domains, we consider this methodology a suitable alternative to RankSVMs. Thus, in order to compare the prediction efficiency between both methods this thesis constructs audio affect models using both RankSVM and ANNs to showcase their efficiency in modelling affect of soundscape audio.

Furthermore, for measuring affect this thesis goes beyond the arousal and valence dimensions. An emphasis on modelling the affective dimension of tension is also presented, where the focus is 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.

2.5 Summary

This thesis proposes initial methods for the realization of multi-faceted procedural content generation, between level architecture and audio, by translating the emotional intent of a designer into playable experiences. This chapter presented four diverging perspectives applied for the realization of this work. The musicological perspective is discussed first, presenting the theoretical background of the impact of audio on both film and digital games. The PCG perspective subsequently follows, surveying the diverging methodologies of content generation applied in both academia and the digital game industry. A background on Computational Creativity is presented next, outlying the perspective of generative systems as autonomous creators, the blending of different creative domains and the definition of intent. Lastly the audio affect modelling perspective is showcased, surveying literature in both affective computing and machine learning.

Chapter 3

Algorithms

This thesis explores the creation of multi-faceted game content by adapting audio to procedurally generated levels for the horror genre. This chapter outlines a set of computational methods used for the creation of said multi-faceted content. First detailing the machine learning algorithms used for the creation of a model capable of ranking sounds based on their perceived emotion and secondly describing genetic algorithms used for the creation of procedurally generated horror levels. Figure 3.1 presents a schematic of the various interconnected methods used in this work.

This thesis argues that the connection between sound and the virtual space is an important component for the creation of interesting and engaging player experiences. Although soundscapes are often associated to the simulation of environmental audio (e.g. sound within a stadium), modern games such as *Doom* (id Software, 2016) have extensively used audio to specifically enhance the player experience. Sound designers tend to meticulously place audio assets throughout a level such as to create a sense of foreshadowing of upcoming events, or for the creation of impactful situations that arise within the level (Stevens and Raybould, 2013). In this thesis we intend to simulate this through the construction of computational predictors capable of ranking diverging sounds based on different perceived emotions. These predictors are then used to assist the level generation process, by choosing which sounds align with the intended emotional progression of the procedurally generated level. Section 3.1 describes the supervised learning algorithms utilised within this thesis: the Rank Support Vector Machine (section 3.1.1) and Artificial Neural Networks with Backpropagation (section 3.1.1). Models are trained with the intent of learning the relationship between low-level statistical features extracted from raw audio signals, and user annotated audio preferences within the affect dimensions of tension, arousal and valence. The data collection process of user annotations is further described in chapter 6, section 6.1 of this thesis.

Feature extraction can often yield an overwhelming number of features, too large for predictors to derive meaningful relations within the feature space. This is certainly true for the audio feature extraction method utilised within this thesis, which is further described within section 6.2. By applying automated feature selection algorithms (AFS), it is possible to reduce the amount of features to a more manageable number. These algorithms automatically experiment with diverging sets of features, such that it maximizes the prediction accuracy of the model. Section 3.2 describes the two AFS algorithms used within this thesis: the Sequential Forward Selection (section 3.2.1) and the Sequential Backward Selection (section 3.2.2).

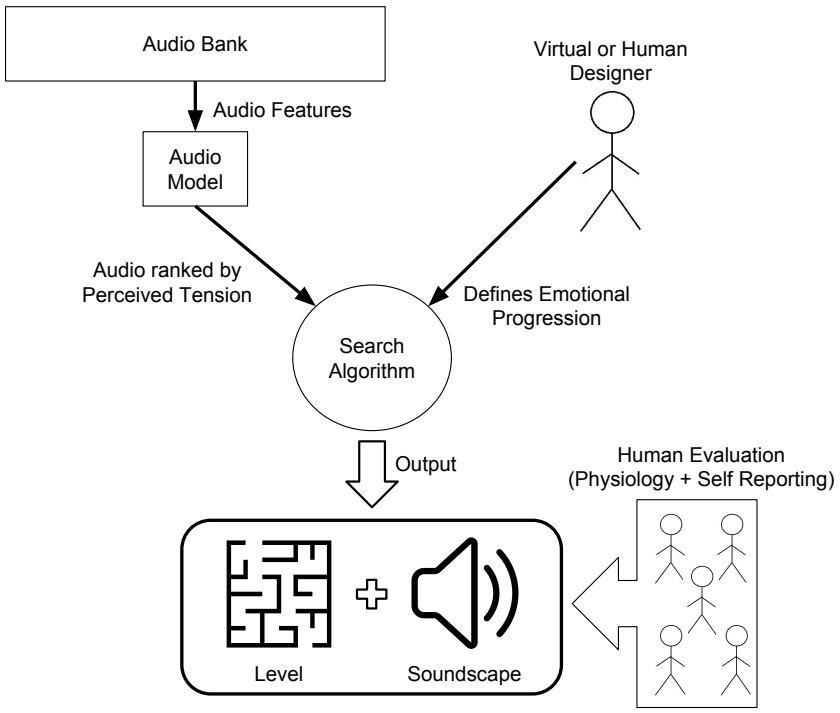


Figure 3.1: Overview of the underlining system presented within this thesis. Audio models are created capable of ranking audio through the mapping of low-level features of audio signals and different affective states. Although 3 different models of affect were created (i.e tension, arousal and valence models) only the tension model was used in the final system. A search based algorithm is used to efficiently generate a level structure and tie audio according to an intent defined by a designer. Multi-faceted levels are then created and evaluated by humans through physiological monitoring and self-reporting.

For the construction of procedural generated levels a search-based approach was chosen, due to its previous successes within the space of procedural content generation (Togelius et al. (2011)). This particular approach also allowed for the creation of fitness functions capable of defining intended emotional level progression. Evolutionary computation allows content to be adapted towards these emotional progressions, while still retaining a degree of novelty since different spaces can also provide the same emotional experiences. Within this work evolutionary computation was utilised for both the creation of system defined experiences and for the generation of levels. Section 3.3 offers a brief overview of genetic algorithms (GA) implemented within this work, while section 4.3 of chapter 4 details how genetic algorithms are adapted for the construction of *Sonancia* levels.

3.1 Machine Learning

Machine Learning (ML) is a field within artificial intelligence that focuses on the development of algorithms that can learn and predict from large-scale data (Mitchell (1997)). The main concept behind the majority of ML algorithms is that trends within data can potentially be learned through performance error minimization. This is achieved by comparing the algorithms prediction accuracy to the actual desired output, and then consistently

adjusting itself in accordance to the obtained error. ML can be performed on labelled (supervised learning) or unlabelled (unsupervised learning) data, or by rewarding/punishing its actions within the acting environment (reinforcement learning).

This work has taken a supervised learning approach for the construction of models capable of ranking audio descriptors according to each affect dimension of the Schimmack and Grob (2000) model. In particular two ML algorithms were explored: Rank Support Vector Machines and Artificial Neural Networks; which are described in detail below.

3.1.1 Preference Learning

Preference Learning (PL) is a ML area, which consists of algorithms capable of learning and predicting how to rank sets of data Fürnkranz and Hüllermeier (2011). Given a set of data samples (elements) and the pairwise comparison of each element within the set (partial ordering), the goal of PL algorithms is to derive a function which is capable of predicting the respective rank of each element. The main advantage of PL algorithms is its capability of learning from a set of preferred and unpreferred labelled items. This is particularly advantageous for the construction of affect models, which rely heavily on human annotated datasets for learning, and has proven to be effective (Yannakakis and Hallam (2011); Yannakakis and Martínez (2015); Yannakakis and Martinez (2015); Martinez et al. (2014); Yang and Chen (2011b)).

Within this work models are trained using object preferences, where each feature vector $q_i = [q_{i0}, q_{i1}, ..., q_{in}]$ of element q_i within the complete set X of possible elements are mapped on a preference order of O_k such as:

$$\forall (q_i, q_j) \in X = \begin{cases} q_i \succ q_j \in O_k & or \\ q_i \prec q_j \in O_k & or \\ q_i \equiv q_j \in O_k \end{cases} \tag{3.1}$$

where $q_i \succ q_j$ is the preference of q_i over q_j (or vice-versa), while $q_i \equiv q_j$ denotes an equal preference of both elements. As an example, in this work an element q_i consists of an audio feature vector of audio asset i, while O_k consists of human self-reports on the perceived affect, such as which sound was considered more tense/arousing/pleasurable. A function capable of transposing element features into an ordered space is subsequently learned, within this thesis both the rank support vector machine (RankSVM) and artificial neural networks (ANN) are used. Furthermore to avoid ambiguous responses only clear preference statements such as $q_i \succ q_j \in O_k$ or $q_i \prec q_j \in O_k$ were taken into account for the construction of audio affect models, ignoring tied preferences such that $q_i \equiv q_j \in O_k$. For the interested reader a more detailed and formal definition of PL can be found in Fürnkranz and Hüllermeier (2011). All models constructed within this work utilized a highly modified version of the Preference Learning Toolbox (PLT) by Farrugia et al. (2015), which was adapted for the context of this thesis. PLT is an open-source accessible software featuring a variety of pre-processing, feature selection and preference learning algorithms such as ANNs and RankSVMs.

Rank Support Vector Machines

A support vector machine (SVM) is a binary classifier algorithm, which through labelled data optimizes a hyperplane capable of segregating seen and unseen data into two diverging groups. First proposed by Cortes and Vapnik (1995), SVMs consist of optimizing a

linear boundary of a weight vector w, capable of separating data samples q_i in a projected space $\phi(X)$. Given a set of labelled data $(d_i, q_i) \cdots (d_n, q_n), d \in \{-1, 1\}$, where d_i is the classification of q_i, an SVM can be formally described as the minimization of function:

$$\frac{1}{2}\|w^2\| + C\sum \xi_i \tag{3.2}$$

$$\begin{gathered}\text{subject to:} \\ \forall q_i \ (w \cdot \phi(q_i)d_i) \geq 1 - \xi_i \\ \forall i \ \xi_i \geq 0\end{gathered}$$

where w is the decision boundary and its norm $\|w\|$, C a constant weight parameter and ξ_i a set of non-negative variables.

RankSVMs is a modified version of the standard SVM, which was first introduced by Joachims (2002). This specific type of SVM attempts to maximize the Kendall's τ between the expected ranking r^* and the proposed $r_{f(q)}$, where the feature space consists of a mapping $(\Phi(q, d))$ between a feature vector q and its ranking label d. Although Joachims (2002) approach was developed specifically for the ordering of web-pages based on the number of clicks, a pairwise approach was also subsequently developed where the difference between feature vectors expressed the preference, which in this case are the support vectors. Formally it consists of minimizing the same equation 3.2 subjected to:

$$\begin{gathered}\forall (q_i^p, q_i^n)(w \cdot (\phi(q_i^p) - \phi(q_i^n)) \geq 1 - \xi_i \\ \forall i \ \xi_i \geq 0\end{gathered}$$

where q_i^p and q_i^n is the preferred and non-preferred objects of a pairwise comparison i. According to Herbrich et al. (1999), by introducing Lagrangian multipliers and solving the constrained optimization problem through the usage of mathematical programming techniques(Mangasarian (1969)), a linear combination of feature vector differences can be expressed as follows:

$$w = \sum_{n}^{i=1} \alpha_i(\phi(q_i^p) - \phi(q_i^n)) \tag{3.3}$$

where $\forall \alpha_i \ 0 \leq \alpha_i \leq C$. Once a w is successfully trained its boundary describes the specific direction of the ordering in projected space (see Figure 3.2). Given a pair (q_i, q_j) the SVM predicts $q_i \succ q_j$, if $w \cdot \phi(q_i) \succ w \cdot \phi(q_j)$. Given that SVMs create linear boundaries on a projected space ϕ it may infer a more complex boundary on the input space. Additionally the kernel trick may also be applied by replacing the dot products by a kernel function: $\kappa(q_i, q_j) = \phi(q_i) \cdot \phi(q_j)$, similarly to the kernel trick within standard SVMs (Herbrich et al. (1999)). It is important to note that if the kernel trick is used then a w value can not be obtained directly, and the construction of an SVM relies specifically on its support vectors.

For the purposes of this work two different kernels were explored:

- Linear: $\kappa_l(q_i, q_j) = q_i \cdot q_j$
- Radial Basis Function : $\kappa_g(q_i, q_j) = e^{-\lambda\|q_i - q_j\|^2}$

For the interested reader more detailed description of RankSVMs can be found Joachims (2002); Herbrich et al. (1999).

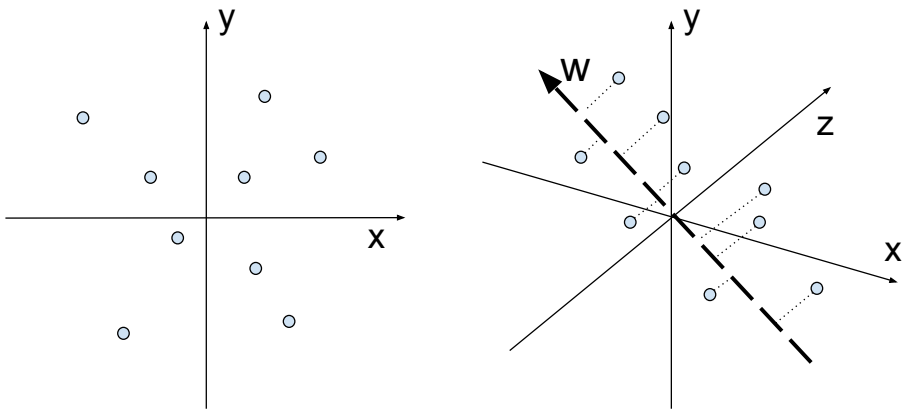


Figure 3.2: Each transformed data point $\phi(q)$ is projected onto $\vec{w}$. The ordering of each projection according to the direction of $\vec{w}$ dictates the global order

Artificial Neural Networks

Artificial Neural Networks (ANN) are a popular machine learning algorithm which consists of a interconnected network. Each interconnection has an associated weight and connects two opposing processing units (neurons). Neurons process information incoming from both previous interconnected neurons and their respective connection's weight, and then pass this information along to the subsequent neurons they are connected to (see Fig. 3.3). This process continues until reaching the output layer, which consists of the final solution values of the network (Russell et al. (2003)). Given a set of real-valued inputs, an output o_j of neuron j is computed such as:

$$o_j = g(\sum_{j=1}^{n} w_{ij} x_i) \tag{3.4}$$

where x_i and w_{ij} is the respective input and connection weight of i and g the activation function of neuron j. Although activation functions may vary substantially, it is common to use both sigmoid or hyperbolic tangent functions due to their monotonic characteristics and that their derivatives are relatively cheap to compute (Bishop (1995); Russell et al. (2003)). The latter is especially important for gradient descent learning such as backpropagation.

Defining the topology (network structure) of a ANN is not a trivial task. The simplest topology available consists of a Single Layer Perceptron, where input neurons are directly connected to an output neuron. Figure 3.4 is a visual representation of a multi-layer perceptron (MLP). An MLP consists of a fully connected network where each neuron from the input layer connects to all neurons in the hidden layer, and subsequently all hidden neurons connect to each neuron in the output layer. The MLP topology has been a popular standard within ML and pattern recognition literature, as it can theoretically approximate to any continuous function (Hornik et al. (1989)). It also presents several efficient options for learning.

Learning consists of the optimal approximation of an ANN's weights, activation functions and topology towards an unknown function. The training phase consists of an automated process over a large set of training samples, where the ANN predicts the output of each sample and subsequently self adjusts according to the error between its prediction

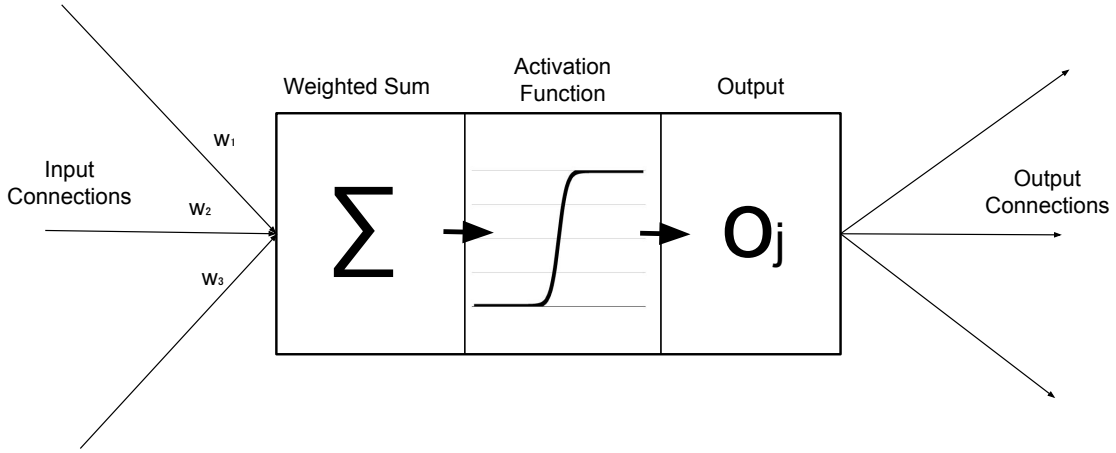


Figure 3.3: Example of a neuron (or perceptron) of a Artificial Neural Network, where a weighted sum between each connection weight (w_{ij}) multiplied with the respective input (x_i) is calculated. The activation function of neuron j is then subsequently applied, with it's output then fed as input to the following neurons.

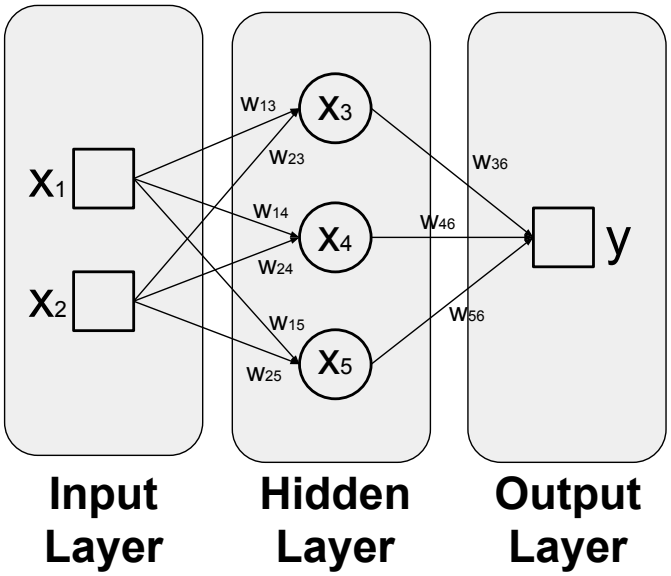


Figure 3.4: Example of an ANN with a Feed-Forward Multi-Layered Perceptron topology, which consists of a fully connected network where information flows in a single direction (i.e. input $\rightarrow$ output). The input layer consists of two inputs (x_1 and x_2) which are fed forward to all neurons in the hidden layer (x_3, x_4 and x_5) and then subsequently fed to the single output (y) in the output layer.

and the actual output. In order to apply learning two main components are necessary: 1) an *error function* capable of measuring the error of an ANN's prediction and the desired output; 2) a *training algorithm* that efficiently searches the solution space for different

configurations.

Calculating the Error Calculating the error of pairwise rankings consists of finding the utility function capable of satisfying:

$$\forall(q_i, q_j), r_{q_i} > r_{q_j} : F(q_i) > F(q_j) \tag{3.5}$$

where q_i is an object and r_{q_i} is its respective rank. For each possible pairwise comparison with diverging ranks, it is assumed that the object with the highest rank is preferred over the lower ranked one. As such an error function is defined as:

$$C = \max\{0, b - (F(q_i) - F(q_j))\} \tag{3.6}$$

where $F(q_i)$ is the expected preference score of the preferred object q_i, while $F(q_j)$ is the expected preference score of the non-preferred object q_j and b is the threshold value. This specific function consists of maximizing the difference between both the preference and non-preference scores. Once this difference is greater than the threshold value, the function becomes equal to 0. Thus by minimizing this function the separation between $F(q_i)$ and $F(q_j)$ increases until the defined threshold b is reached.

Backpropagation Backpropagation first proposed by Bryson and Ho (1969), is an optimization algorithm which calculates the loss function's gradient in respect to all the weights of an ANN over a set number of epochs. The error of training is backpropagated to each weight of the neural network and subsequently adjusted so as to minimize the loss function. Backpropagation may only be applied on ANNs with a predefined topology and activation functions. Within this work the weight update rule follows the suggestion of Martınez (2013) such that:

$$w_n^{t+1} = w_n^t - \lambda_1 \frac{1}{S} \sum_{(q_i,q_j)\in S} \frac{\partial C(U^{w^t}(q), q_i, q_j)}{\partial w_n^t} + \lambda_2 \frac{\partial (w^t)^2}{\partial w_n^t} \tag{3.7}$$

where C is the error function depending on the pairwise objects (q_i,q_j) and the network configuration $U^{w^t}(q)$, S is the set of pairs in the training dataset, $\|S\|$ is the number of pairs in S, w_n^t is the value of weight n at epoch t, λ_1 is the learning rate, while λ_2 is a regularizer weighting parameter. The latter is used to maintain the weights low so as to reduce the chances of overfitting the network.

3.2 Automated Feature Selection

Extracted features can derive several characteristics of the original raw data. However, not all of these might be useful for the learning process, as some features might be irrelevant or even detrimental (i.e. noise). Thus, it becomes crucial to distinguish between these for a more effective learning process. Automatic feature selection (AFS) consists of selecting or discarding features according to an heuristic measure of performance. Unlike other methods of dimensionality reduction such as principal component analysis (Wold et al. (1987)), which projects features into a lower dimensional space, AFS keeps the original dimension intact. Two types of AFS algorithms were used in this thesis: *Sequential Forward Selection* and *Sequential Backward Selection.*

3.2.1 Sequential Forward Selection

Sequential Forward Selection (SFS) consists of sequentially selecting features that are best capable of improving the prediction accuracy, until it ceases to improve. Starting with an empty set of features where $F_0 = \{\emptyset\}$. SFS sequentially adds a feature x_i so that it maximizes the objective function $O(F_k + x_i)$, where F_k is the current combined set of selected features. Once the objective function ceases to increase, the algorithm stops and the current set of features F_k is subsequently selected. SFS was chosen due its robustness, simplicity and its success in previous work (Yannakakis and Hallam (2007); Martínez et al. (2011); Pedersen et al. (2010)).

3.2.2 Sequential Backward Selection

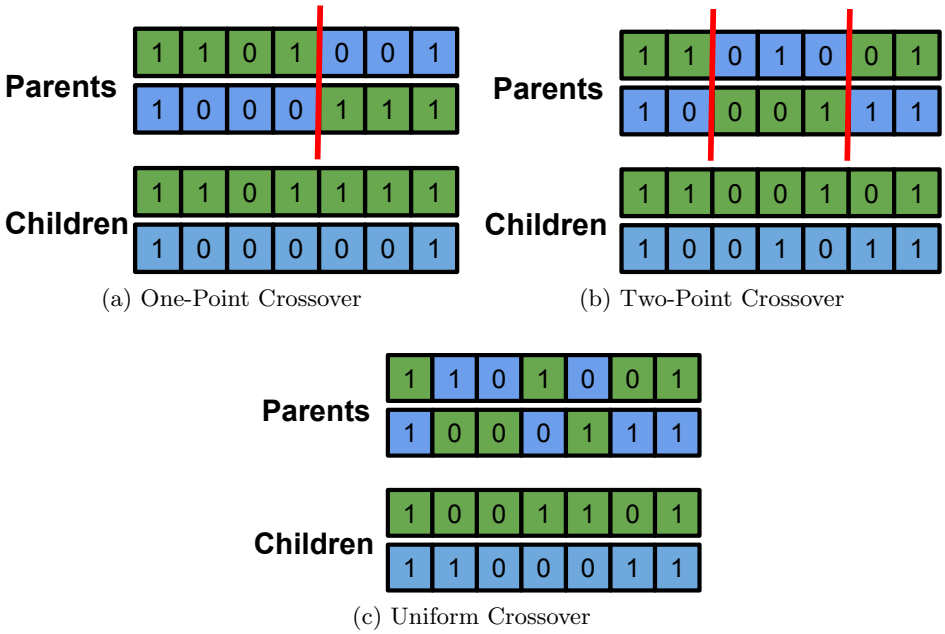
Sequential Backward Selection (SBS) is a "top-down" method, which consists of removing features from the dataset so as to least deteriorate or improve the objective function until reaching a maximum number of defined features l. Alternatively instead of defining a stopping parameter l, features are removed only, and only if the objective function is improved. Given a set of features F_k, SBS sequentially eliminates a feature x_i so as to optimize the function $O(F_k - x_i)$, stopping until no improvement. This particular method was chosen as a performance comparison to the previous SFS algorithm.

3.3 Genetic Algorithms

Inspired by Darwin's natural selection theory, genetic algorithms (GAs) are computational processes that optimize towards goals by *combining* and *mutating* a set of individual solutions, which are then subsequently *evaluated* and *selected* according to a measure of performance. GAs were first proposed by Holland (1975) as a way of transposing the adaptability mechanisms found in natural phenomena into a computational process. While many variations of GAs currently exist in literature, there are key characteristics that define this specific algorithm (Mitchell (1998)). This section provides a brief overview of the genetic process and its standard methodologies, while section 4.3 provides an in-depth description of the specific GA implementation within this work.

The first characteristic is the problem space representation, or more precisely how is a solution represented within the genetic process. Representations, depending on the problem in question, may not be a trivial task and must be capable of representing every solution in the problem space. Furthermore it must present some degree of flexibility and robustness due to several modifications it might suffer during the genetic process. Lastly, genetic representations must be consistent to its "real world" solution once it is decoded (i.e. no information loss). Nomenclature within the GA community often separates an individual's representation into two, one being the *Phenotype* which consists of a "high-level" representation used for evaluating the individual; while the other being a "lower-level" and more robust representation used during mutation and recombination referred to as *Genotype*.

The genetic linkage between generations is an important characteristic of GAs and a fundamental component of evolutionary theory. In GAs this is achieved through processes referred as *Genetic Operators*. The mating process consists of a *Crossover* operator which recombines the genetic material of two diverging individuals into one or more individual children for the following generation. A crossover operator can consist of segmenting the genotype of a selected pair of parent individuals obtained from the current population.



(a) One-Point Crossover (b) Two-Point Crossover

(c) Uniform Crossover

Figure 3.5: Visual representation of three different crossover types.

Each segment is then recombined with the other parent's opposing segment resulting in two new offspring. Genotypes may be divided by either selecting one (One-Point Crossover) or multiple n (n-Point Crossover) cutting points (see Fig. 3.5a and 3.5b). Alternatively by randomly selecting elements from a parent's genotype and then swapping these with the other parent can also be used for the creation of new offspring (see Fig. 3.5c). Additionally given a degree of probability children may undergo through an additional operator called *Mutation*. Mutation consists of slightly modifying one or more elements of the existing genotype, thus providing new alternatives for the evolutionary cycle. It is also possible to apply mutation directly on a selected parent for the creation of new offspring and forgoing crossover, such as the case of one genetic algorithm presented in this thesis (see section 4.3).

Evaluation consists of measuring the "worth" of each individual in the population influencing its probability of being selected for mating. This process insures that certain favourable genetic traits are retained for the following generations, thus progressing towards the desired solution. Evaluation is achieved through a *fitness function* capable of measuring each individual against an intended set of desirable traits, thus influencing the overall optimization strategy of the GA. By using the travelling salesman problem as an example, a fitness function may be defined as the minimization of path distance, where individuals with the shortest path obtain higher fitness values. An in-depth description of each fitness function utilized within this work can be found in section 4.2 and 4.3.

Individual selection and population replacement mechanisms are another important aspect of GAs, and can include either deterministic or stochastic strategies. The latter being more popular as it provides more robustness to the overall algorithm through higher individual diversity. Thus searching space more efficiently and reducing the chances of reaching a local optimum. The selection used within this work consists of a *Roulette Wheel* method, where an individual is selected at random according to a proportional probability of its fitness compared to the entire population. Alternative selection methods include *Tournament*

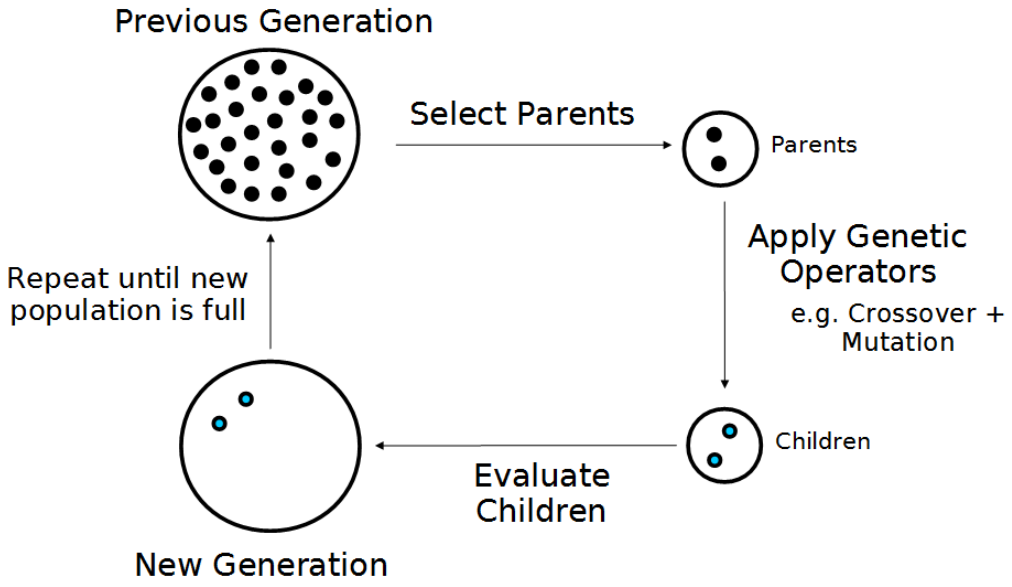


Figure 3.6: Example of a standard Genetic Algorithm Loop.

Selection and *Rank Selection*, where the first runs a mini fitness tournament on a small subset of randomly picked individuals to determine selection; while the latter is similar to roulette wheel but with the selection odds tied to the individual's rank instead of its raw fitness value. Elitism was also employed as a population replacement strategy within this work, safeguarding the top n parents of the current generation and automatically placing these in the following generation. Hence retaining individuals with the current most desirable traits.

As such the first step of the evolutionary cycle consists of creating an initial population (either deterministically or randomly) and then evaluating each individual. Once an initial population is set the generational cycle can begin. First by selecting individuals for mating through a selection mechanism. The genetic operators are then applied on the selected individuals for the creation of offspring, which are subsequently evaluated through the fitness function. This process repeats itself until reaching the total number of individuals necessary to fill the population. An entire generation cycle is represented visually in figure 3.6. This cycle continues until a favourable solution is found, or a maximum number of generations is reached.

3.4 Summary

This chapter presented the principal algorithms used for the construction of multi-faceted procedural levels. The first section described the supervised learning methods utilised for the creation of predictors that rank audio assets according to diverging affective states. This section started with a brief overview of common concepts within the field of both machine and preference learning. Methodologies utilised within this work, rank support vector machines and artificial neural networks using backpropagation, were then detailed in the following sections. Additionally for the reduction of features obtained from raw audio signal extraction processes, two automatic feature selection algorithms were described. Concluding the chapter an overview of genetic algorithms were presented, with the specific

operators and methods used within this work for the procedural construction of playable levels. The following chapter details how these methodologies were adapted and used within the context of this work.

Chapter 4

Sonancia

The *Sonancia* system was specifically developed for the exploration of procedurally generated multi-faceted content, in both the audio and level domains. Although procedural content generation systems have been widely studied as stated by Togelius et al. (2011), rarely they explore multi-facet generation. With *Sonancia* this thesis attempts to dig deeper into this core concept. First by creating a procedural level generation system capable of adapting levels to defined emotional progressions; and then through a soundscape personalization algorithm, allocating audio in accordance to the generated level.

According to Ekman and Lankoski (2009), the survival horror genre is unique in its heavy reliance on sound to convey negative affective states such as shock, disgust, ecstasy, fear and relief. It also focuses on exploration and hiding as players have limited combat ability (e.g. no weapons or limited ammunition). These complex characteristics of player affect raise important challenges for the generation of levels and soundscapes, where the focus is in evoking these types of emotions. For instance, an interesting question is how a level generator can anticipate and influence the affective state of a player, while consistently balancing feelings such as stress and relief; or how players navigate through a level under the effects of stress caused by previously encountered monsters. This thesis tackles these challenges through the exploration of procedural generation of levels and soundscapes in the survival horror genre, simulating horror gameplay during level traversal.

Sonancia generates content for a horror game in the same vein of *Amnesia: The Dark Descent* (Frictional Games, 2010). Players explore procedural haunted manors (i.e. the level) in order to find an objective located within one of the many rooms of the manor. Players are unarmed and must avoid direct confrontation with enemies; therefore monsters act as an instigator of tension and fear, regardless of the character's progression in the game. Haunted manors consist of different rooms, which are separated by walls, and doors that interconnect them. The rooms themselves can be populated by monsters, light sources and the objective (see Fig. 4.4). Furthermore rooms are also affected by sound sources, which are allocated through sonification.

This chapter describes the different interconnecting processes for the creation of multi-faceted horror levels for *Sonancia*. Section 4.1 offers an overview of the system pipeline and the flow of the content creation process. Section 4.2 defines the tension frame concept which is then used by the level generation process, described in section 4.3. The sonification process is then outlined in section 4.4. The chapter concludes with a description of the 3D transformation process in section 4.6, and a summary in section 4.7.

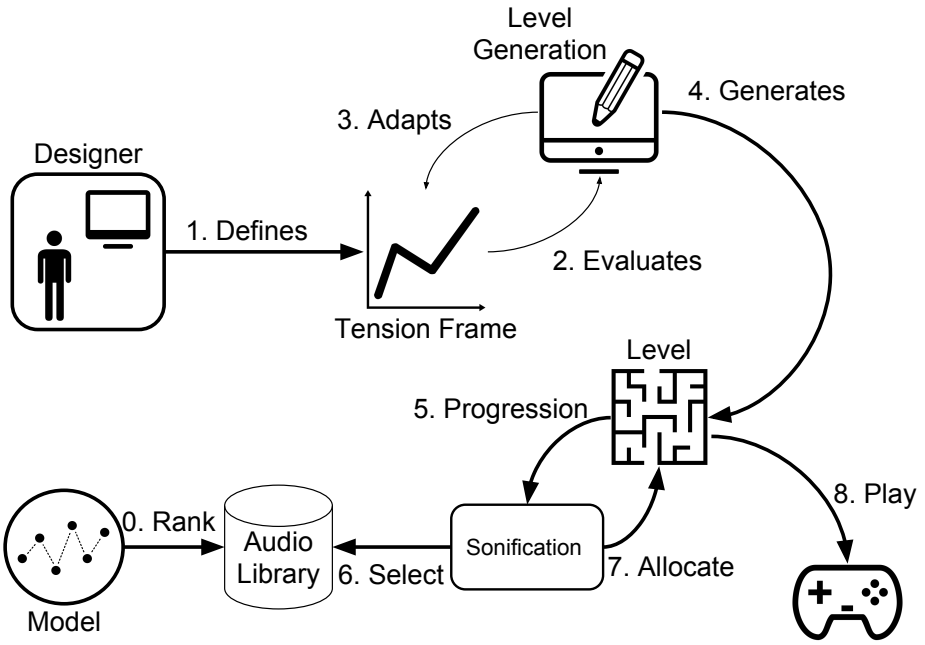


Figure 4.1: The Sonancia Pipeline: **0)** A preference learning model ranks the audio library according to perceived tension. **1)** A human or machine designer defines an intended progression of tension (i.e. the tension frame). **2)** and **3)** The tension frame acts as the fitness function for the level generation process. **4)** A level with a progression that resembles the intended tension frame is generated. **5)** The level's progression is used to inform the sonification process. **6)** Sonification selects different audio pieces from the library in accordance to both the tension ranks and the level progression; **7)** Sonification allocates selected audio pieces within the level, effectively sonifying it. **8)** The level generation is completed and the game is ready to be played.

4.1 The *Sonancia* Pipeline

The construction of multi-faceted content is achieved through multiple interconnecting processes. Figure 4.1 shows the various modules that allow *Sonancia* to create multi-faceted content. The process is kickstarted by a user or machine defined tension frame, described further in section 4.2, and it consists of the intended emotional progression that the procedural level generator should adhere to. Specifically this frame is used as the fitness function for the level generation process, by both evaluating and subsequently adapting levels towards the intended experience (see section 4.3). Once a level is generated a sonification process is applied. Sonification, which is further described in section 4.4, consists of selecting audio pieces based on the predicted global ranking of tension, derived from the machine learned models described in chapter 6 and the tension progression of the generated level. Once a level has been generated and sonified, it is then subsequently transformed into a 3D playable level using the Unity 3D game engine.

For the purposes of this thesis orchestration is explored as an adaptation of audio pieces upon a level architecture, where level construction occurs first and subsequently a "sonification" process is applied. Thus, we are further simplifying the interplay between these two digital game facets by applying a more traditional perspective of sound design in games (Stevens and Raybould, 2013). This perspective consists of adapting audio assets onto the

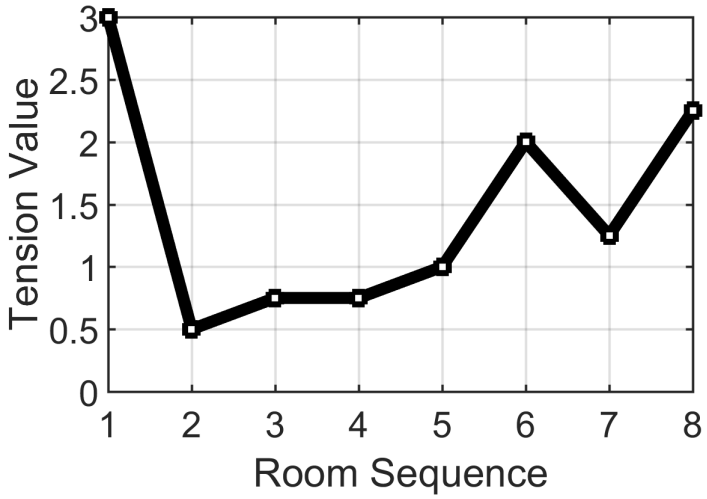


Figure 4.2: Example of a tension frame. The x-axis consists of the total number of rooms. The y-axis consists of a "tension intensity" value. The higher the tension intensity, the more tense that particular section should be.

virtual environment by selecting and placing them according to the context, progression and emotional impact the designer intends. Although other perspectives could be explored, such as applying an opposite approach is possible (i.e. starting from audio progression and adapting architecture), this work focuses on one perspective solely by exploring methods of automating the more traditional process of constructing level soundscapes.

4.2 The Tension Frame

In order to understand the level generation process utilised by *Sonancia*, it is important to first define the *tension frame*. Levels in *Sonancia* are generated through evolutionary computation influenced by framing annotations of intent, allowing the system or designer to define an intended experience the generated level should adhere to.

Intent definition is achieved through a framing device, similar to the FACE model of Colton et al. (2011), which consists of a 2D representation of how tension rises and falls as the player progresses through the level. Within this work the framing device is referred to as a *tension frame* (see Fig. 4.2), and consists of an abstract representation of the intended emotional progression of the level. The x-axis consists of a representation of the main level progression from start to finish, referred in this thesis as the *critical path*. In *Sonancia* the critical path consists of the shortest sequence of rooms traversable by the player, going from the starting location to the player objective. The y-axis consists of an abstract numerical representation of tension in each room of the critical path, within an interval of 0 (no tension) and 3 (maximum tension). The evolutionary algorithm adapts each levels critical path to fit within the intended tension curve. By altering the critical path through the addition or elimination of rooms, monsters or light sources will consequently alter how the player tension rises and falls throughout a level. This thesis will specifically focus on the horror theme, hence frames of tension serves as an amalgam of the most predominant emotions within the horror genre, which according to Ekman and Lankoski (2009), consist of fear, anxiety and stress.

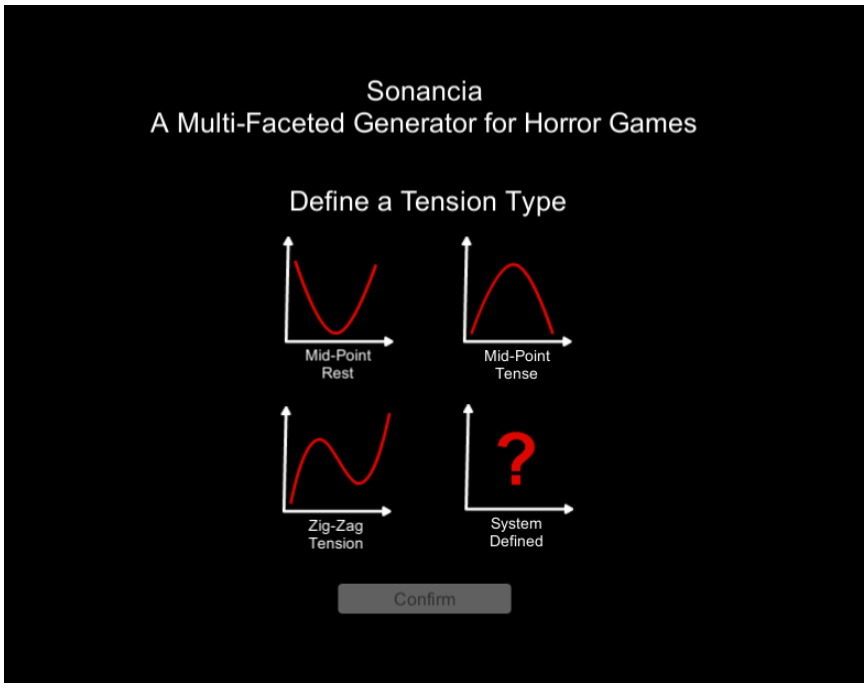


Figure 4.3: The intended tension curve selection screen. Currently the system allows for the selection of three different curve types and a system defined curve.

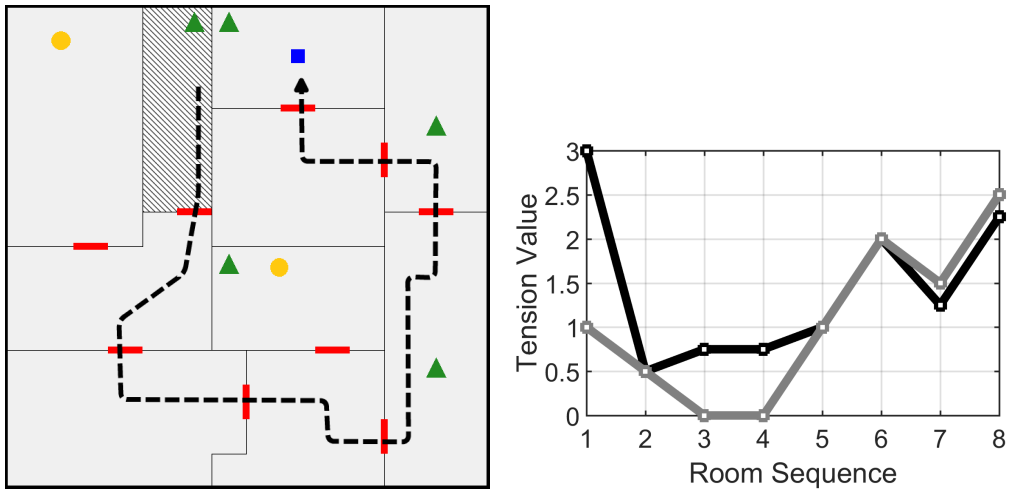
Designer Defined Tension Curve

Sonancia allows designers to select from three different tension frame types, which are defined within the system (see Fig. 4.3). The **Mid-Point Resting** curve consists of creating levels with high tension at the beginning and end of the critical path; **Mid-Point Tense** is the previous' inverse curve, rising the tension until peaking midway through the level; and finally the **Zig-Zag Tension** curve consists of two high tension peaks towards the middle and end of the level. For additional customization, *Sonancia* also allows designers to define the frame through a comma separated value file, where values consist of tension intensity and indexes the room sequence.

System Defined Tension Curve

Sonancia is also capable of generating frames via genetic search, driven by several narrative progression aesthetics. The frame is represented as an array of values between a minimum value of 0 and a maximum value of 3. The array index represents the critical path order, while the array values represent the specific tension value. The GA operators include a roulette wheel selection mechanism with one-point crossover. After recombination each offspring has a 20% chance of mutating, i.e. incrementing or decrementing a single value in the array by 0.25 (provided the result is within 0 and 3). The GA runs for 100 generations with a population of 100 individuals, each initialized with random tension values.

Eight different fitness functions are encoded into the system, inspired by narrative structures and normalized between $[0, 1]$. The **Escalating** and **Decreasing** tension fitness rewards individuals with rooms that have a higher or lower tension value from the previous room, respectively. The **Resting Point** fitness rewards individuals with the deepest tension



(a) Example 2D level

(b) Tension Frame of Fig. 4.4a.

Figure 4.4: Example of a *Sonancia* "haunted manor" level in 2D (Fig. 4.4a). In Fig. 4.4a, the room with the diagonal lines is the starting room, red rectangles are doors, green triangles are monsters, yellow circles are light sources, the blue square is the objective and the black arrow is the critical path (the shortest path between the starting room and objective). The critical path creates a level tension curve (grey) in Fig. 4.4b which must closely match the intended tension curve (black).

'valley', while the **Surprising Moment** fitness rewards the height of the highest 'peak'. The **Cliffhanger** fitness rewards tension curves with at least one peak, where the last room's tension is higher than any of the peaks. The **Denouement** fitness gives high values to individuals if the highest peak is close to the final room (but is not the final room). **Unresolved Tension** fitness rewards consecutive rooms with the same tension. Finally the **Rising & Falling Tension** fitness is proportionate to the number of peaks in the tension curve. To increase the expressiveness of generated frames, the system can choose two fitnesses and apply an **"Or"** or **"And"** operator which sums or multiplies, respectively, the individual fitness scores.

4.3 Level Generation

Given the prior success of search-based procedural content generation (PCG) (Togelius et al. (2011)), a similar approach was taken for the generation of levels in the *Sonancia* system. A genetic algorithm (GA) was chosen specifically due to its optimization ability towards specific goals, while retaining stochastic features for a wider range of level variations. *Sonancia* uses a mutation-based GA, capable of altering the level elements positioning, or even adding/removing them from a level. The GA is also capable of altering the levels layout by changing the shape of each room. Each level is evaluated based on two different characteristics: the critical path progression and the level structure. The first characteristic allows the GA to construct levels around the tension frame, while the latter evaluates the overall structure of a level, penalizing unusable or inaccessible rooms and paths. The following sections describe the adaptation of GAs for the procedural generation of levels in the *Sonancia* system.

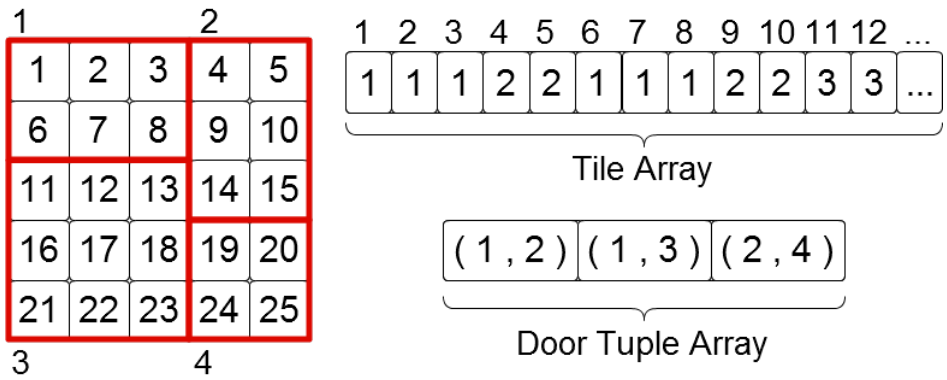


Figure 4.5: (Left) A phenotype of a 5x5 map with 4 rooms. (Right) The genotype representation of the first 12 values of the phenotype on the left, where the index is the spatial location and the integer value is the room identifier. An array of tuples represent the available connections between rooms.

4.3.1 Representation

Haunted manors consist of a layout of rooms that are separated by walls and interconnected by doors (see Fig. 4.4a). Additionally rooms may be populated by several level elements such as monsters, light sources or the player objective. It is important to note that each room can only contain one of each level element type. All rooms are assigned an ID (the room ID), where the room with the lowest numbered ID is the player starting area.

Levels are laid on a grid of tiles, which are represented by an array of integers. Each integer maps to a specific tile of the level, and represents which room occupies that tile. Doors consist of tuple objects, which map the two interconnected room identifiers (see Fig. 4.5). Level elements also consist of tuple objects, where one value describes the type (i.e. objective, monster or light source) and another the room identifier, mapping its placement within the level. Additionally level elements and doors are positioned within their assigned rooms during the genotype to phenotype conversion. Doors position themselves deterministically, as closely to the middle of the interconnecting wall as possible, while level elements are placed on a randomly uninhabited picked tile from their respective assigned room.

4.3.2 Genetic Operators

Sonancia uses a mutation-based GA without recombination, ensuring the validity of each genotype after a genetic cycle is concluded. Recombination methods, such as the crossover operator are disregarded due to its aggressiveness specifically on the genome level representation. Individuals are selected for mutation using a roulette wheel selection method. An elitism operator was also set to retain 3% of the best population for the next generation. A wide range of mutation operators were developed to influence the level structure and the level element placements. Additionally, each mutation has pre-set probabilities of affecting selected individuals. For this thesis the following mutation operators are implemented:

- Wall Shift Mutation: Randomly selects two adjacent rooms and shifts their interconnecting wall by one tile. Shifting direction is randomly selected based on the current possible directions.

- Divide Room Mutation: Randomly selects a room and divides it in half. Rooms are divided in the direction perpendicular to the longest wall.
- Add and Remove Mutation: Selects both a room and an element type at random, with the exception of the objective element (i.e. there can only be one objective). If the selected room already contains that specific element type, then it is removed. If it does not contain that element type, then it is added to the room.
- Move Element Mutation: Selects a random level element from all existing elements within the level. Moves that element to a randomly picked room that does not contain that specific element. If no move is possible (i.e. all rooms contain that element type), then no move is made and the level suffers no alteration.
- Move Objective Mutation: Moves the objective to a randomly selected room.

Mutations that move the objective have a lower percentage of occurrence in comparison to the rest, due to the disruptive nature of these operators; for instance, moving the objective affects the critical path which can highly disrupt the evolutionary progress. After mutation is applied, a flood fill algorithm ensures that rooms are larger than 5 tiles; if not, the gene is repaired to assimilate small rooms with adjacent ones, moving items or monsters as needed.

The above mentioned five mutation types were specifically chosen to allow for the construction of sufficient level variation for players (i.e. Wall Shift), and for exploring the conceptual space of possible level solutions. The divide room mutation allows the levels to increase the number of rooms existent within the level, which can consequently also effect the number of rooms within the main progression. The element manipulation mutation operators (i.e. Add/Remove and Move Elements) directly influence the tension progression, which is further described in the following section. The objective mutation consists of a mutation operator that influences both the tension and structure fitness, the latter is also further described in the following section.

4.3.3 Evaluation

To determine the quality of a level, the fitness function considers both the room layout (*structure fitness*) and the distribution of level elements along the critical path (*tension fitness*). Both fitness dimensions are described below. When evolving levels, the two fitness values are added to determine which individual will be selected for mutation.

The structure fitness consists of evaluating different paths from the player starting room towards the objective. In order to discourage linear levels (i.e. levels with no branching paths), an additional temporary objective is added. The critical path is chosen by picking the objective whose shortest path, from start to objective is the longest. The structure fitness (f_s) is calculated as:

$$f_s = R_s - P_s \tag{4.1}$$

where R_s is the number of rooms that are uniquely traversed by each path from start to an objective; discouraging linear levels; P_s is the number of rooms with no doors, which penalizes disconnected, unusable rooms.

Unlike other game genres, horror attempts to create a sense of unease that slowly builds up over time (Cheong and Young, 2008). *Sonancia* levels attempt to simulate this by analysing the distribution of level elements along the critical path, which is defined as the

level tension progression, not to be confused with the tension frame. Unlike the tension frame, the level tension progression consists of the actual tension representation of a level, derived from the current level element distribution. The tension fitness consists of minimizing the distance between both the tension frame and the level tension progression (see Fig. 4.2). A level tension progression is created by following each room of a level's critical path and increasing tension by 1 if it contains a monster, decreasing it by 0.5 if a light source is present. If a room does not contain monsters, a decay effect is applied decreasing tension by 0.5, in order to simulate the decay of tension.

Matching the tension frame and level progression is achieved by fitting the frame to each room available in the level's critical path, by sampling the frame if the total room number diverges, and then evaluating their similarity via the following tension fitness function (f_t):

$$f_t = \sum_{i}^{r} 1 - \mid L_i - T_i \mid \tag{4.2}$$

where L_i and T_i are the level progression and frame tension values of room i, respectively, and r is the total number of rooms on the critical path. It is important to note that the tension frame also acts as a constraint for the maximum number of rooms in the critical path: levels with more rooms than the x-axis of the intended tension curve receive a f_t value of 0.

4.4 Sonification

Level sonification consists of selecting and allocating different audio pieces within the level so as to accurately match the level tension progression. The main goal of this module is the context specific allocation of short audio pieces based on both the tension value set to each room, as defined by the level tension progression, and the predicted audio ranks defined by the preference learning models.

This section will focus on how sound is chosen and allocated within the level, while chapter 6 describes the methodology responsible of ranking audio pieces based on the emotional model of Schimmack and Grob (2000). Sonification consists of two distinct phases: the audio selection phase and the audio allocation phase.

Audio Selection

To increase audio fidelity and avoid breaking player immersion, it is critical to select sounds in an efficient manner and distribute them throughout the generated level appropriately. To accomplish this each audio piece within the library is ranked according to tension, arousal and valence by an audio model of affect. All ranks are calculated apriori, before both level and sonification processes begin. Based on these rankings the *Sonancia* system provides 4 different selection methods:

- Hall of Fame: Selects the top x audio pieces of the selected emotional dimension. An inverted version also exists (i.e the lower x audio pieces).
- Equidistant: Equidistantly selects x number of audio pieces based on their ranking in the selected emotional dimension.

- Granular: Audio assets are selected according to the minimal difference between the normalized raw emotion value of an asset and the level tension value of a room. With this selection method audio is automatically allocated to the room whose difference is minimal.
- Random: Audio assets are selected randomly.

Audio Allocation

Once a group of audio assets have been selected from the library, it is important to associate a location for these within the level. Currently the system allocates one audio asset for each room of the level, while giving priority of "first pick" to rooms that belong to the critical path. Similar to audio selection, *Sonancia* features several different methods of allocating sound:

- High Ranking: Allocates the highest ranked sounds to the rooms with the highest tension values.
- Low Ranking: Allocates the lowest ranked sounds to the rooms with the highest tension values.
- Random: Randomly allocates sounds to rooms, despite rank and tension values.

Audio Playback

Once sound allocation is complete the level is ready to be played. Each audio asset is directly tied to the meshes of a room within the *Unity 3D* (Unity Technologies) game engine. This allows for each sound to be played at any point within the specific room it is associated to. Once players approach an interconnecting door, the audio from the neighbouring room will start to blend with the audio of the current room, in order to offer players a sense of foreshadowing of the upcoming sound. For the interested reader a demonstration of the system is available here [1].

4.5 Connecting Level and Audio Tension

In order to contextualize the entire *Sonancia* process and its relation to tension, this section offers a top down description of how tension is used to relate both levels with audio pieces. The progression of tension is derived directly through the level generation process and the tension frame defined in the previous sections of this chapter (Section 4.2 and 4.3.3). Each generated level contains a progression, which consists of the intended tension intensity for each room within the main progression. Rooms that belong to the branching paths of a level also have an associated tension intensity similarly derived to the main progression.

The perceived tension intensity of each audio piece is obtained through machine learned predictors, which is further detailed in Chapter 6. Each audio piece has an associated ranking value that is statistically predicted by the models developed within this thesis. These values consist of the associated perceived tension intensity of an audio piece in comparison to all other audio pieces contained in the library.

[1] `https://www.youtube.com/watch?v=S5bOHrdYTx8`

Figure 4.6: Example of a playable *Sonancia* level using the *Unity 3D* (Unity Technologies) game engine.

Both room and audio tension values are cross-referenced and selected based on the different methods discussed in section 4.4. For the purposes of this thesis, the *Granular* selection method was used as it attempts to adapt audio pieces as closely to the room tension as possible, allowing us to study how close the predicted tension of a level and its soundscape is to the actual emotional progression of players.

4.6 3-Dimensional Level Construction

Once a level is generated and sonified a 3D transformation process is applied reshaping the data structure into a 3D fully playable level within the *Unity 3D* game engine. The initial phase consists of constructing the 3 dimensional space from the generated data structure. This is achieved through a list of pre-defined shapes, designed apriori by a human designer, covering every possible tile situation. These shapes consist of wall, floor, corner and door tiles, which are placed according to the respective location defined by the data structure. Subsequently an invisible mesh is then built in accordance to each room's floor structure. These meshes are used to associate each selected sound to each room, so that each audio piece is played exclusively in the room it was selected for (i.e sounds do not flood into the adjacent rooms). The final phase of 3D level construction is the creation of a navigation mesh customized for each procedurally generated level. This allows for non-playable characters such as monsters to wonder within their assigned rooms (i.e. patrolling) and chasing players if they are caught within their line of sight. Each monster is assigned 6 randomly selected way-points within their respective room, which they will randomly patrol towards, if they are not chasing the player.

4.7 Summary

This chapter detailed the system pipeline and each of the interconnecting modules for the procedural generation of multi-faceted horror levels. In particular we described in detail

each of these modules, first by defining the tension frame and how it can be defined by a human or machine designer. The level generation process was then described by detailing the genetic representation, each of the operators implemented and the level evaluation process. The sonification process was subsequently described, consisting of choosing and allocating sounds based on the predicted global rank obtained from audio affect models. The last process described was the level transformation from a genetic data structure to a fully playable 3D level.

Chapter 5

Level Generation: Sensitivity Analysis

Although the usage of procedural content generation (PCG) was once considered a method for minimizing the allocation of hard disc resources, it has since then, become synonymous with gameplay replayability. Notable examples such as *Spelunky* (Mossmouth, 2008), or even *The Darkest Dungeon* (Red Hook Studios, 2016), have ushered PCG into mainstream contemporary digital games. By utilising controlled stochastic properties, such as rule-based or evolutionary systems, new levels can be continuously produced. This constant influx of new unseen levels, restricts the player's ability to memorize them, providing unexpected challenges and gameplay experiences for players, thus elongating a game's replayability.

Digital games are particularly relevant as a multi-faceted medium where visuals, audio, narrative and rule- and level-design come together in an interactive experience Liapis et al. (2014). Not only must these creative domains go well together, but they must provide players with an enjoyable experience: depending on the genre, this experience can be, for instance, frantic in "bullet hell" action games, relaxing in exploration games, or tense and terrifying in horror games (Ekman and Lankoski,2009).

When drawing inspiration from dissimilar creative domains, it is important to find the right patterns to replicate (or re-interpret) in the creative output of the system. While systems can look at structural similarities and associations (Grace et al.,2012), a promising approach is to identify the intentions of the creator of one artefact and attempt to match those intentions in the artefact of the other domain. Towards that outcome, having access to a *frame of reference* for the intentions going into the creative act is ideal. Framing information, as suggested in the FACE model of Colton et al. (2011), can be provided by the creative system itself as "a piece of natural language text that is comprehensible by people". Such framing information can clarify the intentions of the system in its design choices and can make its creativity more easily perceptible (Colton,2008). Moreover, the framing information can act as a guide when transforming media generated by such a creative system into different media.

In the context of digital games, a human game designer's primary concern and frame of reference is the intended player experience. In most games, the intended player experience affects all design decisions: from the colour palette, the responsiveness of the controls, the sound effects, the rewards, to the back-story presented in an introductory cut-scene. Taking a successful horror game such as *Amnesia: The Dark Descent* (Frictional Games, 2010) as an example, the intended player experience is one of dread, of imminent tragedy,

of confusion and constant second-guessing of players' perception and actions. Towards this experience, the visuals include dark colours and dim lights, the audio focuses on ambient noises which foreshadow monsters, the level design has narrow corridors and low visibility while the game rules preclude any way to combat monsters.

In the previous chapter the methodology utilised for the construction of procedural multi-faceted content was presented. This chapter will investigate the applied methodology in several diverging contexts, in an attempt of proving it's ability at constructing divergent horror levels. More precisely, this chapter will present results obtained from several experiments, showcasing the sensitivity of the proposed level generation process. This chapter attempts to answer two questions: does the proposed method optimize towards a designer defined tension frame; is the proposed method robust, despite diverging parametrizations.

This chapter is divided into three sections. Section 5.1 focuses on studying the level generation process on diverging pre-defined tension frames, where both the convergence rate of our proposed methodology is studied, including qualitative and sensitivity analysis of obtained results. Section 5.2 explores the proposed methodology as a meta-creative system, presenting an entire autonomous level creation cycle from intent definition to outputted level. A qualitative study is proposed, in order to showcase the diverging interpretations of level generation, based on the system's initial intended progression. The chapter concludes with discussion section 5.3, and a brief summary and overview of what was previously presented.

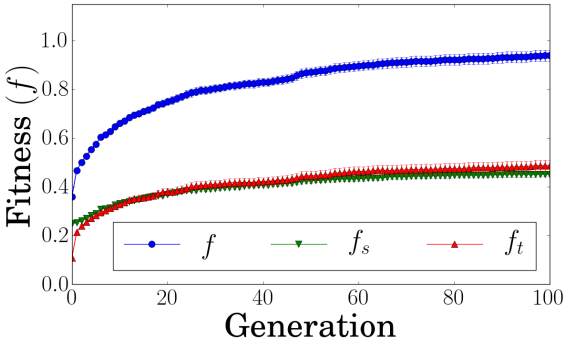
5.1 Generating Levels from Hand-crafted Frames

This section presents several experiments with the objective of stress testing our proposed methodology. The first experiment explores the convergence rate of the proposed method on several diverging tension frames. The following experiment consists of a qualitative study, exploring the level variety of the methodology utilising the same tension frame. The section concludes with an additional convergence study on 2 different map sizes and longer tension frames, exploring the robustness of level generation on larger search spaces.

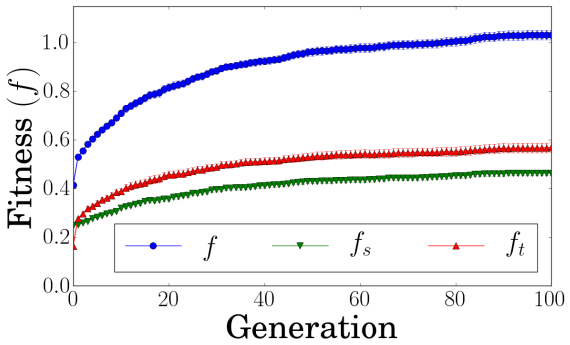
5.1.1 Tension Frame Variations

Four different tension frames were defined apriori, in order to test the convergence rate of evolution on a varied subset of different tension frame types. Figure 5.2 showcases each tension frame that was defined for the experiment presented. For each tension frame 75 independent runs of the genetic algorithm are conducted, where each run consists of 100 generations and contains a population of 100 individuals. Indicatively, figure 5.1 shows the best individuals' fitness values obtained for each defined frame. As it can be seen, similar trends of fitness progressions are found among the 4 experiments.

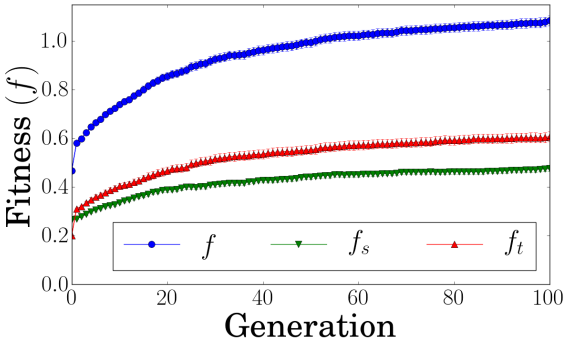
Analysis of the evolutionary progression shows that there is a slight favouritism towards the tension fitness (f_t) in comparison to the structure fitness (f_s). The reason for this is due to the relatively high dependency of f_t towards f_s. More precisely f_t tends to increase only when a set of rooms start to appear within the main progression of a level. f_s on the other hand tends to obtain lower fitnesses, because any significant changes may reflect upon f_t, forcing the overall fitness value to decrease as there is a higher probability of breaking a level's structure. This dependency can be verified by calculating a Spearman's rank correlation analysis between both fitnesses on all experiments conducted, where the average coefficient obtained was 0.74, with each experiment presenting strong statistical significance



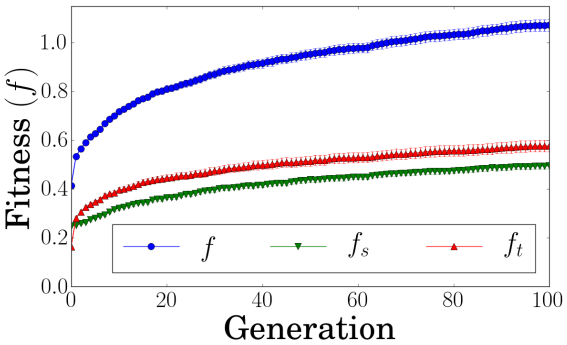
(a) Average Fitness via V-Shaped Tension (5.2e).



(b) Average Fitness via Inverse V-Shaped Tension (5.2f).



(c) Average Fitness via Inverse V-Wave Tension (5.2g).



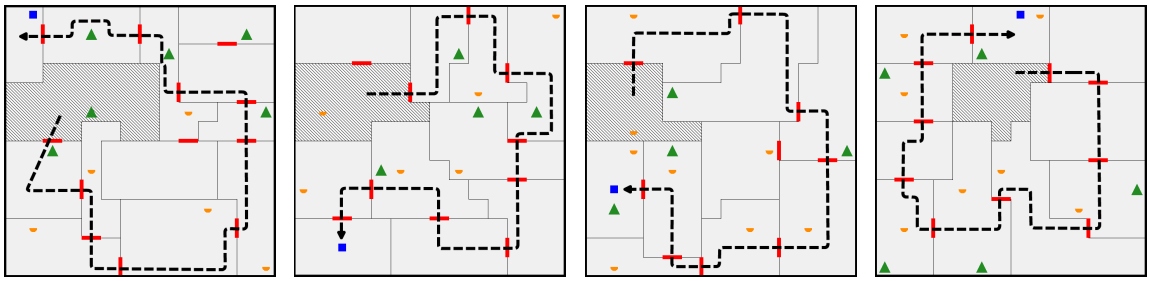
(d) Average Fitness via Linear Tension (5.2h).

Figure 5.1: Evolution of the average total fitness f (blue), and its components f_s (green) and f_t (red) for each tension frame depicted in figure 5.2. Values are averaged across 75 GA trials; error bars show standard error.

(i.e. $p_{value} < 0.05$). This suggests that there is a strong dependency between both fitnesses. Overall f_s consists of the more difficult problem, where it must evaluate levels based on the accessibility of all paths in the level, while juggling f_t and the inaccessible or maximum number of room fitness penalties.

The evolutionary cycle also shows that the overall fitness tends to consistently increase throughout the generations. It is noticeable that the fitness increase frequency begins to diminish as the evolutionary cycle progresses through the generations, although it never fully stabilizes even after 100 generations. The reason for this is because the level generation must work within specific constraints. Specifically the tension decay and the different pre-defined values that increase and decrease the hypothetical tension values of the level progression, makes it difficult for the GA to find a perfect match between both the progression and tension frame. It is important to clarify that this is the intended process, as it allows the level generator to explore different tension progression solutions, while still retaining some designer influence. If exact similarity was consistently achieved, the tension progression of different generated levels would also be similar, despite changes in level architecture. This methodology offers the level generation to have some degree of flexibility and expression without being totally constrained to designer influence, allowing for different level variations using the same tension frame.

Examples of generated levels and their corresponding tension frames are shown in figures 5.2a to 5.2d. As a reminder, it is important to note that the tension value obtained

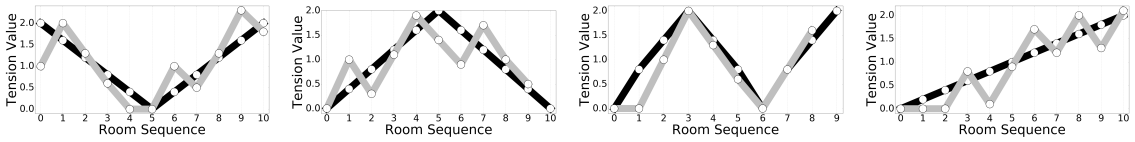


(a) Generated level via 5.2e; $f = 1.51$, $\mu = 0.93$, $\sigma = 0.19$.

(b) Generated level via 5.2f; $f = 1.43$, $\mu = 1.03$, $\sigma = 0.16$.

(c) Generated level via 5.2g; $f = 1.48$, $\mu = 1.08$, $\sigma = 0.17$.

(d) Generated level via 5.2h; $f = 1.53$, $\mu = 1.07$, $\sigma = 0.2$.



(e) V-shaped

(f) Inverse V-shaped

(g) Inverse V-wave-shaped

(h) Linear

Figure 5.2: Generated levels (a, b, c, d), with their respective fitnesses (f) using different designer-authored tension curves (e, f, g, h), represented in black, and their respective tension progressions, represented in grey. Each presented level includes the mean (μ) and standard deviation (σ) of the best obtained individuals over 75 independent runs. Darker rooms represent the players' starting room; green triangles represent monsters; orange semi-circles represent light sources and blue squares the main quest item. A black arrow follows each level's room progression, from start to main objective.

for each level is based off the sequence of gameplay elements that are present within it's main progression. More precisely, by following each room from start to objective room, tension rises by 1 for each monster present in the room, or decays by 0.5 if no monster is present. Furthermore, tension decreases by 0.2 for each light source present within the room. Tension values are restricted to non-negative integers, and will never obtain values below zero.

Figure 5.2a consists of a level generated using a "V-shaped" tension frame (see figure 5.2e). For the sake of comparison both the designer intended tension (black) and the obtained level tension progression (grey) are displayed graphically. This particular level generated the exact total number of rooms defined by the designer. Analysing both the frame and level tension values, a specific phenomenon which similarly emerges in subsequent experiments, can be visualized. This phenomenon, from henceforth, will be referred to as *tension balancing*. Tension balancing consists of the level generators attempt at distributing gameplay elements, in order to manipulate the tension values of the progression, so that each room can obtain a value of tension that is close to the value defined in the frame. The balancing forces evolution to make specific choices, in order to exploit the rise and decay of tension associated by each gameplay element. For example, in this particular level the last room is absent of any monster, as they were pre-emptively placed in the previous rooms. It also demonstrates the algorithms restraint in diverging too far from the defined frame. Furthermore, even though the frame's shape is not fully retained, it is interesting to visualize how the original framing concept is still identifiable within the tension progression. Particularly in this example, how monsters are specifically clustered at the beginning and

end of the main path.

Contrarily to the previous example, figure 5.2b shows a level generated using an inverted version of the previous frame referred to as an "inverted V-shaped" (see figure 5.2f). In this particular example, evolution was not able to obtain the total number of rooms defined by the tension frame. Particularly, in this example the generator heavily utilised light sources to drastically decrease tension in situations of decay, or mitigate the rise of tension in rooms containing monsters. This allowed evolution to concentrate on specific rooms, where almost exact similarity was achieved. In this particular example, apart from the first monster of the sequence, enemies are clustered within the middle of the main path. This follows the ideas presented by the defined frame, where the peak of tension is within the mid-range portion of the level.

Figure 5.2c showcases a level utilising a "wave-like" tension frame, which presents a mid-peak and an end-peak (see Figure 5.2g). This level also presents a 9 room progression, where monsters are more evenly distributed across the main progression. This example also demonstrates how the GA attempts to balance monster-based escalations and de-escalations of tension, so as to closely match the frame, which can be verified visually in figure 5.2g. Furthermore, the three sequential rooms after the first peak showcases how the level generator takes advantage of light sources to aggressively decrease tension, in order to achieve frame similarity.

The final figure (5.2d), presents a level generated through the linear tension frame depicted in figure 5.2h. This level presents a linear structure with 11 rooms in the progression, matching the total defined by the tension frame. It also showcases the adaptability of our methodology in situations where the tension frame minimally varies along the sequence. Given how monsters increase tension more aggressively than the slow rise of the frame, the GA attempts to counter this aggressiveness by accompanying each monster with a light source, effectively reducing tension of that particular room. Furthermore, decay is also exploited in the latter rooms, where a sort of "zig-zag" effect is used to keep the progression within a close range of the room values defined by the frame. It is interesting to visualize how the GA often plays with the transitioning values of tension between rooms, where strategic position of monsters and its subsequent decay may result in a higher similarity value, which at first glance might appear contradictory.

Figure 5.3 shows the mean and standard deviation of several level characteristics, including the total number of: rooms, accessible rooms, inaccessible rooms, traversable paths, monsters and light sources. Analysing the figure it can be seen that a similar pattern emerges among all the different tension frame experiments, with no significant differences between them. Level generation tends to average at around 8 rooms in total, with a deviation of about 1 to 2 rooms, suggesting that despite achieving the total room count, the level generator consistently prefers to stay at around 2 rooms below the total. Additionally, the level generator tends to be effective in filtering inaccessible rooms, where it averaged at 0 with a low deviation throughout the experiments. As expected the "maze-like" properties were not a priority for the GA, as can be seen in the possible paths statistics. On average the generated levels tend to present 1 to 2 paths in total, where occasionally a third path might be generated or even a forth path, although the latter is rare.

Table 5.1 showcases the Spearman's Rank correlations of level characteristics and gameplay elements in relation to fitness f. As expected, each tension frame experiment showed a significant correlation between the overall fitness and the total number of rooms present within a level. This suggests the fitness value of levels tend to increase as more rooms are added into the level. Interestingly, this correlation increases if the focus is solely on

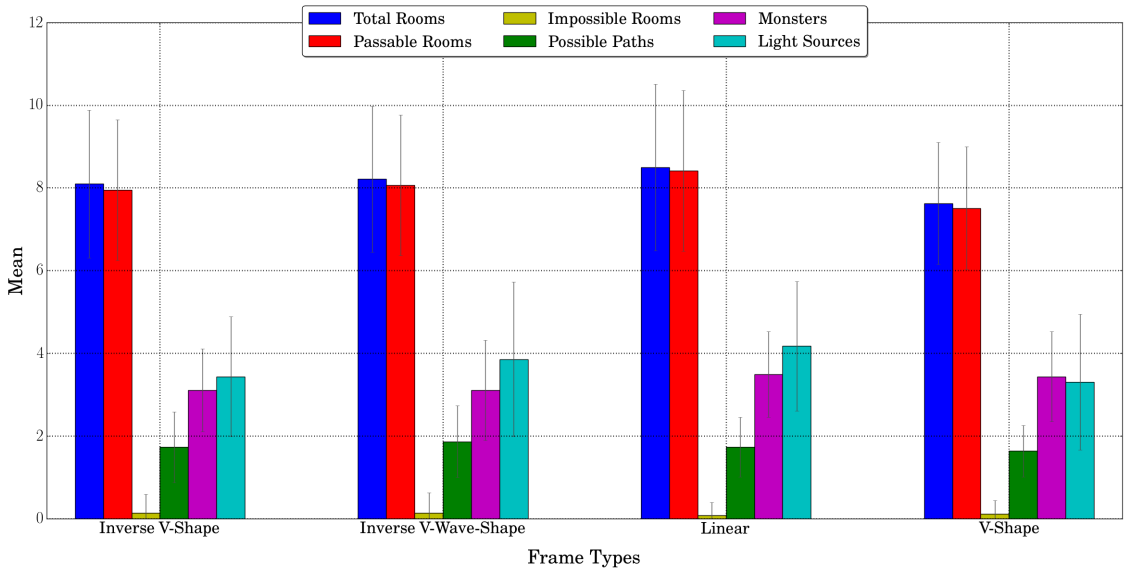
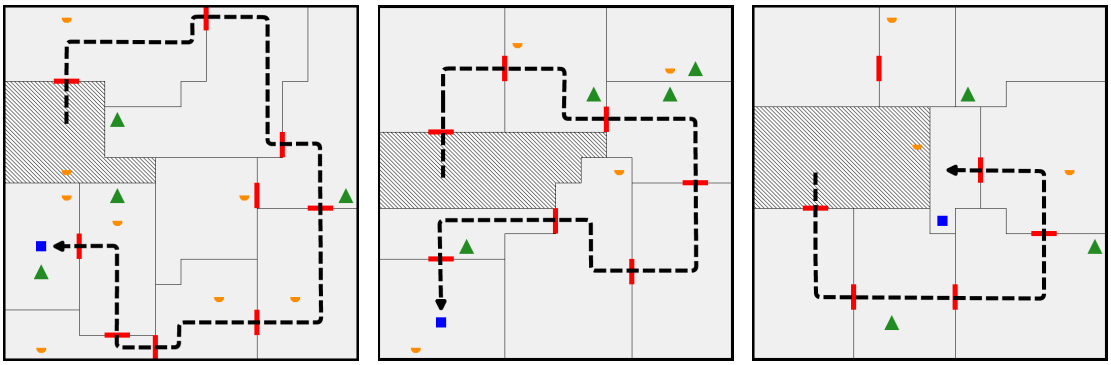


Figure 5.3: The mean and standard deviation of the level characteristics from the best individual of each 75 run over the diverging tension frames.

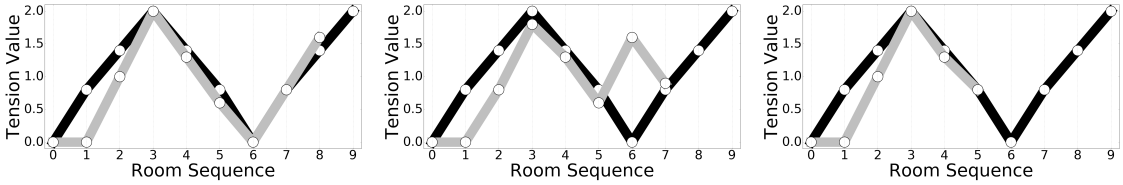
Table 5.1: Spearman Rank Correlations between the overall fitness value obtained from the best individuals of each run and their level elements. Bold values represent statistically significant correlations ($p_{value} < 0.05$).

Tension Frame	Total Rooms	Accessible Rooms	Alternating Paths	Total Monsters	Total Light Sources
V-Shape	**0.76**	**0.8**	-0.06	0.06	-0.06
Inverse V-Shape	**0.77**	**0.84**	0.11	**0.59**	**0.41**
Inverse V-Wave-Shape	**0.66**	**0.73**	0.22	**0.37**	**0.5**
Linear	**0.87**	**0.88**	0.28	**0.52**	**0.59**

accessible rooms, i.e. rooms players can reach and traverse. This suggests that our earlier assumptions about the structure fitness (f_s) is correct, where the construction of new accessible room structures has a greater impact on the overall fitness, even though graphically it appears to under-perform in comparison to f_t. For three out of the four experiments, the quantity of gameplay elements does suggest some degree of correlation between the overall fitness, however it is important to keep in mind that this is highly dependent on the type of frame used. In the particular example of the V-Shape experiment, this correlation is absent, due to being the only tension frame that begins with an aggressively high value, and because the majority of the rooms within this frame sequence, tend to present tension values below 1.2. For this frame it makes sense that evolution is particularly more conservative with the monster placement, while light sources specifically might not directly influence the fitness, due to rooms already tending towards a value of 0 tension, as for example rooms 4 and 5 of figure 5.2a. One of the current weaknesses of the proposed algorithm, is it's difficulty in constructing multi-path levels. Even though the obtained correlations are not statistically significant, this lack of significance can be justified by the fact that the majority of levels did not consistently present more than one alternating path in addition to the main path.



(a) High Fitness Level; $f = 1.48$ (b) Mean Fitness Level; $f = 1.08$ (c) Low Fitness Level; $f = 0.73$



(d) Tension of figure 5.4a (e) Tension of figure 5.4b (f) Tension of figure 5.4c

Figure 5.4: Three diverging levels generated using the "Inverse Wave" tension frame of figure 5.2g, for the comparison average, high and low fitness levels.

This particular limitation is a consequence of the algorithms focus on the main progression primarily, whereas future work could subsequently improve this limitation by focusing on the development of alternate solutions, capable of exploiting "maze-like" properties within the proposed levels.

5.1.2 Non-Diverging Tension Frame

An important aspect of PCG in contemporary games is its ability to provide diverging content. Figure 5.4 showcases 3 different generated levels using the "Inverse Wave" tension frame. Level 5.4b shows an example of a level close to the average fitness values, while level 5.4a and 5.4c consist of levels with a high and low fitness value (i.e. the outliers), respectively.

The examples shown exemplifies how the total number of accessible rooms can directly influence the fitness values of levels. For example levels 5.4b and 5.4c show an equal number of total rooms, however the latter level contains 3 totally inaccessible rooms, which is heavily penalized by the structure fitness (f_s). An unexpected consequence of a high room count in individuals, is that the GA must consistently work around the maximum room limit, increasing the probability of levels obtaining harsh penalties. Consequently, potential high fit candidate solutions have a greater risk of being discarded from the population, because it surpassed the total number of room threshold. A potential solution to increase the consistency of our methodology would be to alter the room threshold penalty from a hard constraint ($f = 0$) to a softer constraint such as an exponential value. More precisely, if the number of rooms from a level's progression deviates too far from the intended number of rooms, a harsher penalty is applied to the level. This method would offer the GA some leeway to work on progressions with higher room count, without an instant penalization.

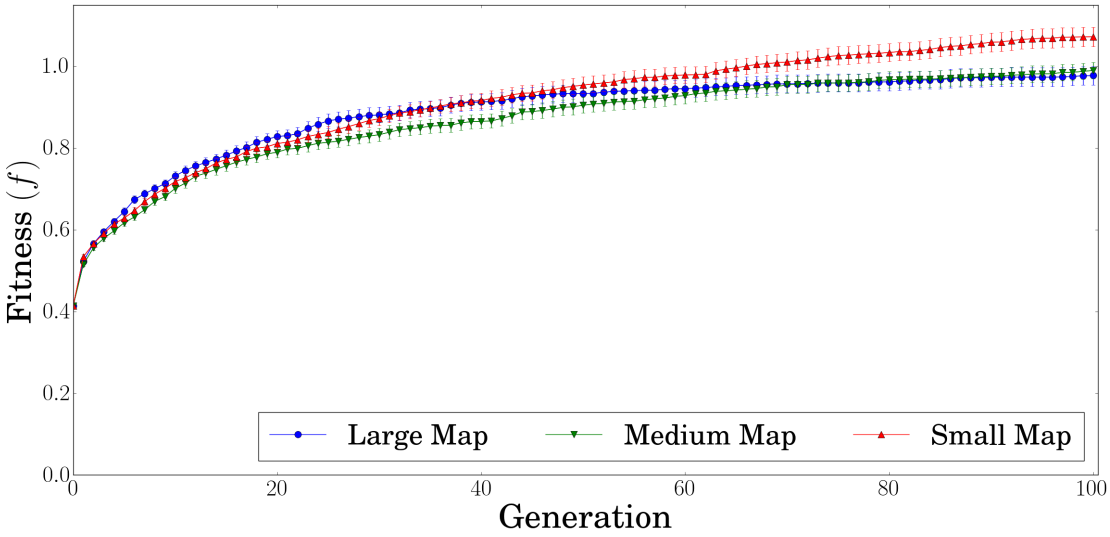
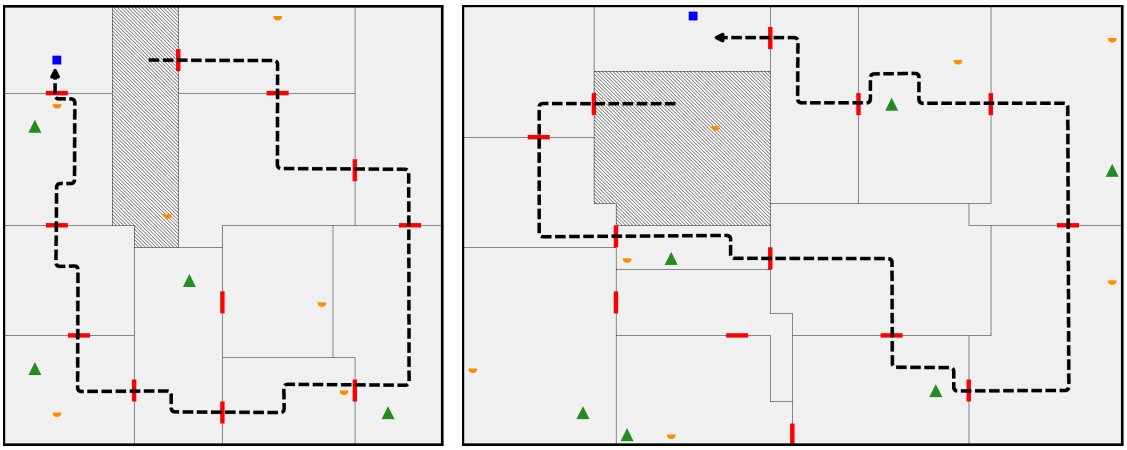


Figure 5.5: Evolution of the total fitness (f) across three different level sizes using the linear tension frame of figure 5.2h. Values are averaged across 75 GA trials; error bars show standard error.

Although "maze-like" properties were more common within high-fit individuals, it was not a consistent tendency, where most levels presented a long linear path with a singular branching room, similarly to level 5.4a. Given the low non-significant alternate path correlations obtained in table 5.1, the absence of "maze-like" properties was not particularly surprising. Given the systems focus on interpreting a designer intended progression, f_t tends to be prioritized above f_s, as fitness tends to exponentially grow if both work in concordance (i.e. f_s constructs the rooms to place the elements f_t evaluates). Although the emergence of maze-like properties was less frequent as initially intended, the analysis does suggest the viability of the proposed system in constructing different level variations utilising the same tension frame. Close to average and high fitness individuals do tend to present the desired elements defined, although future improvements would consist of a rework on some of the mutation functions and fitness penalization parameters, and potentially modify f_s so as to provide the GA more freedom to explore branching paths without directly affecting f_t. In terms of diversity, the majority tend to present diverging characteristics if it is among the mean and high fit value range. Common patterns are frequently found in the lower fit individuals, where levels tend to be quite similar to figure 5.4c. While there are differences between levels in terms of tension progression, most levels tend to create similar pattern variations of the tension frame, which was expected. These variations usually consist of alternating the monster placement throughout the progression, where for example some levels place monsters sooner or later than others.

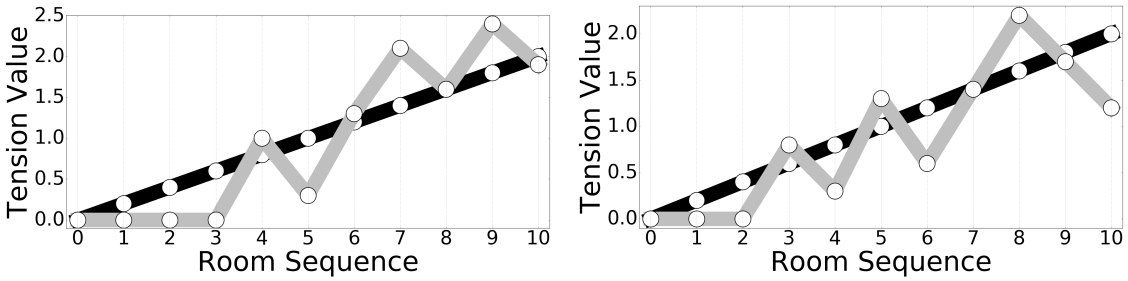
5.1.3 Diverging Level Sizes

In order to test the efficiency of *Sonancia* across different level shapes and sizes, three types of levels are tested: a small level (14x14 tiles) as per the previous experiments, a medium level (20x20 tiles), and a large level (20x24 tiles). Although the standard *Sonancia* level consists of the small variant, these experiments allow us to gauge the extensibility of our



(a) Medium level; $f = 1.37$

(b) Large level; $f = 1.71$



(c) Tension Frame and Progression of figure 5.6a

(d) Tension Frame and Progression of figure 5.6b

Figure 5.6: The fittest levels of different sizes generated using the linear tension frame of figure 5.2h

methodology for more complex map variants. Similarly to the previous experiments all level types were run for 100 generations with 75 independent trials; all experiments reported in this section use the linear tension curve as depicted in figure 5.2h.

The fitness mean over the 75 independent runs for each diverging map size is shown in figure 5.5. The obtained evolutionary progressions of all three map size variants show that they follow almost identical patterns, where the mean values for each respective generation are similar to each other. Suggesting that map size alone does not directly influence or hinder the evolutionary process.

Examples of a medium and large map are shown in figure 5.6. Figure 5.6a presents the fittest medium level obtained over the 75 runs, which consists of an 11 room progression. Apart from the longer processing time, in general, generated medium levels did not present any significant structural or tension-relevant divergences when compared to the small generated variant. Figure 5.6b showcases the fittest large level obtained over 75 runs, also consisting of a 11 room progression. Interestingly, this level presents one of the highest obtained fitness values out of all level experiments conducted, and exemplifies how a particularly long branching path can heavily influence the overall fitness, alongside f_t. Applying a Spearman Rank Correlation analysis on levels of diverging sizes (see table 5.2) suggests that this particular example is an exception, and that similarly to previous experiments large size levels tend to present similar characteristics to smaller sized generated maps.

A thorough analysis of both table 5.2 and figure 5.7 confirms that diverging sizes did not

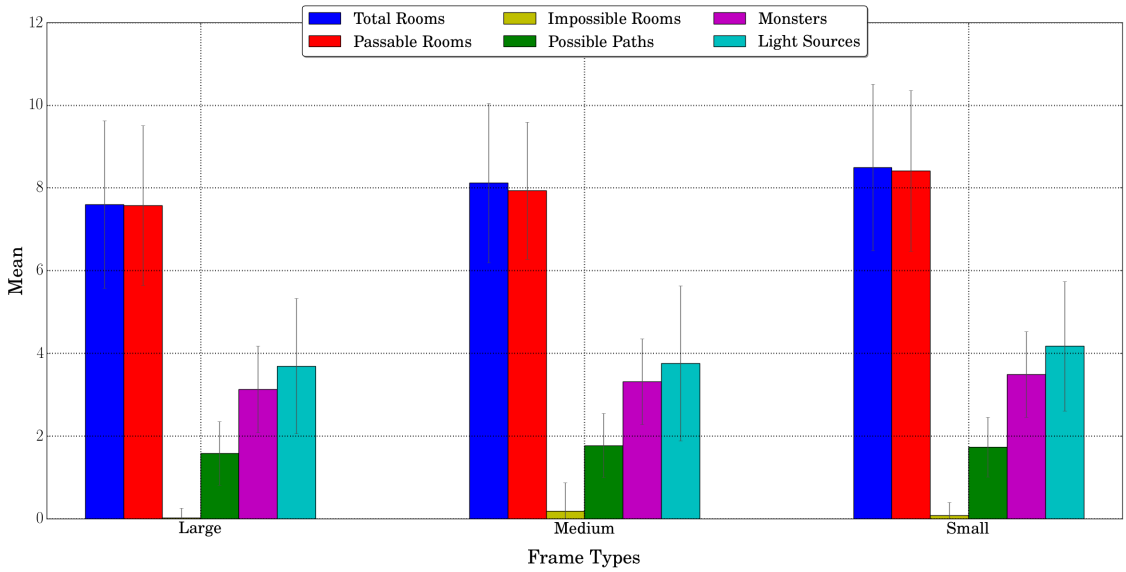


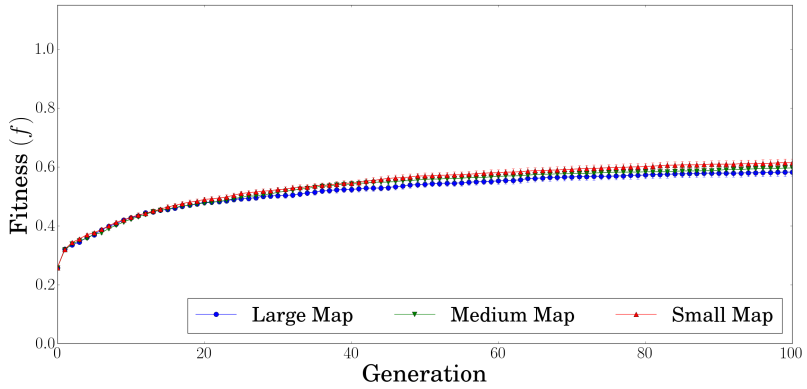
Figure 5.7: The mean and standard deviation of the level characteristics from the best individual of each 75 run over the diverging level sizes.

Table 5.2: Spearman Rank Correlations between the overall fitness value obtained from the best individuals of each run and their level elements. Bold values represent statistically significant correlations ($p_{value} < 0.05$).

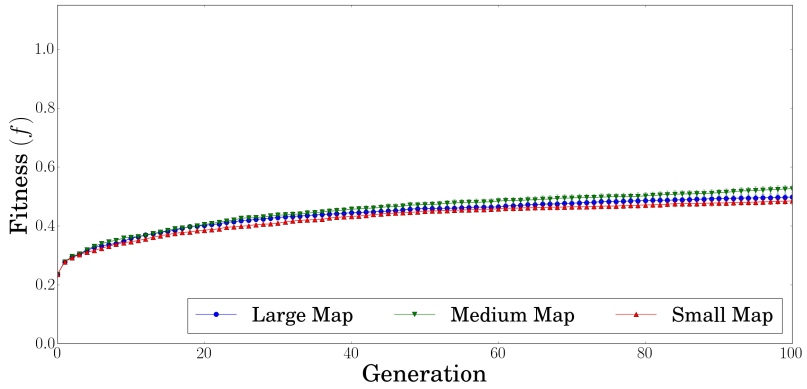
Tension Frame	Total Rooms	Accessible Rooms	Alternating Paths	Total Monsters	Total Light Sources
Large	**0.81**	**0.81**	0.08	**0.24**	**0.31**
Medium	**0.64**	**0.71**	0.14	**0.29**	**0.62**
Small	**0.87**	**0.88**	0.28	**0.52**	**0.59**

significantly influence the tendency of the genetic algorithm. Medium maps present a lower correlation value between the fitness and both room count types, however by analysing the average and standard deviation of these map characteristics, it was found that they did not particularly diverge significantly from the other map size variants. Specifically, large maps presented an average accessible room count of 7.5, with a deviation of 1.9, while medium maps averaged at 7.9 with a deviation of 1.6, while the smaller sized maps verified a higher average of 8.4 with a 1.9 deviation. This discrepancy suggests that despite obtaining structures similar to other map size variants, the medium map experiments had more difficulty managing f_t, which similarly happened the experiment utilising the V-Wave-Shape frame (figure 5.2g).

Analysing the obtained results it can be concluded that changing the map size did not significantly hinder the level generation process, obtaining similar results throughout all different variants. However, considering that more space exists for the construction of more intricate room sequences, it can be argued that the generator did not utilise the full potential of available space. Given that the tension frame was the same for each different map size, the similarity in room quantity was expected and also limited by the fitness functions.



(a) Average Fitness via Linear Long Tension (5.2e).



(b) Average Fitness via Inverse V-Wave-Shape Long Tension (5.2f).

Figure 5.8: Evolution of the total fitness (f) across three different level sizes using the long variations of Linear and Inverse V-Wave-Shape tension frame. Values are averaged across 75 GA trials; error bars show standard error.

5.1.4 Varying Tension Frame Length

This section will explore the level generation process of the same different map sizes (i.e. small, medium and large) utilising two significantly longer versions of the Linear and Inverse-V-Shape tension frames, where the total number of rooms is 20 and 22, respectively. Similarly to previous experiments each size and frame was run 75 times independently for 100 generations.

Analysis of figure 5.8 suggests that longer frames causes the genetic algorithm to struggle and prematurely convergence. Unlike previous experiments, there is a substantially stagnated rise within the evolutionary progression, which can be found in both frame experiments. Interestingly, this particular experiment showcases the importance of f_t in guiding the evolutionary cycle of the proposed methodology, and also presents it's limitations. Longer tension frames directly affect the influence of f_t on the overall fitness function f, where assuming that f_t could achieve exact similarity to a 20 room frame, each perfectly similar room would contribute 1/20th of fitness value towards f. As was the case in previous experiments, with shorter tension frames, f_t forces the structure fitness f_s to commit to certain structural patterns, thus progressing the evolutionary cycle closer towards the de-

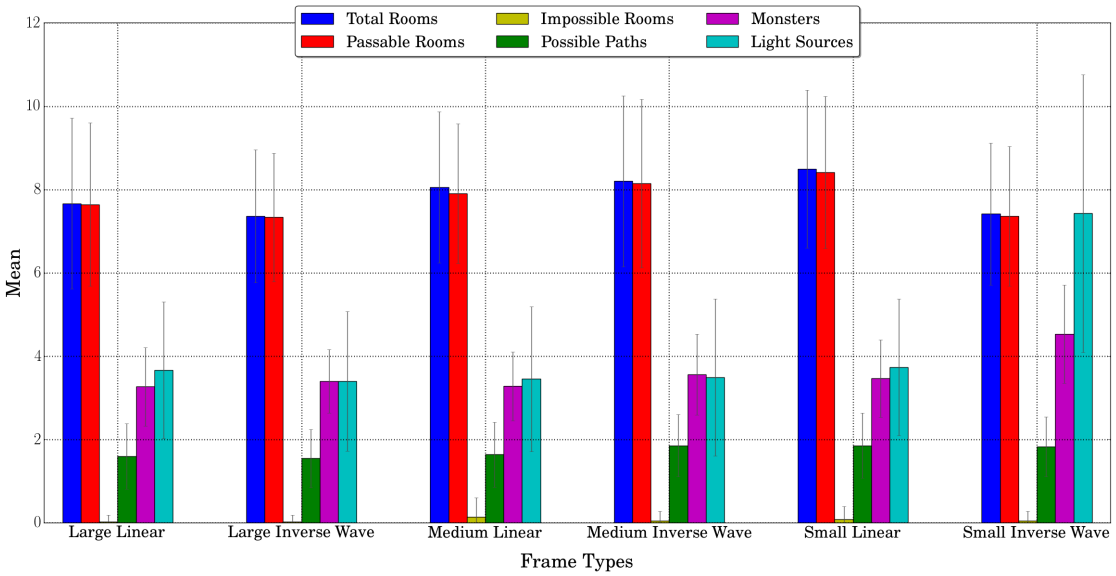


Figure 5.9: The mean and standard deviation of the level characteristics from the best individual of each 75 run over the diverging level sizes and long tension frame types.

fined tension frame. Given that the overall influence of f_t is lower, it forces f_s to constantly explore due to it not being constrained to a specific structural pattern. This causes f_s to consistently take precedence over f_t, such as changing room structure or door placement due to a higher fitness gain in comparison. Consequently, this forces the evolutionary cycle to ignore the tension frame and focus on the construction of rooms, even if it is detrimental to the overall level generation. As previously mentioned, level structure construction is a harder problem as it is highly dependent on two things: 1. Long sequences of rooms, with at least 5 tiles or more; 2. Doors must interconnect between each adjacent room of that sequence to be traversable. By increasing the size of the main progression path, it also increases the search space of total solutions, as more intricate patterns may emerge. Consequently this causes the sub-space of solutions optimal for *Sonancia* levels to diminish, as the total space of solutions increases.

Despite the longer tension frame, level generation consistently obtained an average total of 8 to 10 accessible rooms, similar to averages obtained in the smaller frame experiments (see figure 5.9). This further suggests that room limitation may not be specifically tied to the aggressiveness of mutation operators, and may be directly related to the overall fitness function utilised, as the GA struggles to find levels with larger room numbers even with a longer tension frame, effectively eliminating harsh penalizations. Although this particular limitation is not detrimental to the overall study of this thesis, as the standard *Sonancia* level and it's frames consist of an average of 8 to 10 rooms in the progression, it does question the extensibility of the proposed method for larger more intricate maps. Further improvements on the level generation process could be applied in order to further increase its efficiency on larger spacious maps, some suggestions are further given within the discussion section (see section 5.3).

Table 5.3 shows the Spearman's Rank Correlations, similarly to previous experiments. No significant differences occur despite the more stagnant evolutionary progression, with the exception of the medium sized inverted v-wave-shaped experiment, which presents a

Table 5.3: Spearman Rank Correlations between the overall fitness value obtained from the best individuals of each run and their level elements. Bold values represent statistically significant correlations ($p_{value} < 0.05$).

Tension Frame	Total Rooms	Accessible Rooms	Alternating Paths	Total Monsters	Total Light Sources
Large Linear	**0.86**	**0.86**	0.09	**0.55**	**0.38**
Medium Linear	**0.81**	**0.84**	0.11	**0.52**	**0.25**
Small Linear	**0.78**	**0.82**	0.2	**0.54**	**0.43**
Large Inverted V-Wave	**0.94**	**0.94**	-0.11	**0.39**	**0.45**
Medium Inverted V-Wave	**0.92**	**0.93**	**0.43**	**0.64**	**0.65**
Small Inverted V-Wave	**0.84**	**0.85**	-0.05	**0.43**	**0.59**

higher correlation between the number of alternating paths and fitness. However, it is important to keep in mind that obtained correlations are from individuals with particularly low fitness values, which have a tendency of presenting more branching paths simply because the evolutionary cycle is exploring potential pathways through the placement of doors and rooms around the starting room.

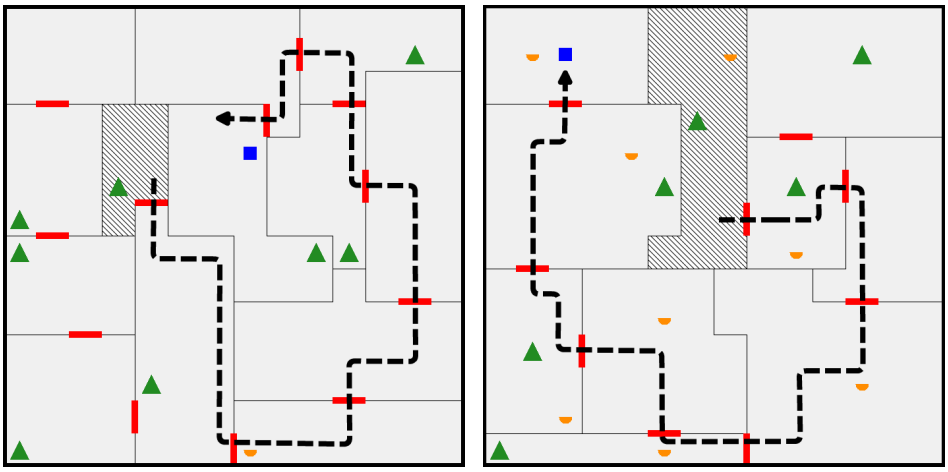
5.2 Generating Levels from Generated Tension Frames

This section showcases results obtained through the tension frame generation procedure. Each step from creating a framing of tension to the generation of levels based on this frame are presented in this section, for four different types of frames which were obtained through experimentation.

For each experiment the system runs independently 75 times, and the framing fitnesses were selected (and often combined) by the system without human intervention. Once a framing fitness is selected, tension frames evolve for 100 generations to guide level generation. Subsequently level generation runs for 100 generations using the tension frame previously generated. For brevity, we discuss the fittest individuals (tension frames and levels) for four different fitness frames; these provide the most varied and interesting results. The highlighted examples of the system's frames were provided in text as follows:

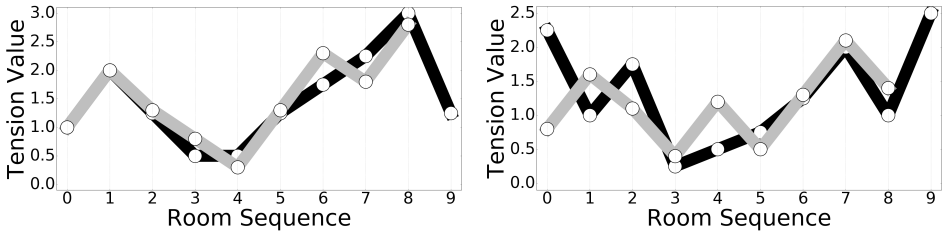
1. "I want an experience with a denouement."
2. "I want an experience with a cliffhanger."
3. "I want an experience with both a surprising moment and a point of rest."
4. "I want an experience with decreasing tension or a cliffhanger."

The following sections will describe (in the above order) the tension frames and levels created following this computer-generated frames of tension.



(a) Best level for Denouement; $f = 1.68$ $\mu = 1.02$, $\sigma = 0.22$

(b) Best level for Cliffhanger; $f = 1.35$ $\mu = 1.02$, $\sigma = 0.18$



(c) Tension for Denouement

(d) Tension for Cliffhanger

Figure 5.10: Generated levels with their respective frame (black) and level progression (grey) tension values for single aesthetics.

5.2.1 Framing Denouement

Denouement (or conclusion) is encoded aesthetically as a fitness function which rewards when the highest peak in the tension frame is near the last room (but not the last room, as that would not form a 'peak' per se). Observing the fittest intended tension frame in figure 5.10c, the frame (in black) matches this specification as the highest peak (at 3) is on the penultimate room of the tension progression.

Figure 5.10a shows the fittest level obtained for the generated tension frame discussed above; the level progression is shown in figure 5.10c, in grey. It is immediately obvious that the obtained tension progression does follow the intended frame, once compared side-by-side. Unfortunately, the key characteristic of the denouement is cut short at the end of the tension progression, due to the level generation not being able to match the exact number of rooms to that of the defined frame. Hence, in this particular example the level ends resembling more of a cliffhanger type progression, than denouement.

Nevertheless, the progression does consistently match the frame's intent, despite it being an overly tension aggressive frame. Given that the generated frame is particularly smooth, without sudden aggressive rises and falls, it allows for the level generator to be less mindful of balancing monsters, light sources and decay. This level, also presented a substantial number of monsters, which goes in-line with the high values of tension the final rise demonstrates.

The result in figure 5.10a contains 2 monsters in the first two rooms, and then no

monsters until three rooms after — allowing the player to relax from the initial tense encounters. Another intense sequence appears with 2 additional monsters that appear in the following two consecutive rooms, with a brief resting point afterwards. The level ends with a "final" monster protecting the player's objective.

5.2.2 Framing the Cliffhanger

The cliffhanger is encoded aesthetically as a fitness function which rewards tension frames with at least one peak, where the last room's tension is higher than any of the peaks (acting, thus, as the cliffhanger). Observing the fittest intended tension frame in figure 5.10d, the frame (in black) matches this specification as it contains a tension peak in the 7th room (with a value of 2.0) and it ends with a higher tension value in the final room (2.5).

It should be noted that the frame starts at a higher value in room 1 (2.25) than each peak of the frame; this is due to the fact that the first room does not register as a peak (peaks compare tension values with both neighbours), causing this artefact. Interestingly, this specific curve actually resembles the progression of most TV episodes with cliffhangers, where the episode starts by quickly resolving the previous cliffhanger, then building up tension until the problem is resolved but with a complication (the cliffhanger) for the next episode.

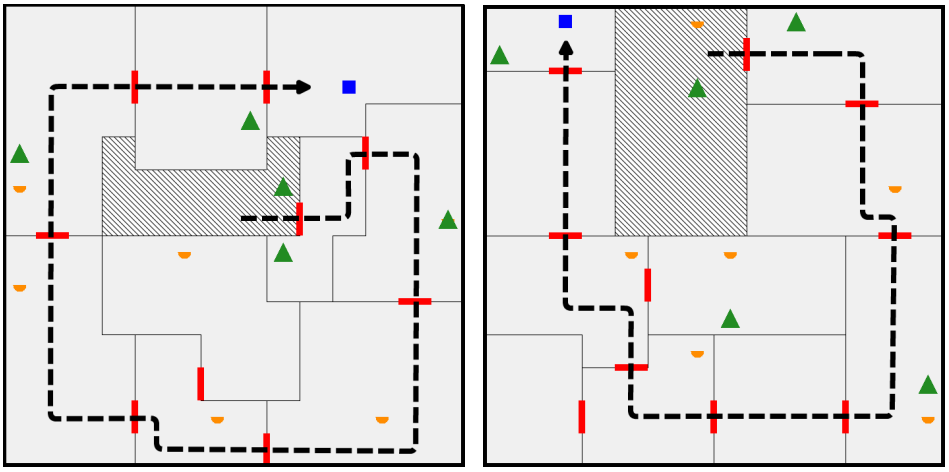
Figure 5.10b shows the fittest level generated using the obtained tension frame discussed above; its level tension progression is shown in figure 5.10d, in grey. Similarly to denouement, the level progression closely matches the intended frame for the cliffhanger aesthetic, with the exception of the last room. Both start at the maximum possible tension (for levels, this is 1 if there is a monster in the first room) and then drop the tension value in the subsequent rooms, only to increase it around rooms 5 until 6, culminating at the highest value (ignoring the first room in the intended curve). Interestingly, the progression applies a similar tactic found in the previous linear frame experiments within rooms 3 to 5, rising and falling tension to keep up with the linear increase of the frame. Ironically, due to the fact that the level generator was not able to construct the exact number of rooms present in both the Denouement and Cliffhanger level progressions, these ended resembling mostly their counterparts, as they depend substantially on the distribution of tension of the last room.

The result in figure 5.10b has 5 monsters on the level progression, distributed near the start and end of this path. This causes an initial tense moment for the players when they start the level, then lets them relax with 2 subsequent empty rooms, a small tense spike occurs afterwards, until reaching a climax with two monsters towards the end of the level.

5.2.3 Frame of Surprising Moments and Resting Points

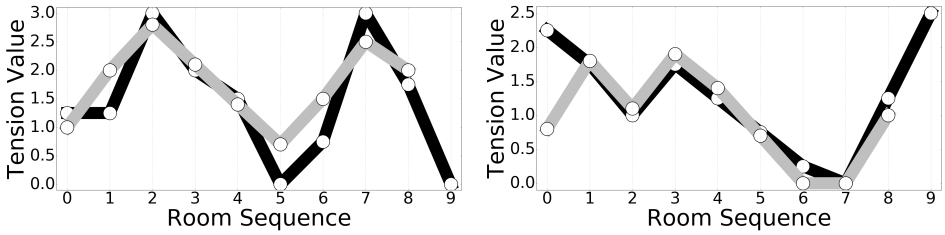
When combining fitness functions, the "and" combination forces both fitnesses to have high scores: in this case, the surprising moment aesthetic rewards high 'peaks' while the resting point aesthetic rewards deep 'valleys'. Indeed, both aesthetics are present in the intended tension frame of figure 5.11c as it exhibits the highest peak (height of 3) and the lowest possible valley (depth of 3, considering the tallest adjacent peak). The aggressive changes in tension were expected, as both fitnesses directly reward high peaks and deep valleys; their combination unsurprisingly causes tension to soar from a value of 0 to 3 within the span of three rooms.

Figure 5.11a shows the fittest level for the tension frame discussed above; its level



(a) Best level for Surprise and Resting Point; $f = 1.38$, $\mu = 0.97$, $\sigma = 0.2$

(b) Best level for Decreasing or Cliffhanger; $f = 1.37$, $\mu = 0.99$, $\sigma = 0.17$



(c) Tension for Surprise and Resting Point

(d) Tension for Decreasing or Cliffhanger

Figure 5.11: Generated levels with their respective tension frame (black) and progression (grey) for combined aesthetics.

tension progression is shown in figure 5.11c, in grey. The level progression attempts to match the intended frame, considering the constraints the obtained progression follows the rises and falls of tension of intent. The fact that each room can only have one monster (and each monster increases tension by 1) causes the progression to steadily decrease or increase consistently, in order effectively achieve the values of peaks and valleys. The structure of the level progression retains both a resting point at room 5 and the surprising moments in room 2 and 7, matching the intent of the provided frame.

The level of figure 5.11a contains a total of 5 monsters within the progression, placed primarily at the start and end of the path. This yields two high spikes (surprising moment) in room 2 and 7 after encountering three and two consecutive monsters, respectively. Midway through the progression a resting point lasts for a total of three rooms, allowing the player to relax from the aggressive tension rise of the first three rooms.

5.2.4 Framing Decreasing Tension or a Cliffhanger

Combining fitness functions with an "or" in this system multiplies the two fitness scores together. This will still reward the presence of both features but since it is less aggressive than a sum of the scores (as in "and"), it may reward either fitness equally. The fittest tension frame in figure 5.11d, for instance, predominantly presents a cliffhanger pattern,

although a decreasing tension also exists from room 3 to 7. The cliffhanger and decreasing tension are conflicting objectives, as the former rewards an increase in tension both for the presence of a peak and for the final room. Therefore, the frame in figure 5.11d attempts to balance between the two by predominantly having a decreasing pattern aggressively reducing tension to 0, once a peak has been already established early within the progression. Once tension is not able to be reduced further, a sudden tension spike satisfies the cliffhanger pattern.

Figure 5.11b shows the fittest level for the intended tension frame discussed above; its level tension progression is shown in figure 5.11d, in grey. The level tension progression matches the frame quite accurately with the exception of the first room, which due to the monster constraints is impossible to achieve. The level extensively utilised light sources to match the decay of tension with that of the frame, as can be verified from room 4 to 6. Interestingly, despite the expected differences when the level generator interprets the intended frame, the aesthetics match quite accurately between both the tension frame and tension progression. The level progression predominantly has decreasing tension, with no cliffhanger but at least one peak (thus fulfilling one of the requirements for a cliffhanger). In this particular example, the cliffhanger was cut short but given the constraint limitation would simply obtain an equal value to the highest peak, thus not fulfilling the second requirement.

The level of figure 5.11b has 4 monsters within the room progression, placed at the start, middle and end of the progression path. The three monsters in the first, second and fourth rooms trigger a very tense experience to the player, but the decreasing tension aesthetic allows them to relax for the next remaining rooms before facing the last monster within the final room.

5.3 Discussion

Results obtained from level generation shows the potential of generating levels based on a designer-defined intended structure of tension. All experiments show the GA attempting to balance monster and light source distribution, while also taking into account the decay effects, in order to approximate the progression towards the intended frame as closely as possible. Interesting and unexpected patterns also occurred, due to the inability of perfectly matching between both; allowing for a degree of control while still providing variability. The majority of experiments showcased the potential of the level generation methodology put forward, where a variety of diverging frame patterns and level sizes obtained similar results despite these variations. Overall the methodology put forth performed successfully on the majority of experiments, although certain limitations do exist. Parametric balance is an important aspect of evolutionary computation (Michalewicz,1995), and in situations where long frames were present it severely limited this balance. Certain aspects of the GA could be reworked in order to mitigate the lack of influence of f_t on longer tension frames, however we believe that a different type of representation could potentially be more beneficial. Given the importance of room and path construction, it would be beneficial to have a representation where rooms are not directly tied to the constraints of tile spaces within the map, and dependent on adjacency and connections with the rooms that are directly nearby. We hypothesize that this is the particular weakness in the current solution, and by allowing rooms to be more robust and easily constructed would significantly mitigate this limitation. Furthermore, by eliminating the bounds of the map (i.e. the hard limit on

width and height), would also allow the GA to focus specifically on the rooms instead of the tiles, without restrictions of space and adjacency. Rooms could be thought as similar to "Tetris" pieces that could be created and combined through evolutionary means. However, such a solution would mean that each genotype would have varying sizes and structures, complicating potential crossover and mutation operators. Furthermore the precise location of each room within the hypothetical map would need to be constantly tracked as well.

The last section of the chapter highlighted four example tension frames which were associated with one or multiple fitness dimensions. The results showed that the intended tension curves created by the system matched the patterns in the narrative structures they were based on. The generated levels in many cases matched the frame (if not value-for-value) but the limitations of the level progression calculation could cause deviations. Operating on its own framework and its limitations, the level generator attempts to reinterpret the tension frame by balancing the placement of monsters, light sources or empty rooms, in order to match it. At a high-level, all generated levels exhibited the intended aesthetics of each frame. Observing results with other fitness dimensions of framing, we found that *Escalating*, *Decreasing* and *Unresolved Tension* fitnesses created the least variability in the tension curves. This was expected, as these fitnesses reward small incremental changes in the tension or no changes (for Unresolved Tension). Both the *Surprising Moment* and *Resting Point* fitnesses created more variations in the tension curves but both showed similar patterns: a drastic change of tension (from 0 to 3 or vice versa) between two adjacent rooms (similar to Fig. 5.11c). This pattern is impossible to replicate in the levels, leading to more free-form interpretation of the intended curve by the level generator. An interesting emergent solution to attain less aggressive tension changes was when fitnesses were combined: for instance, combining any fitness with the Escalating or the Decreasing fitness yielded curves with smoother changes in tension. Due to a less strict evaluation formula, the *Denouement*, *Cliffhanger* and *Rising & Falling Tension* created the most diverse curves. Peaks very often varied in tension, and in some cases the entire curve would have low values of tension, or only high values. The level generation process was usually successful in translating the intent into playable levels. The formula for deriving the level tension curve is quite limiting, with a single value for increasing and two values for decreasing tension, but this constraint spurred evolution to creatively interpret the intended frame to balance the presence and absence of monsters. However, having more ways of affecting the level tension curve by different increments would likely allow for more interesting variations in the generated levels.

5.4 Summary

This chapter presented the results obtained from several experiments realized on *Sonancia*'s level generation process. The objective of these experiments was to present the capabilities of the implemented methodology in the creation of procedurally generated levels. The first section presented the results for testing the methodologies efficiency in creating levels capable of following a defined experience. Furthermore this section also presented experiments designed to test the generality of our implementation. First by showcasing different levels that were generated using the same defined experience, and then showcasing generated levels of different shapes and sizes. This chapter concludes by showcasing the system's ability as an autonomous creation system, absent of human intervention. An entire cycle, from tension frame generation to procedurally generating a level, are described for 4 diverging experiments. First the tension frame is generated utilising one or two combinations of di-

verging fitnesses, inspired by different literary styles. The tension frame is subsequently used for the generation of a fully system defined level. This showcases the process of interpreting a literary style into a framing device (i.e. the tension frame), and then interpreting the frame into a fully playable level with gameplay constraints. The final section offered discussion and reflection of all studies presented in this chapter, while presenting the pros and cons of each experiment, and offering potential solutions in fixing some of the limitations found.

Chapter 6

Modelling Affect of Audio

Audio within the digital game industry has evolved substantially over the years, providing increasing fidelity and diverging in its specific usage within the medium. Often used for the simulation of soundscapes in virtual environments, audio has also been used to complement the gameplay experience, providing an additional layer of emotional engagement during play (Collins,2013). This work will concentrate on the latter, constructing soundscapes that complement the emotional progression intended for each procedurally generated level, as a way of increasing the player engagement during gameplay. Ideally audio should follow an indented emotional progression defined by designers, so that during play both the on-screen visuals and the player's own actions sync with the audio experience. For the purposes of simplification, this work studies the construction of soundscapes that adapt to the intended emotional progression of the generated level itself, without taking into consideration the player's real-time agency.

This chapter describes the methodology followed for the construction of supervised learning models, capable of ranking different audio assets according to the tension, arousal and valence affects. These models will subsequently be used to aid the procedural level generation process for the construction of multi-faceted levels. In order to learn a relationship between an audio piece and its perceived emotion, it is necessary to construct a data sample capable of describing this mapping. Section 6.1 describes how human annotations were collected through crowdsourcing for the construction of this dataset. Furthermore, Section 6.2.1 details the methodology of extracting descriptive features of audio signals, which serves as the inputs for constructed models. The presentation of statistical results obtained from the crowdsourcing annotations (Section 6.4) and the different constructed models for two diverging experiments: the sound rank experiment (Section 6.5) and the sound & effect ranking experiment (Section 6.6), subsequently follows. This chapter then concludes with an in-depth discussion of the obtained results (Section 6.7) and summary (Section 6.8).

6.0.1 System Overview

Figure 6.1 shows the system overview utilised for the construction of preference learned audio affect models. 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

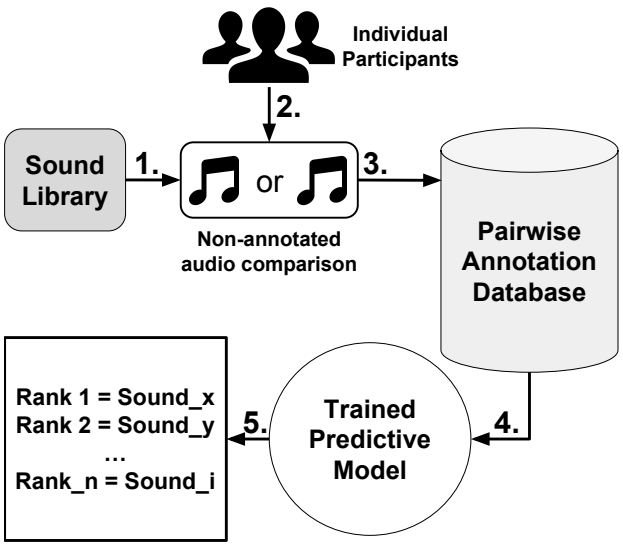


Figure 6.1: The system pipeline presented in this paper: 1) The sound library provides a pair of sounds; 2) Participants compare sound pairs based on the perceived tension, arousal and valence; 3) Participant annotations are kept in an annotation database; 4) Annotations are used to train predictive models; 5) The trained predictive models predict a global ordering of unseen sounds.

vector machine was used to create a predictive global ordering of sounds according to the perceived affect of tension, arousal and valence.

6.1 Data Collected

In this thesis data was collected for two different purposes: first for the creation of computational predictors, and second for experimental validation purposes further described in Chapter 7. Data collection is an important process for the creation of supervised learning models, due their reliance on annotated data used for guiding the learning process. Data obtained through human participants consist of controlled experiments, where they are exposed to distinct conditions for the elicitation of certain emotional states. Participants are then tasked of annotating the experience by self-reporting the perceived emotion felt or through physiological monitoring (Martínez et al.,2011). The following section will describe the methodology employed for the retention of user annotated data, to construct datasets used for the supervised learning of audio affect models.

6.1.1 The Audio Library

All audio assets were chosen from the existing database of 97 sound created assets for the horror genre. Audio files consist of short audio loops averaging 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.

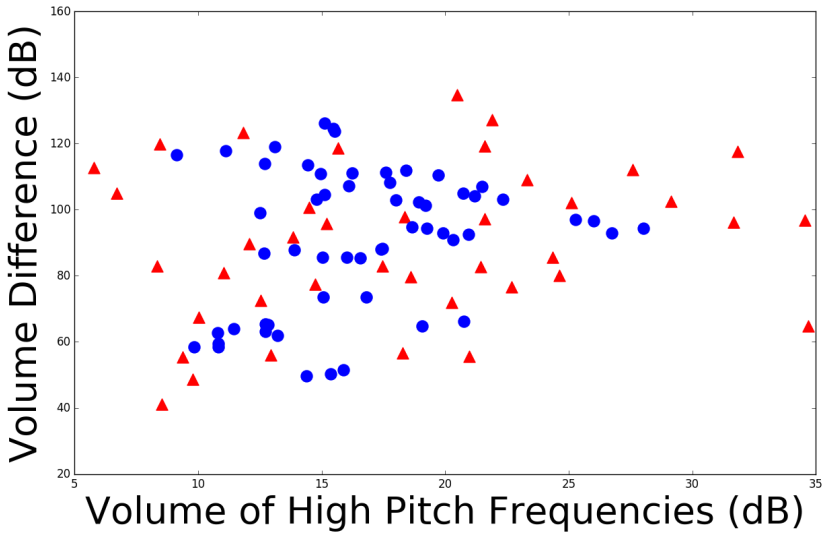


Figure 6.2: Scatter plot of the entire audio asset library. Triangles and circles are the selected and unselected audio assets, respectively.

A spectrum is the power density (measured in decibels, Db), which measures the intensity and consequently "the loudness" of each frequency band, which in turn affects the overall pitch of sound.

According to Garner et al. (2010) 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. 6.2). 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 6.2).

6.1.2 The Digital Signal Processing 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 (Stevens and Raybould,2013), for the creation of novel sounds (e.g. the roar of a dinosaur) or for adding more emphasis to the base sound conveying 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[1] for the interested reader.

In this study we intend to explore how effects can influence the perceived emotion in comparison to the original audio signal, and if a data-driven approach can potentially learn this relation between an effect, the audio piece and the perceived emotion. In this way,

[1] `https://goo.gl/kfHP7Y`

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 either heavily or lightly affect the original audio signal. 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.

6.1.3 Annotating Audio

To effectively obtain a 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 (Shaker et al.,2013;Li et al.,2013). For that purpose, a website[2] 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 logs 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 (Yannakakis and Martínez,2015; Martinez et al.,2014;Yannakakis and Martinez,2015). In the context of this thesis, 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 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

[2]http://sonancia.institutedigitalgames.com

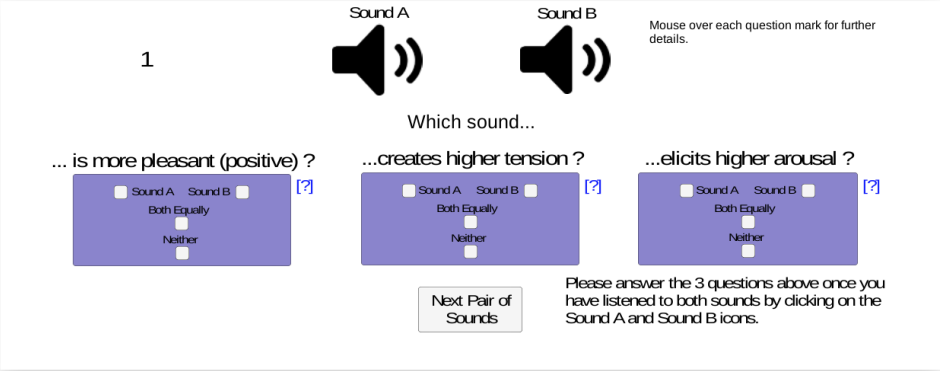


Figure 6.3: The crowdsourcing annotation tool for sounds. The top two icons allow users to select and play one specific sound of the selected pair; only one sound can play at a time to avoid cacophony. The 4-AFC questions below ask the participant to rank valence, tension and arousal, respectively. Once 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.

least amount of repeated sounds during their time annotating, 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 6.3 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 given by the permutation of n ($n = 40$), being the total number of sound assets in the library and r, the combination size ($r = 2$ being a pair): $P_2^{40} = 1560$. The total number of sound asset pairs required for the sound effect experiment consists of 1280 which is the product of 40 sounds times 32 effects per sound.

6.2 Feature Extraction

Feature extraction is the process of transforming a raw signal into a set of manageable input features used for prediction algorithms. Audio and skin conductance signals, such as the ones used in this thesis, present a large number of points over a given time. This type of signal would present both varying degrees of dimensionality, as one signal is never exactly the same in length as another, but also an abundance of inputs, which would significantly impact the efficiency of learning. The following sections describes how features were extracted from audio assets in the context of this thesis.

6.2.1 Audio Signal

Low-level descriptors (LLD) are statistical representations of information extracted from an audio signal. Several examples of LLDs include the root-mean-square level, spectral centroid, zero-crossing rate, among others (Peeters,2004). Although features can be extracted globally, i.e. over the entire signal, McKinney et al. (2003) states that for the purposes of music classification better results are achieved by segmenting the signal, with some degree of overlap between the previous and following segments, and thus extracting statistical values from these LLD vectors. The reason is due to the tendency of fluctuation present in these signals, hence this approach can more easily take into consideration the multiple variations of each unique audio signal. Usually there are three levels of extraction granularity which are associated to 1) selecting arbitrary points in the signal; 2) defining sequential regions (i.e. frames); 3) using pre-segmented regions. In this thesis we opted for the second approach due to its popularity within literature (McKinney et al.,2003; Eyben et al.,2013; Eyben,2016). Several statistical values can be derived directly from a raw LLD vector, however in most cases it is necessary to perform some degree of pre-processing such as normalization and temporal smoothing. In this work each audio sample is normalized within a range of $[-1, 1]$ where each segment, from henceforth referred to as a frame, consists of a $25ms$ fragment of the signal with an additional overlap of $10ms$. Each value of the LLD vector is then smoothed using the following temporal smoothing function:

$$x_s(n) = \frac{1}{W} \sum_{i=-(W-1)/2}^{(W-1)/2} x(n+i) \tag{6.1}$$

where $x_s(n)$ is the smoothed value of frame n and W is the number of context frames (or windows) over which the function is computed. Within this work $W = 3$ as it is considered a reasonable number for the removal of windowing artefacts (Eyben et al.,2013). A first order Δ regression coefficient is also calculated for each $x_s(n)$ value creating an additional LLD vector such as:

$$\Delta^W(n) = \frac{\sum_{i=1}^{W} i \cdot (x(n+i) - x(n-i))}{2 \sum_{W}^{i=1} i^2} \tag{6.2}$$

where a default window size of $W = 2$ is set as per the suggestion of Young et al. (2006).

For the purposes of audio feature extraction the openSMILE tool was used (Eyben et al.,2013). OpenSMILE is an open-access audio feature extraction tool that has been widely used for speech emotion recognition Schuller et al. (2011, 2012, 2013). All features extracted follow the 'INTERSPEECH 2009 Emotion Challenge' feature set of Schuller et al. (2009) consisting of 384 statistical features across 32 types of LLD vectors. The interested

reader is referred to Eyben (2016) for more in-depth and formal descriptions of each LLD used within this thesis. LLDs extracted of both the x_s and Δ^W consist of:

- The Root-Mean-Square of signal energy (RMS-Energy);
- Twelve Mel-Frequency Cepstral Coefficients (MFCC-1 to MFCC-12) at a 16-bit sample range $[-32767; 32768]$;
- Zero-Crossing Rate (ZCR);
- Fundamental frequency computed by the Cepstrum (F0);
- Probability of voicing (voiceProb)

From each LLD vector 12 statistical features are obtained, resulting in a combined feature set of 384 features. Statistical features used consist of the following:

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

6.2.2 Feature Selection

Due to an overwhelming amount of extracted features, a form of feature selection either manually or automatically is required to efficiently train a predictive model. Features extracted from audio were used to construct two diverging 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.

Table 6.1: 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)

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 (El Ayadi et al.,2011), two variants of both base and effect audio datasets were created, consisting of only the MFCC statistical features. Both SFS and SBS were also used to further reduce the dimensionality space of the datasets. These algorithms are described fully in section 3.2.

6.3 Statistical Analysis of Crowdsourced Annotations

The crowdsourcing platform was heavily disseminated over social media platforms, including Twitter, Facebook and Reddit; scientific conferences and within the University campus. 1009 annotations were 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 6.1 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 lack and defective annotations. The pruning methods used are further described below.

6.3.1 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 6.2. For the tension and arousal questions 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.

Following the methodology of Yannakakis and Hallam (2011), in order to apply supervised preference learning ambiguous annotations were discarded (i.e. Both Equally and

Table 6.2: The preference distribution of the crowdsourced sound ranking experiment.

Affect	A	B	Both Equally	Neither	Total
Tension	187	216	34	16	453
Arousal	170	219	29	35	453
Valence	168	166	10	109	453

Table 6.3: Baseline performance for each of the three affective dimensions calculated as the higher value between the times (in percentage) sound A and B was preferred.

Tension	Arousal	Valence
53.1%	54.9%	50.2%

Table 6.4: Rank-correlations of annotations between all pairs of affective dimensions.

Tension-Arousal	Tension-Valence	Valence-Arousal
0.25	-0.45	-0.13

Neither). Following pruning the total resulting base sound annotations amounted to 403, 389 and 334 for tension, arousal and valence, respectively.

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 6.2, 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.

Table 6.4 shows the rank correlation between all three affective combinations for this experiment. 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.

6.3.2 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, capable of predicting 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

Table 6.5: The preference distribution of the crowdsourced sound and effect ranking experiment.

Affect	A	B	Both Equally	Neither	Total
Tension	249	457	221	82	1009
Arousal	227	444	216	122	1009
Valence	294	238	141	336	1009

Table 6.6: Baseline performance for each of the three affective dimensions calculated as the higher value between the times (in percentage) sound A and B was preferred.

Tension	Arousal	Valence
68.1%	69.3%	57.4%

Table 6.7: Rank-correlations of annotations between all pairs of affective dimensions.

Tension-Arousal	Tension-Valence	Valence-Arousal
0.46	−0.42	−0.15

be referred to as a "base sound".

Table 6.5 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 alter 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 is specifically because the current annotation dataset associated effected sounds to sound A, which eventually influenced the participants' decision making.

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 seldom be 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 value for each of the three affective dimensions is calculated as the maximum value of two numbers: the amount of times sound A was preferred to B versus the

amount of times sound B was preferred to A. Table 6.6 shows the baseline performance obtained. 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.

Table 6.7 details the rank correlations between the different affective dimensions. Similarly to the previous experiment both the valence-arousal and the tension-valence rankings are negatively correlated, although with slightly differing results. However, the correlation between tension-arousal increases substantially from the previous experiment. This is most likely due to the influence of some effects on the volume of the base sound, which could potentially make the effected sound much louder. Louder sounds tended to be perceived as both more tense and arousing in comparison to the lower volume sound.

6.4 Global Order of Sound Rank Annotations

The 40 sounds are ranked based on the human-annotated tension preferences. The *global order* is derived through the pairwise preference test statistic Yannakakis and Hallam (2011) 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 and -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 6.4 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. 6.4). 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[3]. 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. Garner et al. (2010), 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[4]. 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

[3] `https://goo.gl/Z2ihfo`
[4] `https://goo.gl/IbY0gf`

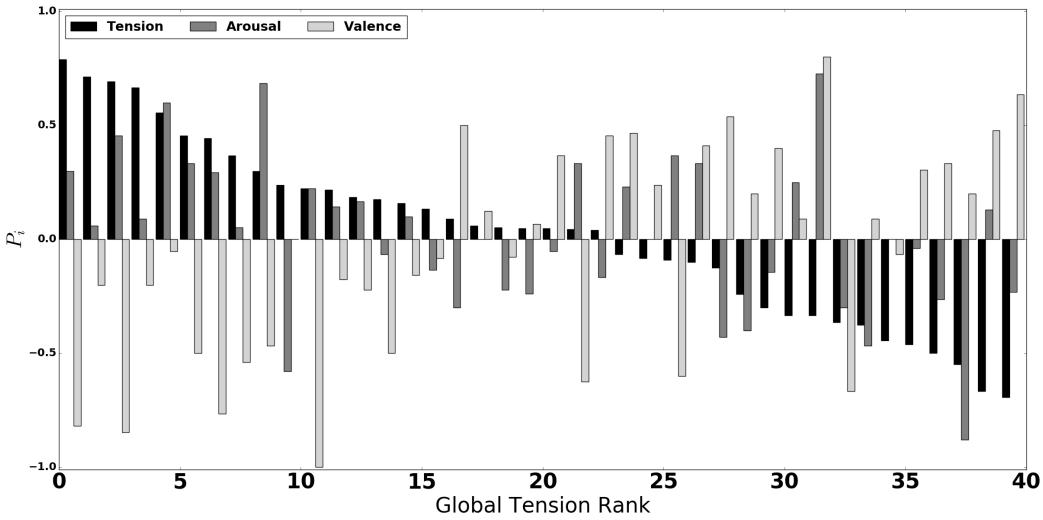


Figure 6.4: The global order and distribution of the annotated sounds in each affective dimension: tension (black), arousal (grey) and valence (white). The *y-axis* consists of the preference score value (P_i) and the *x-axis* consists of the sound rank according to the tension dimension, ordered by the most to less tense sounds.

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[5]. 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 Garner et al. (2010), and the obtained global ranks, the kendall's τ correlation coefficient was calculated Wilkie (1980). Table 6.8 shows the correlation and p-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.

6.5 Learning to Rank Sounds

The construction 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:

[5] `https://goo.gl/E7VIu0`

Table 6.8: 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)

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 section we present the best average accuracy obtained for each affective dimension but we also provide the accuracy of the best fold in parentheses.

6.5.1 Rank Support Vector Machine – Sequential Forward Selection

Figure 6.5 shows the average 5-fold cross-validation accuracy of the two different RankSVM kernels employed (Linear and RBF) using Sequential Forward Selection (SFS). 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 despite the kernel used.

Interestingly arousal was the most difficult to predict of the three affective dimensions, which was surprising considering that literature states otherwise (Yang and Chen,2011b). 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. 6.5 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 (Miranda and Castet,2014), 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. 6.5). The best average accuracy of 72% (79%) was obtained using an RBF 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

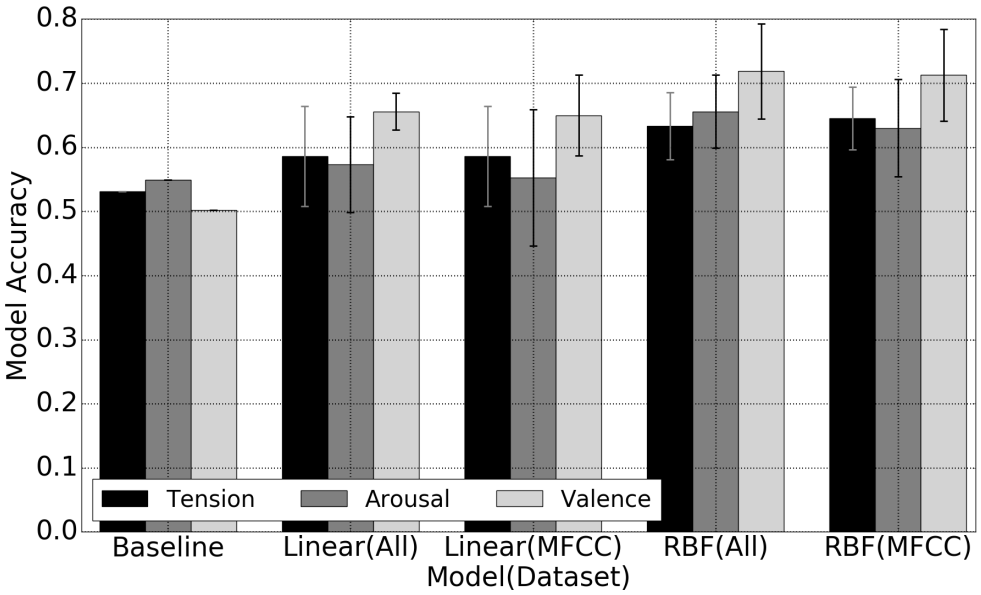


Figure 6.5: Learning to rank sound: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models, employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Forward Selection (SFS), feature selection algorithm is applied in all experiments reported. The presented accuracies for RBF consist of the best accuracy obtained through extensive parametrization testing.

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.

For brevity Table 6.9 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 6.9 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). 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

Table 6.9: 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}	$F_{Max}(\Delta MFCC_3)$ $\mu\Delta MFCC_{10}$ $\mu MFCC_6$ $Max(MFCC_1)$

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.

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.

6.5.2 Rank Support Vector Machine – Sequential Backward Selection

Similarly to the previous experiment, Fig. 6.6 showcases the average 5-fold cross-validation accuracy obtained, for both the Linear and RBF RankSVM kernels utilising Sequential Backward Selection (SBS). Tension performed slightly worse, in comparison to the previous experiment, where the best average obtained was 61.7% (67.9%), with SBS applied exclusively on the MFCC LLDs and using the RBF kernel set to a gamma value of 0.1. However, it is important to note that the models trained between the diverging datasets presented minimal differences for the tension affect, where the best average obtained from the All dataset was 61.2% (65.4%) with the same exact RBF parameters. Interestingly, using SBS slightly improved and also worsened the Linear and RBF models respectively, where the most accurate models obtained consistently stayed within the 60% to 62% accuracy range.

Out of all the three diverging affective dimensions, arousal was the only one to see a slight accuracy improvement in comparison to the previous SFS experiment, specifically with the Linear RankSVM variant. Similar to tension, the obtained accuracies tended to closely mimic models trained without feature selection, however it is noticeable that SBS did slightly influence accuracy, as an increase is present in comparison. Also the majority of models were able to predict with an accuracy above the baseline, which are rarely achieved without feature selection. The best average obtained was 65.5% (70.5%), with SBS applied on the MFCC dataset using the Linear kernel. In comparison, the All dataset performed slightly worse, where the best obtained model achieved an average accuracy of 61.9% (66.6%). Considering that the MFCC dataset contains a lower number of features which are also included in the All dataset, suggests that the SBS could have been slightly more aggressive at removing irrelevant features from the feature space, in order to obtain better overall accuracies.

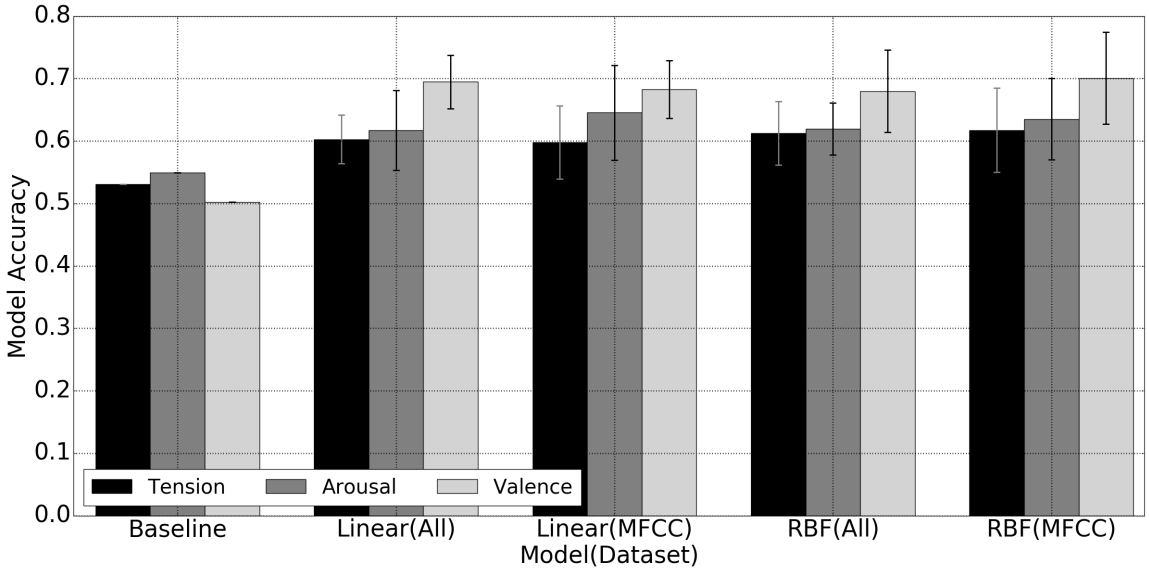


Figure 6.6: Learning to rank sound: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models, employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Backward Selection (SBS), feature selection algorithm is applied in all experiments reported. The presented accuracies for RBF consist of the best accuracy obtained through extensive parametrization testing.

The valence affect performed worse in comparison to the previous feature selection method, although the Linear variants obtained slightly better accuracies. This specific trend is shared among all three different affect models, where linearly trained RankSVMs performed better on a larger feature space than RBF. The best average obtained for valence was 70% (77.6%) with SBS applied on the MFCC data using the RBF RankSVM variant. Interestingly, applying SBS on the All dataset with Linear RankSVM obtained very similar results, with an average accuracy of 69.4% (73.1%).

Unlike SFS, SBS proved to be a much more reserved feature selection algorithm, often not even removing features from the selected dataset. This is particularly apparent in the RankSVM experiments, such as the ones presented above, where a very low number of features were removed, often using almost the totality of available features for training. Although, this does not necessarily suggest that each feature within the dataset is a necessary requirement, as some features might impact the accuracy by relatively small margins, and could also consist of "noisy" features. A more aggressive SBS algorithm might prove better than the one proposed within this thesis, such as allowing a margin of leeway for removing features where the error difference between the trained models with and without a specific feature is below a certain threshold value. Contrarily, the arousal models slightly benefited from the added number of features, which also suggests that certain affective models might improve from an added feature set in comparison, and thus a less aggressive feature selection might be favourable. A particular downside of SBS is the large computational effort required in comparison to SFS, meaning that parameter optimization can take significantly longer. This is particularly problematic when using machine learning algorithms such as artificial neural networks, where a large number of parameters exist. Hence a problem may

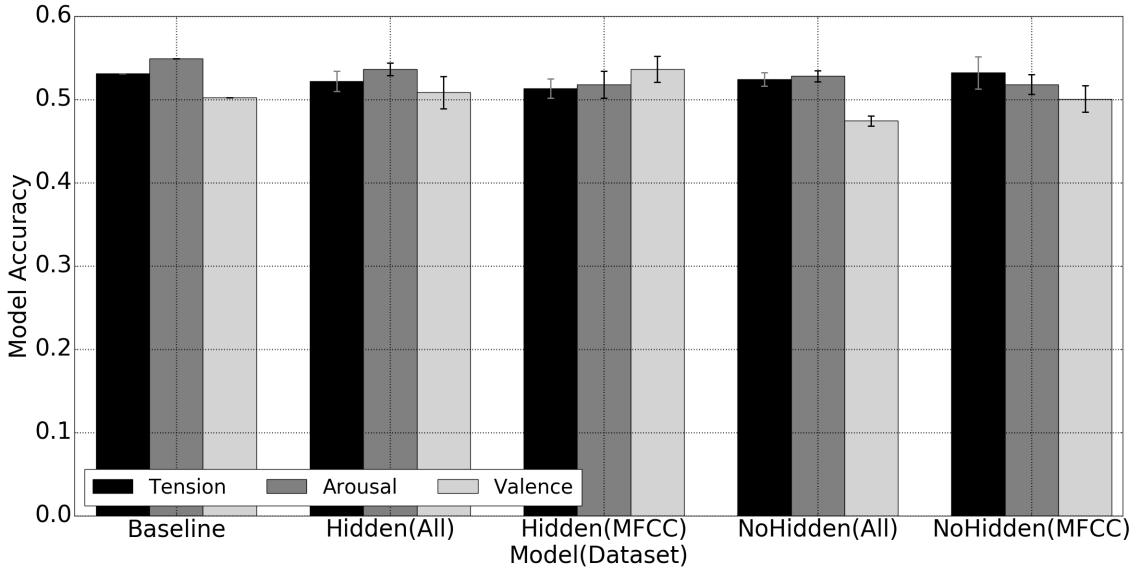


Figure 6.7: Learning to rank sound: The test accuracy mean with the standard error of 5 independent runs of the 5-fold cross-validation of ANN models, employing diverging topologies across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Forward Selection (SFS), feature selection algorithm is applied in all experiments reported. The presented accuracies for the hidden layers consist of the best accuracy obtained through extensive neuron parametrization testing.

arise, the more aggressive an SBS feature selection algorithm is, the more computational effort it will require to run each feature combination. For this reason showcasing each chosen feature would go beyond the limitations of this thesis. The total number of features selected from each best fold in the tension, arousal and valence affective dimensions are 287, 287 and 286 from a total of 288 MFCC features, respectively. It is apparent that SBS did not effectively select features, and was too reserved in its feature extraction process. The best folds utilised almost the totality of available features, which had a minor influence on the trained models. Interestingly, the best results and most variance came from models whose gamma parameters were particularly low (i.e. < 0.1). A more aggressive SBS algorithm could potentially improve these results, although at the cost of larger computational effort.

6.5.3 Artificial Neural Networks – Sequential Forward Selection

For this experiment several Artificial Neural Network (ANN) models were trained using backpropagation with the All and MFCC datasets. Unlike RankSVMs, ANNs rely on the initial randomization of weights, and for this reason ANN models with identical parameters were run independently 5 times. This specific number was chosen in order to keep the computational effort low, while still providing some preliminary statistical results as both feature selection algorithms, particularly SBS, can exponentially increase the number of hours required for just a single run to complete. Learning rate is set to 0.1, while the number of epochs to 200. Several different topologies were experimented including without a hidden layer (NoHidden), or with exactly one hidden layer (Hidden).

Figure 6.7 shows the average 5-fold test accuracy over 5 independent runs of ANNs trained utilising SFS on both the All and MFCC datasets. It is immediately apparent that

ANNs performed significantly worse than RankSVMs in all three affective dimensions, where the majority of models rarely achieved or surpassed the baseline values. In-line with the literature, these results reinforce the reasoning behind why SVM type algorithms are often favoured, such as the work of Yang and Chen (2011a,b), against other machine learning algorithms such as ANNs. Furthermore, it is apparent that the majority of ANNs did not effectively learn, as the hidden layers obtained very similar accuracy values to the non hidden layer experiments. Although several diverging and more aggressive parameters were tested in addition to those reported in this thesis, no significant observable improvements were obtained.

Given the ambiguity of emotion recognition, specifically in the context of sound, we hypothesize that to efficiently train an ANN capable of such predictions, a larger subset of training data would be required. This limitation severely hinders the viability of utilising ANNs for sound emotion prediction models, which depend on ambiguous and contradictory human participant annotations of affect. Given this limitation, and the overwhelming number of audio features extracted from audio, it is difficult for an ANN to effectively obtain accurate results on small datasets.

For the tension affect, the best average obtained was 53.2% (64.1%) on a no hidden layer ANN topology, utilising the SFS feature selection algorithm on the MFCC LLD features. It is important to note that the remaining models did not significantly differ, where similar accuracies ranging from 48% to 53%, were also achieved by models with diverging topologies. In the majority of topologies tension did not manage to surpass the baseline, capable of only matching it. Comparatively to the other affect models, tension obtained similar results with minor differences.

The best obtained average for the arousal affect was 53.6% (64.1%) utilising a hidden layer with 25 neurons and the SFS feature selection algorithm on the All dataset. Unlike tension, arousal was not capable of matching the baseline, consistently obtaining average accuracies slightly below it. However arousal did achieve better results with the hidden layer, although the accuracies did not diverge substantially across the different experimental runs.

The valence affect produced the most variance out of three. The least accurate models consisted of topologies without a hidden layer, where the lowest obtained accuracy consisted of models trained with the All dataset. The best average accuracy obtained for valence was 53.6% (65.7%) using a hidden layer with 25 neurons and the SFS feature selection algorithm on the MFCC dataset.

Table 6.10 showcases the selected features of the best fold from each affective dimension with the most accurate average. The arousal model was the only exception, where the best results were obtained using the All dataset. Similarly to the RankSVM experiments, the arousal affect models consistently utilised RMSEnergy type features, which were rarely used by the other affect models. Valence and tension also repeated similar trends from the RankSVM experiments, where the majority of features chosen were MFCC related, where the most accurate folds exclusively utilised the MFCC dataset.

6.5.4 Artificial Neural Networks – Sequential Backward Selection

In order to compare the influence of feature selection on the accuracy of diverging ANN models, a new experiment was conducted. Several ANN models were trained in the same conditions of the previous SFS ANN experiment, where the only exception was the usage of SBS rather than SFS. Figure 6.8 presents the obtained average and standard error of the 5-fold cross-validation test accuracy over 5 independent runs, where each ANN model

Table 6.10: The selected features of the most accurate fold with the best obtained average accuracy model parameters of each affect.

	Tension	Arousal	Valence
Model	NoHidden(MFCC)	Hidden(All)	Hidden(MFCC)
Selected Features	$apr_e(\Delta MFCC_3)$ $\mu MFCC_2$ $F_{Min}(\Delta MFCC_{12})$	σR_{Egy} $Min(MFCC_2)$ $\sigma F0$ $F_{Max}(\Delta MFCC_8)$ $Min(MFCC_6)$	$F_{Max}(MFCC_5)$ $\mu MFCC_1$ $\lambda MFCC_2$ $F_{Max}(MFCC_1)$ $apr_s(\Delta MFCC_2)$ $apr_e(\Delta MFCC_3)$ $apr_e(MFCC_2)$ $F_{Min}(\Delta MFCC_{10})$ $Max(\Delta MFCC_8)$

is trained with features selected by the SBS feature selection algorithm from either the All or MFCC datasets. Although RankSVMs still outperformed ANNs for all affects, the SBS algorithm did however improve the general accuracy of ANNs in comparison to using SFS. Particularly, by using SBS it is apparent that some degree of learning was achieved, where the hidden layer topology consistently outperformed models without a hidden layer, while also obtaining accuracies above the baseline, which was rarely achieved by SFS models.

We hypothesize that the discrepancy between both feature selection algorithms, can be inherent to how SFS specifically selects features. Considering that SFS is a particularly greedy algorithm, where features are added based exclusively on how it influences the obtained accuracy in comparison to other features, can substantially increase the probability of these models overfitting towards the chosen feature subset. More precisely, overfitting occurs when a model is too closely related to the training data failing to learn a more general trend, which then performs poorly on unseen data points. By comparing the training and validation accuracies this hypothesis was validated, where the accuracy difference between training and testing ranged from 10% to 20%, with the exception of the no hidden layer topology where models tended to underfit. Contrarily, this trend was less noticeable in the SBS models, where the accuracy difference between both training and test sets were often below 10%. Given that SBS is a more reserved feature selection algorithm compared to SFS, does suggest that the larger feature set allowed ANNs to retain a higher degree of robustness, which in turn allowed these models to perform better on unseen data.

Out of all three affects tension performed worse, where a slight increase in the average accuracy was observed in relation to the previous SFS experiment. The best average accuracy obtained for tension was 56.5% (65.4%), with a topology of 25 neurons in the hidden layer and SBS applied on the MFCC dataset. Although tension performed slightly better with SBS, the overall increase was not substantial comparatively to both arousal and valence models. A common trend found among both the RankSVM and ANN experiments, including the SFS and SBS variants, was that tension proved to be the most consistently difficult affect to learn relatively to both arousal and valence.

For arousal the best average accuracy obtained was 59.7% (68.1%) with a topology of 50 neurons in the hidden layer and SBS applied on the MFCC dataset. Similarly to the other affect models, arousal achieved higher accuracies and consistently surpassed the baseline

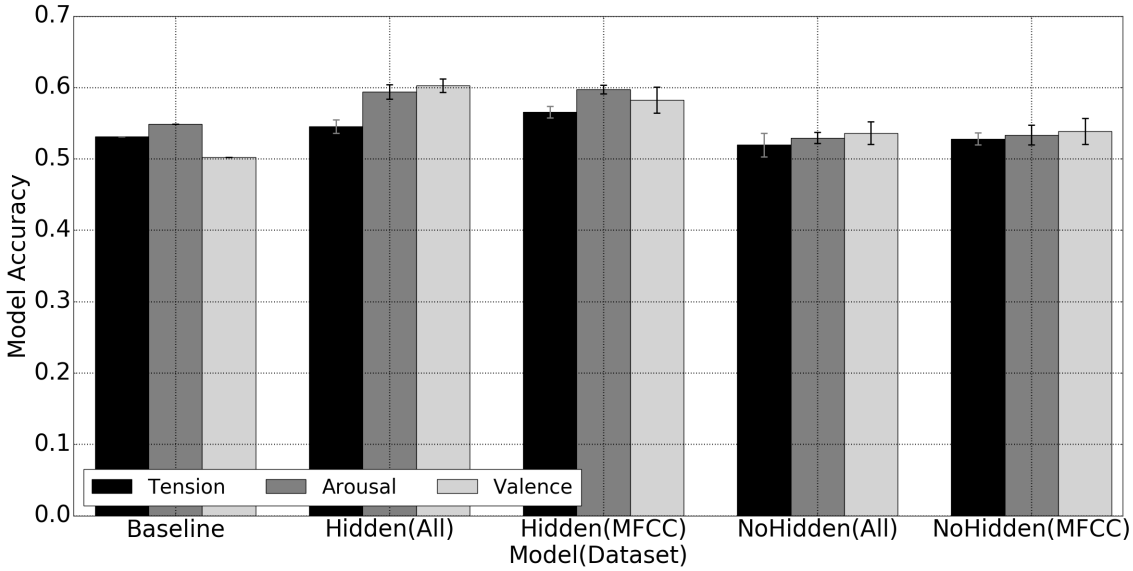


Figure 6.8: Learning to rank sound: The test accuracy mean with the standard error of 5 independent runs of the 5-fold cross-validation of ANN models, employing diverging topologies across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Backward Selection (SBS), feature selection algorithm is applied in all experiments reported. The presented accuracies for the hidden layers consist of the best accuracy obtained through extensive neuron parametrization testing.

value with a one hidden layer topology. On the other hand valence achieved an average accuracy of 60.2% (73.1%), also with a topology of 50 neurons in the hidden layer, where SBS was used on the All dataset. Similar to the previous experiments, valence continued the trend of achieving the best average accuracy in comparison to the other two affects.

Similarly to the previous RankSVM SBS experiment, the features selected for training models consisted of almost the totality of available features in either the All or MFCC datasets. However, unlike the previous experimental results the reserve nature of the proposed SBS implementation yielded better accuracies overall in comparison to SFS. This suggests that the algorithms reserve nature at removing features, thus allowing larger feature set, made it difficult for models to overfit the features around the training data. The best obtained fold for tension and arousal retained 286 and 285 out of a possible 288 MFCC features, respectively, while valence only removed one feature, retaining 383 out of a total 384 features in the All dataset.

6.6 Learning to Rank Sounds and Effects

This section presents the predictive accuracy obtained from training various SVM and ANN models with either SFS or SBS feature selection algorithms, capable of ranking both base sounds and how their perceived affect is influenced by different effects, and between the base sounds.

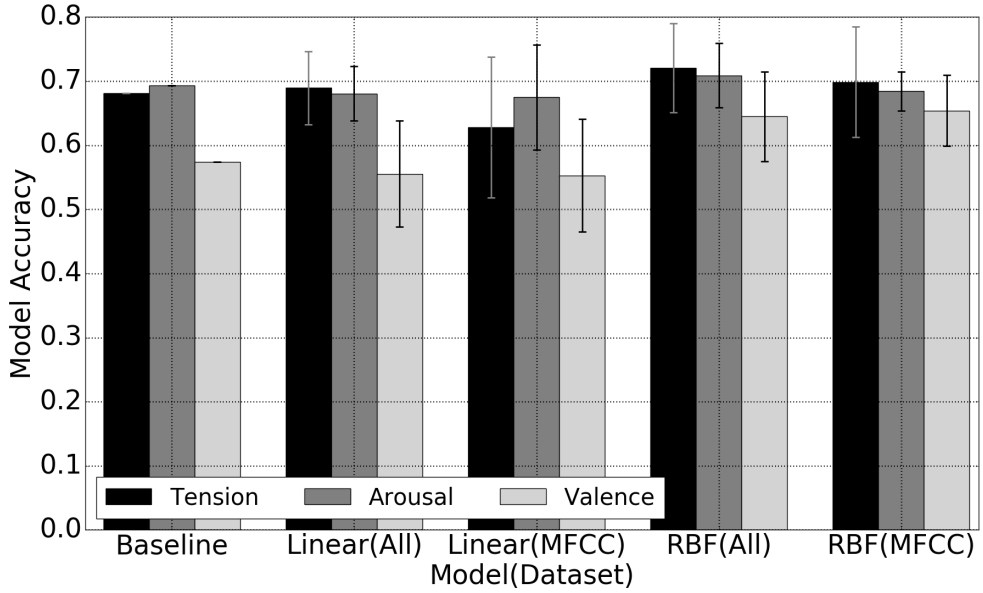


Figure 6.9: Learning to rank sound and sound effects: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Forward Selection (SFS), feature selection algorithm is applied in all experiments reported. The presented accuracies for RBF consist of the best accuracy obtained through extensive parametrization testing.

6.6.1 Rank Support Vector Machine – Sequential Forward Selection

Figure 6.9 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 to loudness (Miranda and Castet,2014). 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 aver-

Table 6.11: 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	$M_{in}(R_{Egy})$ $apr_s(MFCC_9)$ $Effect_2$ $F_{Max}(MFCC_6)$ $F_{Min}(MFCC_{11})$ $\sigma F0$ $M_{in}(MFCC_2)$ $Rg(MFCC_4)$ $\sigma MFCC_3$ $\mu VProb$ $M_{ax}(MFCC_6)$ $\sigma\Delta MFCC_6$	$M_{in}(R_{Egy})$ $\mu MFCC_7$ $apr_e(MFCC_1)$ $\mu\Delta MFCC_4$ $apr_e(ZCR)$	$kt\Delta MFCC_2$ $kt\Delta MFCC_3$ $apr_e(MFCC_{11})$ $apr_o(\Delta MFCC_{12})$ $F_{Max}(MFCC_6)$ $Rg(MFCC_4)$ $Rg(MFCC_8)$

age 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.

Similarly to section 6.5.1 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 index binary number.

In this particular experiment the SFS was less biased towards MFCC statistics, even though it is quite substantially present. Interestingly the effect input binaries did not prove to be particularly helpful for affect prediction, with only the tension model taking it into account. Additionally, in the majority of RBF(All) models presented related statistical features to R_{Egy} much 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 (Garner et al., 2010), while a too high or low pitch can cause a sense of discomfort impacting the valence state.

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 6.11 are analysed. Table 6.12 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 6.12 can be listened to here [6].

[6] `https://goo.gl/72dFd9`

Table 6.12: 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 affect RankSVM models, using the SFS feature selection algorithm.. For brevity the highest ranked effect or base sound is chosen for analysis.

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

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 did not see much influence from effects, where the base sound is often within the general rank vicinity of its effects, which tend to have a minor impact on the base sound.

6.6.2 Rank Support Vector Machine – Sequential Backward Selection

In order to compare the performance of both feature selection algorithms an additional RankSVM experiment was conducted with same conditions of the previous RankSVM experiment, where the only exception was the feature selection algorithm. Figure 6.10 presents the average accuracy obtained for tension, arousal and valence over 5-folds. Surprisingly, the majority of models consistently obtained averages above the baseline, despite the variant utilised. Particularly, the valence models benefited substantially from the added number of features, where in the previous SFS experiment valence models only surpassed baseline values once the RBF kernel was applied. The tension models also similarly improved, consistently outperforming the other two affects despite the RankSVM variant used, an aspect that had not been seen in previous experiments, where tension often underperformed in comparison. Arousal models performed slightly better than SFS, although without major differences comparatively to the previous experiment.

The tension Linear and RBF RankSVM model variants obtained almost identical performances, albeit each on two diverging datasets. More precisely, the best average accuracy obtained was 72.7% for the Linear(All) and the RBF(MFCC) using a gamma value of 0.01, where the SBS feature selection algorithm was applied to both. Although technically Linear(All) performed better due to the rounding error of decimal values, it is insignificant comparatively to the results obtained from the RBF(MFCC). The best folds obtained from both the Linear(All) and RBF(MFCC) were 79.1% and 80%, respectively. Although these results do not significantly differ from the best average obtained using SFS, the added features did particularly influence the majority of the models obtained, in particular Linear models.

Comparatively to the other affective states, arousal did not vary significantly from the previous experiment, although the majority of folds using SBS were capable of more ef-

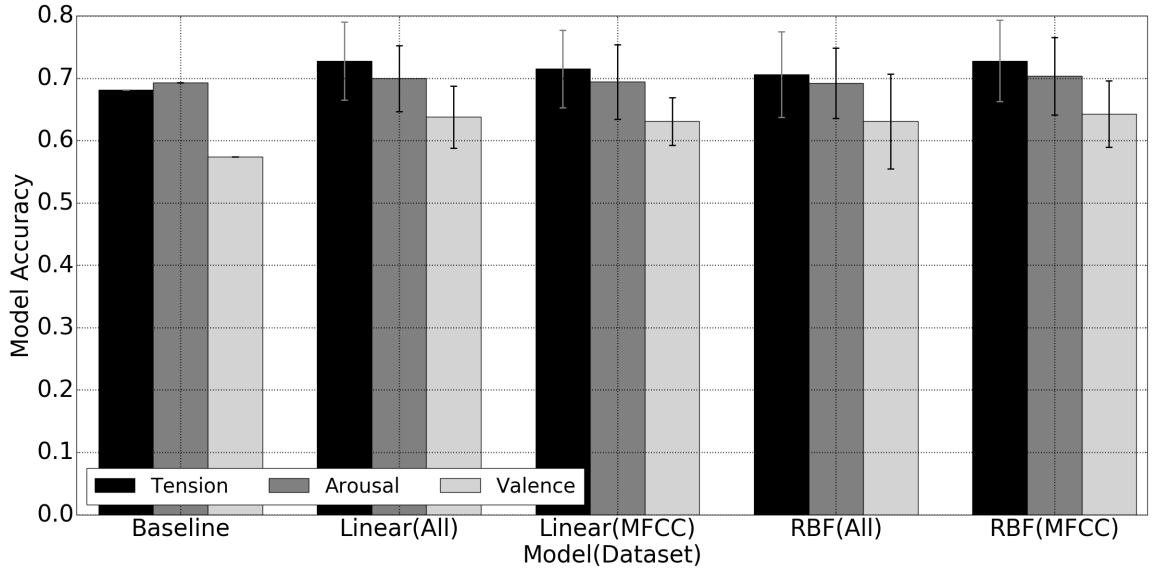


Figure 6.10: Learning to rank sound and sound effects: The test accuracy mean and 95% confidence intervals of the 5-fold cross-validation of RankSVM models employing two different kernels (Linear and RBF) across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Backward Selection (SBS), feature selection algorithm is applied in all experiments reported. The presented accuracies for RBF consist of the best accuracy obtained through extensive parametrization testing.

fectively matching the baseline accuracy. The best average accuracy obtained was 70.3% (75.4%) with RBF set to a gamma value of 0.01 and SBS applied to the MFCC feature set. A difference of approximately 1% between the best average for the SFS experiment was obtained, further suggesting that both feature selection algorithms performed similarly, despite the substantial difference of features used.

Valence models utilising SBS performed slightly better across the different RankSVM variants in comparison to SFS, where models consistently surpassed the baseline values which was previously only achieved by using the RBF kernel. The best average accuracy obtained was 63.7% (71.7%) with an RBF kernel set to a gamma value of 0.01 and SBS applied exclusively on the MFCC features. Although improvements were seen across the different RankSVM types, particularly the Linear variant, applying SFS with the RBF kernel still proved superior, albeit the difference was of only 2%. Interestingly applying SBS did improve the overall accuracy, even though the number of removed features was low it did influence models to improve past the 60% mark, which models without any feature selection applied rarely achieved.

Continuing the trend within previous SBS experiments, very few features were removed from the dataset, where almost the totality of both feature sets were used for training. In particular the best fold obtained for tension used 386 from a total of 388 features in the All dataset. The best folds for arousal and valence used the MFCC feature set, where 289 and 290 features were used out of a total of 291.

Table 6.13: 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 affect RankSVM models, using the SBS feature selection algorithm. For brevity the highest ranked effect or base sound is chosen for analysis.

	Tension	Arousal	Valence
Effects	Rk 1 (Reverb)	Rk 1 (Reverb)	Rk 1 (Reverb)
	Rk 14 (Echo)	Rk 2 (Reverb)	Rk 2 (Reverb)
	Rk 28 (Chorus)	Rk 4 (Reverb)	Rk 3 (Reverb)
	Rk 30 (High Pass)	Rk 29 (Reverb)	Rk 6 (Reverb)
Base Sound	Rk 51	Rk 6	Rk 33

Rank Comparison of Sound and Effects

Table 6.13 shows the rankings of a base sound and its 4 highest ranked effects within the predicted global rank, of the best fold obtained for tension (Linear), arousal and valence. For the interested reader all sounds presented in table 6.12 can be listened to here [7].

Although sounds differed between the predictive tension rankings of the SFS and SBS models, the top ranked sounds featured similar characteristics. Particularly, the highest ranked sounds often consisted of higher pitch variations of the base sound. In the previous SFS example, effects tended to lower the pitch from the base sound often removing the screeching characteristics of the sound. In this particular example effects worked in an opposite fashion, where they tended to increase the pitch of the base sound making the sound more uncomfortable to listen, where the chorus effect was the exception. The arousal predictive rankings were similar to the previous experiment, where the base sound picked was identical. Furthermore the effected sounds were also similarly closely ranked. Interestingly for the valence predictive rankings, one particular sound dominated the highest ranks, where the effects helped the sound become more pleasurable by eliminating some of the higher frequencies, and through the application of reverberation it effectively elongated this particular sound and added several characteristics that provided a sense of "mystery".

6.6.3 Artificial Neural Networks – Sequential Forward Selection

This section presents the obtained accuracies of different ANN models trained using both the sound and effect participant annotations. Alike previous ANN experiments presented in this chapter, all models were run independently 5 times on both the All and MFCC feature sets, utilising either the SFS and SBS (Section 6.6.4) feature selection algorithm. All parameters remained unchanged from the previous experiments.

Figure 6.11 showcases the best average accuracy over 5-folds of the diverging ANN variants. Preliminary analysis suggests a similar pattern found in the previous ANN ranking sound experiment, with ANNs significantly underperforming in comparison to RankSVMs. Another similarity was that ANNs with hidden layers tended to significantly overfit with SFS. Even though overfitting frequently occurred with RankSVM models, the difference between test and validation, through preliminary analysis, tended to range from 2% and 9%. In this particular experiment ANNs tended to overfit substantially higher, where arousal models presented the highest discrepancy between test and validation. On the other hand

[7] `https://goo.gl/vPaoIW`

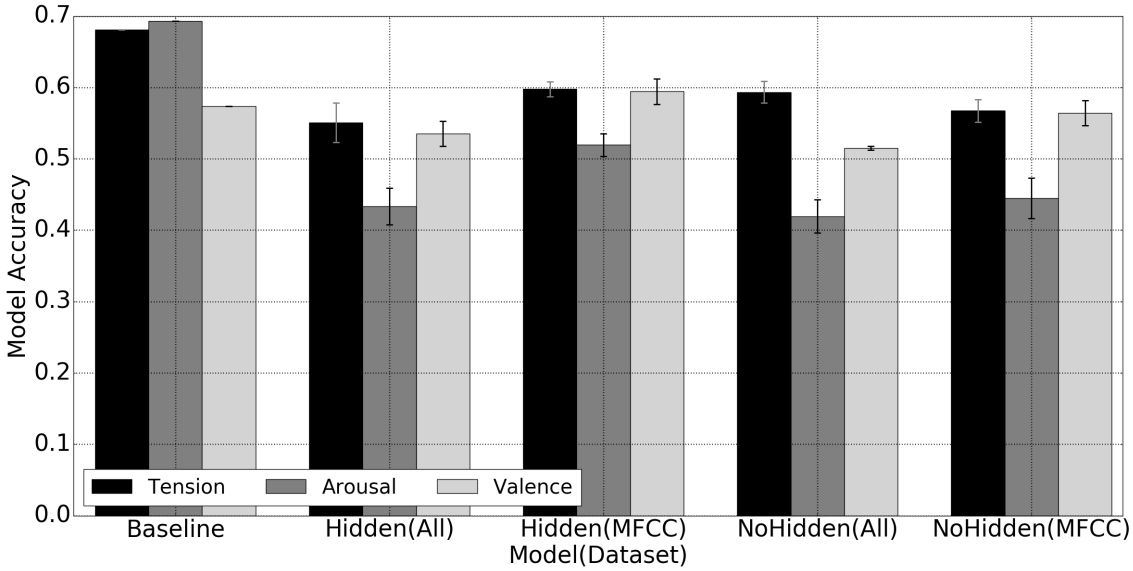


Figure 6.11: Learning to rank sound and sound effects: The test accuracy mean with the standard error of 5 independent runs of the 5-fold cross-validation of ANN models, employing diverging topologies across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Forward Selection (SFS), feature selection algorithm is applied in all experiments reported. The presented accuracies for the hidden layers consist of the best accuracy obtained through extensive neuron parametrization testing.

tension and valence presented the least amount of overfitting, where tension in particular obtained validation and test discrepancies around the same values of the RankSVM experiments. This aspect is also reflected in the results obtained, where tension and valence models tended to obtain better accuracies in general than arousal.

For tension the best average obtained was 59.7% (73.8%) with a hidden layer topology of 25 neurons, where SFS was applied on the MFCC feature set. The no hidden layer topology also performed significantly well by applying SFS on all available features, obtaining a 59.3% (69.3%) average accuracy. Unlike the previous ANN SFS experiments, the no hidden layer rarely underfit during training obtaining accuracies above 5% to 10% of validation.

SFS was particularly detrimental to the arousal affect, where the majority of models underperformed in comparison to the other two affective dimensions. The best average accuracy obtained was 51.9% (71.4%) with a hidden layer topology of 50 neurons, and the SFS feature selection algorithm applied on MFCC features exclusively. Overfitting was a particular problem for this affective dimension, where some models obtained discrepancies of over 20% between validation and testing. Although this particularity was observed in the previous ANN experiments, the skewness present in the data for arousal annotations also potentially influenced the tendency for these models to overfit. Certain tendencies which are present within the training data are likely influenced by specific features, which in turn influences the SFS algorithm. By optimizing the training data towards these specific features, constructs a model capable of optimizing efficiently towards seen data, but inefficient on unseen data with slightly diverging tendencies.

Valence performed similarly to tension, where the best accuracies obtained for both affective dimensions were from models that used exclusively the MFCC LLDs. The best

Table 6.14: The selected features of the most accurate fold with the best obtained average accuracy model parameters of each affect.

	Tension	Arousal	Valence
Model	Hidden(MFCC)	Hidden(MFCC)	Hidden(MFCC)
Selected Features	$apr_s(MFCC_7)$ $Effect_2$ $\sigma\Delta MFCC_3$ $M_{in}(MFCC_8)$ $Effect_0$ $F_{Max}(MFCC_9)$	$apr_o(\Delta MFCC_4)$ $Effect_2$ $Effect_0$ $M_{in}(MFCC_1)$ $\sigma MFCC_{11}$ $F_{Max}(MFCC_8)$ $M_{ax}(\Delta MFCC_{11})$	$apr_o(\Delta MFCC_1)$ $F_{Max}(MFCC_1)$ $M_{in}(MFCC_6)$ $ktMFCC_2$

Table 6.15: 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 affect ANN models, using the SFS feature selection algorithm. For brevity the highest ranked effect or base sound is chosen for analysis.

	Tension	Arousal	Valence
Effects	Rk 48 (Low Pass)	Rk 10 (Chorus)	Rk 1 (Reverb)
	Rk 88 (High Pass)	Rk 14 (Chorus)	Rk 26 (Reverb)
	Rk 108 (Chorus)	Rk 16 (High Pass)	Rk 27 (Reverb)
	Rk 114 (High Pass)	Rk 17 (Chorus)	Rk 31 (Reverb)
Base Sound	Rk 1	Rk 1	Rk 933

average accuracy obtained was 59.9% (72.9%) with a one hidden layer topology of 25 neurons and SFS applied to the MFCC feature set.

Table 6.14 showcases the selected features of the most accurate fold, with the best average accuracy for each affective dimension. Considering that all models obtained the best accuracies with the MFCC feature set, it is unsurprising that the majority of selected features consists of MFCC statistical features. Interestingly, valence was the only affective dimension to discard the sound effect descriptors, while both tension and arousal utilised the same 2 (out of 3) effect descriptors as features.

Rank Comparison of Sound and Effects

A comparison between the highest ranked effects and their respective base sound is shown in Table 6.15. These rankings are derived from the predictive global order of the best fold for each affective dimension. For the interested reader all sounds presented in Table 6.15 can be listened to here [8].

Tension in this particular example ranked all effected sounds lower than the base sound, whilst arousal also did the same, with the only exception being that effected sounds were closer to the base sound and higher in the rankings. Filtering effects were particularly influential for tension, where an apparent influence on the overall volume of the sound can be discerned. The reason for this is that filtering effects tend to remove frequencies

[8] `https://goo.gl/pZhFiJ`

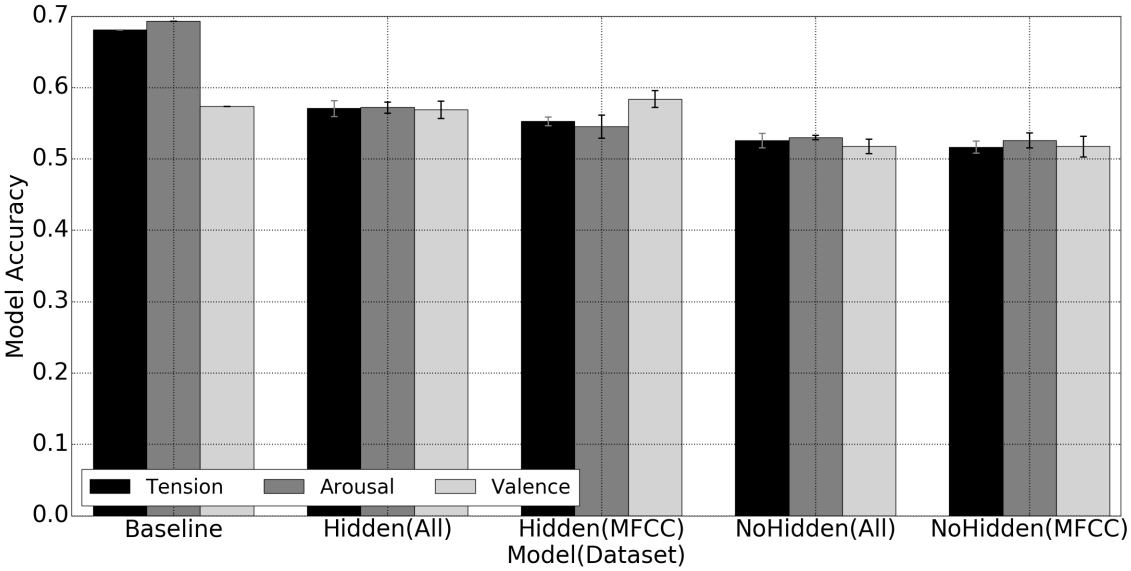


Figure 6.12: Learning to rank sound and sound effects: The test accuracy mean with the standard error of 5 independent runs of the 5-fold cross-validation of ANN models, employing diverging topologies across two different sound features: all features (All) and only the MFCC features (MFCC). The Sequential Backward Selection (SBS), feature selection algorithm is applied in all experiments reported. The presented accuracies for the hidden layers consist of the best accuracy obtained through extensive neuron parametrization testing.

at certain defined thresholds, which can both influence volume and the audible pitches of the base sound. Arousal on the other hand was particularly influenced by chorus type effects, which can also heavily impact volume and the harmonic features of the sound itself. Like many other effects the chorus can vary substantially between different base sounds, where it is often dependent on certain characteristics of sound being available, particularly harmony. Contrarily valence showed a larger discrepancy between the base sound and its effects within the predictive ranking, where effected sounds ranked significantly higher than the base sound. The reverberation effects similarly to the previous experiments also ranked highly, suggesting that these types of effects can be particularly influential on this affective dimension.

6.6.4 Artificial Neural Networks – Sequential Backward Selection

Figure 6.12 presents the average 5-fold accuracies obtained from ANN models using the SBS feature selection algorithm on both All and MFCC features. In this particular set of experiments the arousal models were able to match the accuracies of the other affective dimension models, performing substantially better than the previous SFS experiment. Although in general ANNs still underperformed relatively to RankSVMs with either feature selection utilised.

For tension the best average accuracy obtained was 57.1% (64.8%) with a topology of one hidden layer with 25 neurons, where SBS was applied on the entire set of features. In particular SFS achieved better accuracies than SBS, where the smaller subset of features chosen by the latter tended to outperform the larger more robust subset chosen by SBS.

Table 6.16: 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 affect ANN models, using the SBS feature selection algorithm. For brevity the highest ranked effect or base sound is chosen for analysis.

	Tension	Arousal	Valence
Effects	Rk 1 (Reverb)	Rk 1 (Reverb)	Rk 1 (High Pass)
	Rk 24 (Pitch Shift)	Rk 6 (High Pass)	Rk 15 (Reverb)
	Rk 81 (Reverb)	Rk 7 (Reverb)	Rk 30 (Chorus)
	Rk 83 (Reverb)	Rk 8 (High Pass)	Rk 41 (High Pass)
Base Sound	Rk 93	Rk 5	Rk 87

Arousal was the only affective dimension that saw an increase in performance, relative to the previous SFS experiment. The best average obtained was 57.2% (67.9%) with a one hidden layer topology of 50 neurons, with SBS applied on the All feature set. SBS performed substantially better than SFS, where the larger subset prevented training models to substantially overfit. Unfortunately, the majority of arousal models trained still consistently underperformed comparatively to the baseline.

The best average accuracy obtained for valence was 58.4% (67.1%) with a one hidden layer topology of 25 neurons, where SBS was applied on the MFCC statistical features. In the same vein of tension, SBS models obtained slightly worse accuracies comparatively to SFS, although not significantly.

In terms of features selected SBS still consistently removed a very minimal number of total features from either MFCC or All datasets. The most accurate ANN fold for both tension and arousal used 384 out of a total of 388 features in the All dataset. The best fold obtained for valence exclusively used MFCC features, where only one feature was removed from the total 291 features available.

Rank Comparison of Sound and Effects

Table 6.16 presents the highest ranked sound in relation to its base sound and other high ranked effects. Rankings are obtained from the predicted global rank of the most accurate fold of each affective dimension. For the interested reader all sounds presented in table 6.16 can be listened to here [9].

The chosen tension sounds specifically exemplifies how effects increased the pitch of the base sound, which in itself already presented high frequencies. The screeching characteristics of effected sounds such as the ones ranked 1st and 24th go in-line with the literature (Garner et al.,2010), where higher pitch sounds tend to be perceived as more tense. Arousal interestingly, also placed higher pitched sounds at the top of the rankings, where volume in this particular example was less of a factor than previous examples. Valence on the other hand ranked sounds with a lower volume higher, where effected sounds often presented a quieter version of the base sound, a tendency that has also been consistent in previous experiments.

[9] `https://goo.gl/RR6pse`

6.7 Discussion

Music-elicited emotion recognition is a complex task due to the ambiguous nature of human emotions and the 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 (2000) 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 in the learn to rank sound and effect experiment.

For the RankSVM 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 ($\approx 1\%$). 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.

The lack of annotations was particularly detrimental for ANN trained models, where the average validation accuracy often ranged around 50% mark, i.e. random prediction. Despite heavy parametrization tweaking little change was obtained in the overall accuracies. Applying SBS did increase the accuracy in relation to SFS, but it was still insufficient compared to RankSVM models. Similarly to RankSVMs, the added complexity of participant annotations made it difficult for ANN models to learn a relation between features and annotations. A larger training set could potentially improve accuracies from ANN models, as more data could help reduce the ambiguity seen within the data. However, we still remain sceptical as annotations of emotion in sound is a highly ambiguous task, which to effectively train ANNs would require a significant dataset. This in conjunction to the large quantity of features makes this particular problem difficult for such models.

The success of RankSVMs against ANNs lies in the fact that SVMs use the training data directly, where "prototypical" hyperplanes are derived from how data points are arranged in space. This allows SVM models to easily construct a metaphorical division of data, allowing

it to make more accurate assumptions between two diverging feature vectors. ANNs on the order hand train models through a "trial and error" optimization, where the relation between features and annotations are learned by optimizing the weights based on the error. This makes ANN models particularly data dependent, where the data itself and even the ordering can influence how the model is trained. In a way RankSVMs circumvents this limitation by optimizing the hyperplane towards training data and not an error function, and thus assumes that unseen data will be roughly similar.

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.

The underperformance of ANNs persisted in the sound and effect experiment, where the majority of average accuracies obtained from ANN trained models were below 60% and unable to match the baseline. Although some affective dimension models benefited from diverging feature selection algorithms more than others, the accuracies still underperformed in comparison to RankSVM. The best average accuracy obtained for tension was 59.7% a substantial difference from the best obtained accuracy with RankSVM models ($\approx$ 72%). This trend continues in other affective dimension experiments, where RankSVMs consistently outperformed ANN trained models by a margin of 10 to 20 percent. Arousal in particular achieved 57.2% accuracy only with the SBS feature selection, while valence only accomplished an average accuracy of 59.9% with SFS.

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.

Crowdsourced 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 presented two diverging strategies (i.e. more vs less aggressive), more types of sound features could be investigated. The work of Schmidt and Kim (2011) suggests the use of deep belief networks for the selection of optimal music emotional features. Although for this particular problem training a deep learning model would

be impossible, as it would require significantly more data than what is currently available. However, using a pre-trained model could yield a better set of features than the ones used in this thesis, and more specific to audio emotion recognition.

As a potential final step towards realising affective interaction via sounds in horror games, additional applications utilising the presented predictors could be used for developed tools capable of assisting sound designers directly. Such models could 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 (see Chapter 7). Other potential application domains include experience-driven generated games (Yannakakis and Togelius, 2011) 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 (Holmgård et al., 2015), for realising effective game-based learning (Lopes et al., 2014; Khaled and Yannakakis, 2013) or alternatively for enabling an AI-assisted game design approach (Yannakakis et al., 2014) that can suggest soundscapes which are expected to elicit particular emotive patterns.

6.8 Summary

This chapter showcased the overall process used in this thesis for the construction of audio affect models within the genre of horror, and their subsequent results. The first half of this chapter defined the methodology for the construction of crowdsourced machine learned audio affect predictors, where an overview of the model construction pipeline was initially described. An in-depth description of how data was collected through the crowdsourcing system subsequently followed, including how audio and effects were selected for annotation. Additionally the crowdsourcing experiment protocol was then defined for two types of experiments: one comparing between base sounds, and another comparing between the same base sound, where one is influenced by an effect. Following, the details of the step-by-step audio ranking process and UI were presented, where the details of how important information was displayed and each sound played to annotators. In order to efficiently learn the relations between annotations and sound, key characteristics capable of describing different sound types are necessary. The exact methodology used to extract features from the selected audio assets was described, where each feature consists of a statistical descriptor of the raw audio signal. For experimentation, two feature datasets were created: the All data set consisting of the entire set of features extracted, and the MFCC dataset consisting exclusively of MFCC statistical features. The latter half of the chapter consisted of primarily statistical analysis, where the results of the crowdsourcing experiment and the trained models were presented. An in-depth analysis of the annotations was conducted for both crowdsourcing experiments, where demographics, time taken by each participant, the distribution of preference and the affect correlations were thoroughly analysed. For the purposes of preliminary analysis a global ordering was derived from the preference annotations of the base sound comparison experiment, where the top and lower ranked sounds are showcased and discussed. A comprehensive statistical analysis of the different machine learned predictors follows, where the average validation accuracy is compared between different combinations of feature sets, feature selection algorithms and parametrizations. The chapter then concludes with an in-depth discussion of the various experiments and results obtained.

Chapter 7

User Evaluation

The construction of emotional playable experiences is one of the main focal points of the *Sonancia* system. Both audio and level architecture work in tandem for the realization of a pre-defined intended experience, which the system interprets and subsequently adapts to the gameplay constraints. To test the viability of the *Sonancia* system, several human participants were invited to play-through a level generated by the methodology presented within this thesis. This chapter will present both the process utilised for experimentation and the subsequent results obtained. The experimental protocol is fully defined in section 7.2, where a detailed overview of each step of the protocol is given. Through these experiments, several aspects of player emotion instigated through play and how these relate to the tension progression of the level will be investigated. Furthermore, we also explore the differences between felt emotion, obtained through physiology, and the perceived emotion, obtained through user annotation.

Section 7.1 describes both the collection process and the types of data collected for user experimentation. The three different types of data collected consists of: psychophysiology skin conductance signals, participant annotations of gameplay, and logs of game state information. The use of psychophysiology sensors has been often utilised in the digital game space. For example the work of Martínez et al. (2011) argues that psychophysiology signals can be used for the construction of in-depth player models, which can subsequently be used for the personalization of gameplay experiences. Additionally, the work of Yannakakis et al. (2010) also utilised psychophysiology for the detection of frustrating situations, that would arise during gamaplay. Psychophysiology has also been used for therapy, such as in the work of Holmgård et al. (2015), where a serious game was created in order to measure the anxiety of soldiers suffering post-traumatic stress disorder. For this thesis, psychophysiology was utilised to obtain a measure of player experience, to effectively validate several aspects of generated levels. More precisely, skin conductance signals obtained during play, was used to measure the player's actual emotional experience, and compare it with the tension progression of the level (i.e. the expected experience), and the player's gameplay annotations (i.e. the perceived experience). The methodology utilised for the collection of skin conductance signals is fully detailed in section 7.1.1 of this chapter. The participant's perception of their own emotion can often differ from that of psychophysiology, which can provide interesting insights on player behaviour. It can also dictate the intensity of emotion, as a strong emotional event has a higher probability of being remembered post-mortem. A wide variety of annotation methodologies exist, in order to ease this process for partici-

pants. Some of the most common methods utilised have consisted of rating each experience individually, through a Likert scale for example (Likert,1932); or by ranking between two diverging experiences (Yannakakis and Hallam, 2010); or even by describing their experience post-mortem using written words or questionnaires. This thesis intends to explore an alternative methodology of annotating player experience, inspired by the work of Clerico et al. (2016). This particular system allows participants to review a video of their gameplay, and annotate the experience directly in real-time using a "wheel-like" interface. Further details are offered in section 7.1.2, where a complete overview of the tool is given.

An important process of working with participant data is both the filtering of corrupted and unusable data, and the extraction of viable features capable of giving insight to the study at hand. Several processes were utilised for the extraction of statistical features of both skin conductance signals and real-time player annotations. Section 7.3 presents an overview of the data cleaning process and the feature extraction methods for analysis.

This chapter concludes with both section 7.5 and section 7.6, where an extensive discussion of the obtained results is presented, and a brief summary of the entire chapter is given, respectively.

7.1 Data Collection

To analyse player tendencies and the effects of *Sonancia* levels on the player experience, an efficient data collection process is an important step of user experimentation. Data obtained through human participants consist of controlled experiments, where they are exposed to distinct conditions for the elicitation of certain emotional states. Participants are then tasked of annotating the experience by self-reporting the perceived emotion felt or through physiological monitoring. Additionally, the game itself also keeps track of the player's actions and game state, through a continuous logging system. So in total three types of data are collected within this experiment: skin conductance signals (i.e the psychophysiology data), self-report annotations and experiment logging. The following sections will introduce each data collection methodology accordingly.

7.1.1 Collecting Physiology Signals

According to Andreassi (2013), "*Psychophysiology is the study of relations between psychological manipulations and resulting physiological responses, measured in the living organism, to promote understanding of the relation between mental and bodily processes.*". In the medical field physiology has been heavily used for the diagnosis or monitoring of involuntary human bodily functions, such as electrocardiograms (ECG) which are commonly used for heart-rate monitoring. Psychophysiology consists of using physiological devices by measuring the reactions of human bodily functions according to psychological stimulates. Common devices used in psychophysiology have consisted of ECGs, electroencephalography (EEG) and skin conductance (SC). Within this thesis we focused exclusively on monitoring SC, also referred to as Galvanic Skin Response (GSR) or Electrodermal Activity (EDA). Considering the context of this thesis within the horror genre, which target emotions such as stress and fear, SC has been consistently used as a good indicator for these specific emotions (Healey and Picard (2005); Hernandez et al. (2011); Holmgård et al. (2015)), additionally SC monitoring devices are particularly easy to setup and non-intrusive for participants. For this reason SC was used as ground truth for validating user experience during play.

Figure 7.1: The Empatica E4 wristband used for monitoring skin conductance.

SC is monitored through the Empatica E4 device (Garbarino et al. (2014)), which consists of a bracelet-like apparatus akin to a wristwatch (see Fig 7.1). SC consists of measuring the amount of sweat secreted by an individual's sweat glands through the conductance of small electrical pulses from the participants wrist. SC is measured in μS (micro Siemens) within a range of $[0.01, 100]\mu S$ at a sampling rate of 4Hz, where high μS values indicate high arousal (i.e. high conductance), while low μS values indicate low arousal (i.e. low conductance).

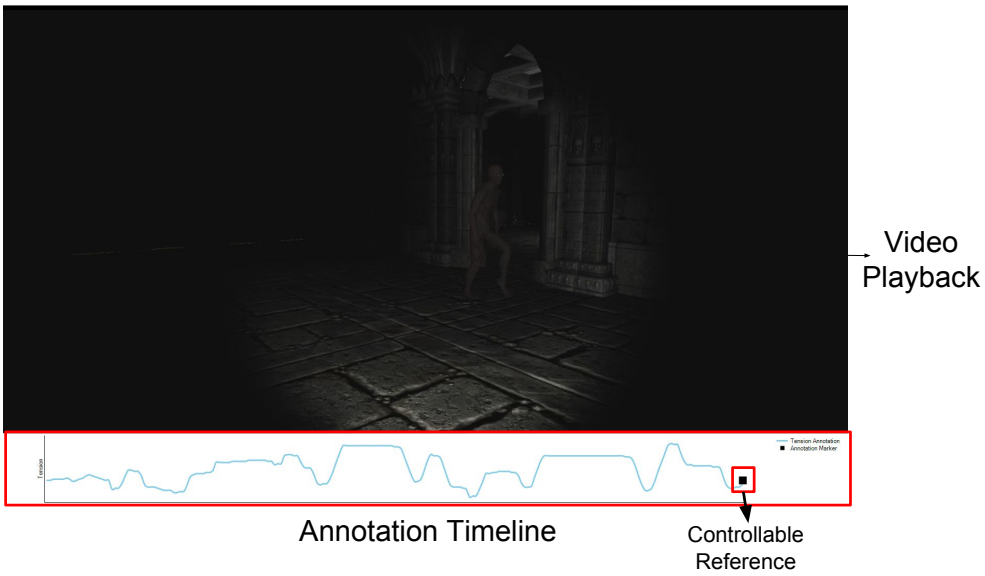
Communication between the Empatica E4 and the game is achieved through a bluetooth connection, allowing the E4 device to communicate with a local server without the need of cables or wiring. Once the E4 device is registered by the server, it stays connected until it is disconnected manually or the device's battery runs out. Through a client software program which was implemented specifically for this experiment, the game can directly communicate with the server in order to receive and send data from the connected Empatica E4. This allows the game logging system to control the flow of information received by the device, whilst also allowing each data point to be tagged with the game's local timestamp.

7.1.2 Annotating Gameplay

For the purposes of comparing the ground-truth with the participant's own emotional perception, a video annotation software was developed. First proposed by Clerico et al. (2016), this tool allows participants to watch the recorded playthrough of a previous playing session and annotate in real-time the perceived intensity of emotion (see Fig 7.2a). The entire annotation is controlled through a "wheel-like" controller (see Fig 7.2b), allowing participants to meticulously increase or decrease emotional intensity by turning the wheel, similarly to how volume is controlled on a stereo system. Unlike the work presented in Clerico et al. (2016), annotation is not limited to a particular range, allowing participants to continuously increase or decrease the amount of intensity. This was opted due to several situations arising, where participants wished to increase/decrease tension but could not do so due to the bounded limitation. With our approach participants may work with a broader range of emotional intensity. Once a playthrough is annotated the software automatically applies a min-max normalization on the entire annotation.

7.1.3 Player and Game State Logging

During play, the system will keep track of the player interaction, and other types of emergent behaviours. This information is logged through a synchronous method, where the current



(a) Real-Time Video Annotation Software. Participants annotate their emotional experience (controllable reference) by watching a recorded playthrough (top). The trail of their annotation can be seen below for the participants own reference.

(b) The Griffin PowerMate wheel interface used for video annotation.

Figure 7.2: Figure 7.2a is an image of the real-time annotation software, allowing participants to annotate their emotional experience using the PowerMate controller (fig. 7.2b) in real-time, while watching a video of their playthrough.

state of the game is logged every 5 in-game time frames; and through an asynchronous method, where events are logged the instant they occur. All game logs are accompanied by a timestamp and log-type identifier, allowing for the efficient post-processing of the different log types. Each log-type identifier can be accompanied by a diverging set of parametric values, which offer specific information of the different behaviours occurring during the game.

The following parameters are tracked by the synchronous logging system:

- Player Status Update: Logs the positional information of the player, including the current room, tile and (x, y) position. The player's current health and the number of times a player has died is also logged.
- Monster Status Update: Logs the positional information of each monsters within the level, including the current room, tile and (x, y) position. Each monster is identified through an ID number, which is also included in the log. Additionally, the current monster behaviour is also logged.
- Level Start / Level Complete: This is the only log-type that overwrites the synchronic-

ity by logging the exact timestamp of the level start and complete event. This ensures that data from each playthrough can be parsed, in the situation of partial data loss.

All of the asynchronous logs come accompanied with the object's positional information, similarly to the synchronous logs it includes the object's current room, tile and (x, y) coordinates. Followed are the parameters tracked by the asynchronous logging system:

- Player Input Log: Logs the specific key pressed by the player.
- Player Room Change Log: Logs the specific point in time when the player enters a new room. It is important to note that this log also represents the audio change, as audio assets are directly tied to each room, meaning that once a player changes room a new audio asset is being listened to.
- Player Tile Change Log: Logs the specific point in time when the player steps over a new tile.
- Player Damage Log: Logs the specific point in time the player receives damage. This log also keeps track of the monster identifier that damaged the player.
- Player Dying Log: Logs the specific point in time the player dies from a monster. This log also keeps track of the specific monster identifier that killed the player.
- Monster Patrol Update: Logs the specific point in time the monster x starts a new patrol route.
- Monster Field of View Update: Logs the specific point in time the player enters monster x's field of view.
- Monster Behaviour Update: Logs the specific point in time the monster x changes behaviour, out of four possible behaviours: patrolling, chasing, seeking and attacking.
- Monster Audio Behaviour Update: Logs the specific point in time the monster x emanates a patrol or chasing growl. This log also tracks if the sound is heard by the player (i.e. audio sphere of influence collides with the player).
- Player Start: Logs the specific point in time the player initiates play.
- Player End: Logs the specific point in time that play stops. It also keeps track of how play terminated, i.e. time limit reached, or completing the level through the objective.

To ensure that all logs are written efficiently, and that the main game does not suffer frame rate slowdowns, each logging system works on its own separate thread. The game communicates with each thread through a queuing data structure, where logging events are allocated as they appear in-game. Each logging thread will continuously write to their respective file as events appear, only terminating once the game has ended and all events in the queue have been written.

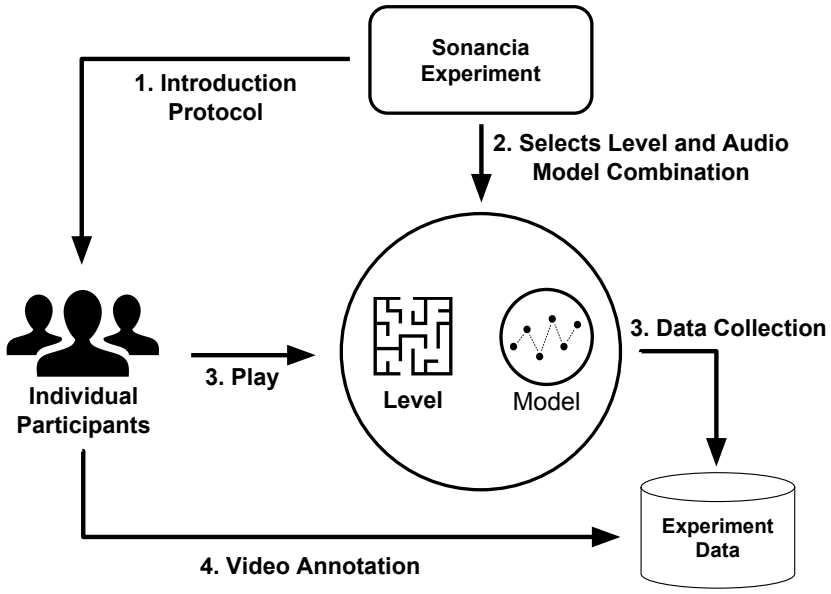


Figure 7.3: Overview of the experimental protocol used for the validation of the *Sonancia* system. The protocol is divided in 4 different phases: 1. The Introduction Protocol; 2. The selection and the ordering of three different audio model and level combinations to be played by the participant; 3. The participant playthrough and data collection phase for each level + model combination; 4. The participant video annotation phase for each level + model combination.

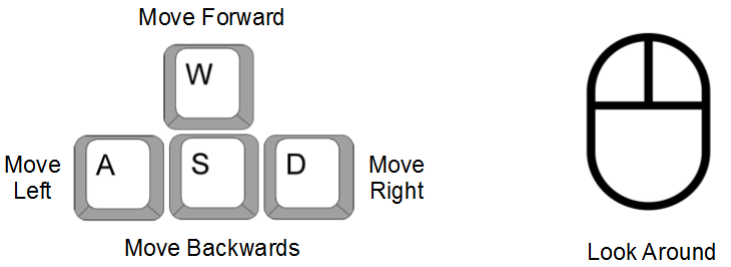
7.2 Experimental Methodology

An experimental protocol was defined apriori to ensure an efficient and consistent process for each participant partaking in the experiment. Figure 7.3 shows an overview of the various phases of the protocol, which will be described in detail within this section. The first step consists of an introduction protocol, where the experimental procedure is presented to each participant and the data collection devices are synchronized with the game (see section 7.2.1). Once all devices are connected, the level ordering is chosen. Each level consists of a variation of the same level utilising a different audio model (see section 7.2.2). Each participant plays through a level, while the system is logging their data (see section 7.2.3). Once a level has finished, the participant annotates a recording of their own gamplay of that level (see section 7.2.4). The latter two processes are repeated for the remaining level variations.

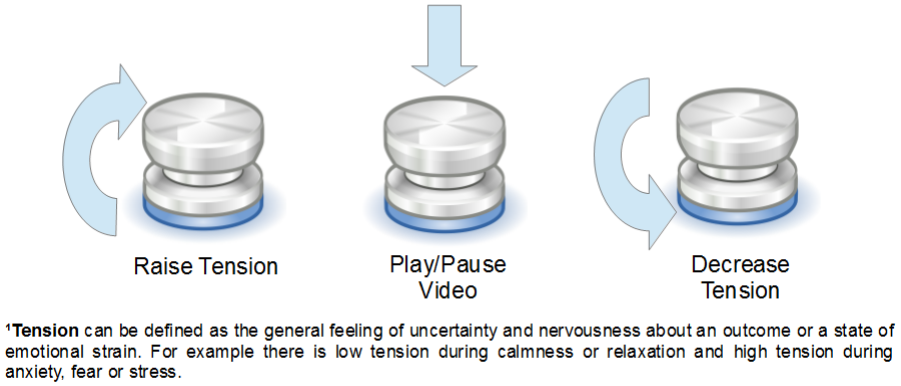
7.2.1 Introduction Protocol

The introduction protocol is a set of actions, allowing participants to both get acquainted with the systems utilised during the experiment, and to set up the data collection processes (i.e. skin conductance hardware). It was important to get this process as streamlined as possible, in order to be efficient so as not to waste the participants time, and also guarantee that data collection was functioning at the start of each experiment.

The first step of the introduction sequence consists of signing a consent document, allowing for the ethical usage of the participant data in future publications and studies.



(a) *Sonancia* controls schematic presented to participants.



(b) Video Annotation Tool controls schematic presented to participants.

Figure 7.4: Control schematics presented to each participant as part of the introduction protocol.

Once consent is given, participants are then asked to fill a demographics survey with the following questions:

- What is your gender?
- Which age group do you belong in?
- How do you feel about the horror genre?
- Are you a musician or sound producer?
- How often do you play games?
- Does the prospect of playing a horror game frighten you?
- Do you have any hearing problems?
- Have you drunk coffee, energy drinks or any substance with caffeine today?
- What hand is your dominant?

Following the demographics survey, each participant is given a briefing about both the objective and controls of the game. Once the participant understands the basics of the

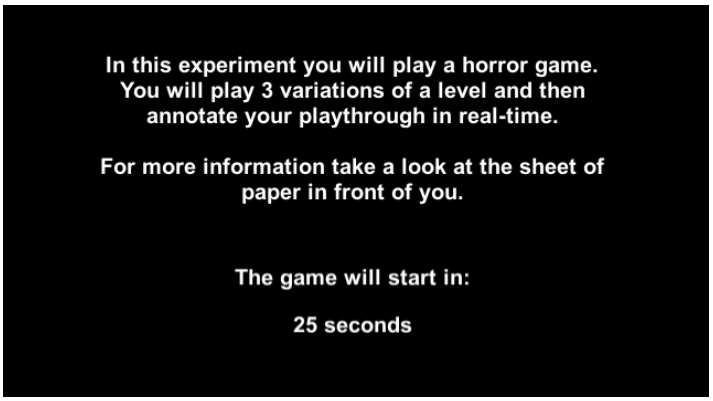


Figure 7.5: Initial experiment screen with a brief description of the experiment. A baseline sequence of the skin conductance is also captured during the entirety of this screen. The sequence lasts for a total of 30 seconds.

game, the video annotation software is presented. Participants are encouraged to test the controls of the annotation tool, to get used to the wheel controller and the procedure for loading the gameplay recordings. To keep participant interaction to a minimum during the actual experimentation phase, cheat sheets are provided detailing the controls of both the game and annotation tool (see figure 7.4).

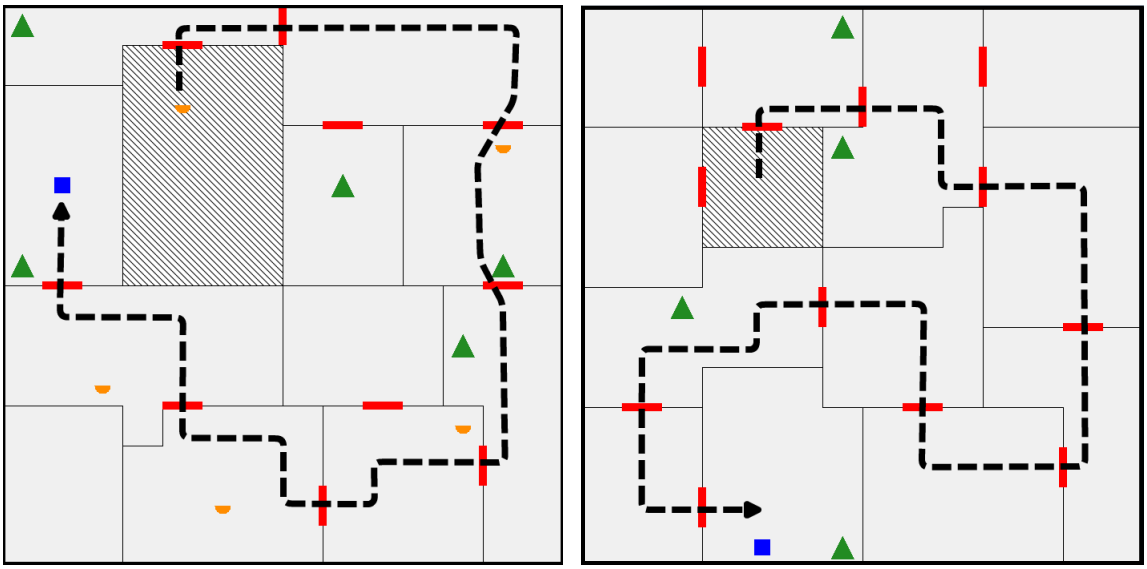
Once participants are comfortable with the experimental process, they are asked to place the Empatica E4 device onto their wrists. To ensure that data is being collected, a temporary startup screen was created in order to visualize the specific data points being sent to the game in real-time. It is important to note that due to the limited resources available, participants were not able to conduct the experiment in an ideal scenario, i.e. a noise cancelling sound booth. Although, to combat this limitation, participants were provided with high quality noise cancelling headphones, specifically the "Audio-Technica ATH-50x", which allowed to filter the majority of unwanted surrounding noises.

Once a participant places the headphones, they are left undisturbed for the remaining duration of the experiment, with a few minor exceptions: when a software complication arises; or a question needs to be asked by the participant. Figure 7.5 shows the first screen presented, which provides both a brief description of the experiment and registers a baseline value of the participants skin conductance. When the timer expires the first level of the sequence is loaded into play. It is important to note that both the audio and gameplay will only start after the level is fully loaded and the participant confirms by pressing the *start* button appearing on screen.

7.2.2 Level Sequence Selection

A total of two diverging levels were pre-generated by *Sonancia* specifically for experimentation, which could then be loaded into play during the experiments. Figure 7.6 showcases both of the levels utilised for the user evaluation experimentation. The first level (see figure 7.6a) utilises both light sources and monsters to influence the tension progression, while the second level (see figure 7.6b) focused solely on monster distribution.

Each participant was tasked of playing a sequence of three variations of a particular level. Each variation consists of a different audio selection method, for the application of sonification. The methods consist of:



(a) Generated *Sonancia* level utilising light sources, referred to as Level 1; $f = 1.0$.

(b) Generated *Sonancia* level with monsters only, referred to as Level 2; $f = 1.3$.

Figure 7.6: The two diverging pre-generated *Sonancia* levels used for user experimentation. Green triangles are monsters, yellow half circles are light sources, while the black arrow is the level progression.

- The Predictive Ranking: Audio assets are ranked by a machine learned model, which was obtained through the experiments realized in chapter 6. Ranking was done by the most accurate fold obtained for tension.
- The Random Ranking: Audio assets are ranked through a random gaussian distribution.
- No Sound (Diegetic Only): No audio asset is assigned. Audio played consists of in-game sounds only, such as monster growls, monster footsteps or the player's own footsteps.

Each level and its variations were evenly distributed before participant experimentation, this was done to ensure that each variation sequence and level were played by different participants. The system also ensured that each sequence assigned to the participants was loaded accordingly.

7.2.3 Participant Playthrough

As previously stated in chapter 4 the objective of the game is to reach the objective room and interact with the item placed within. Visually this item is represented as a statue, lying in the centre of the objective room. Once the player interacts with this statue, by pressing the pre-defined key, the level concludes. Along the path towards the objective players must avoid monsters that patrol the various rooms of the level. Monsters will never patrol outside of their respective rooms, but will chase players in adjacent rooms if they are caught within their line of sight. Monsters present four different behaviour types (see figure 7.7): patrolling, chasing, seeking and attacking. Patrolling is the standard monster behaviour,

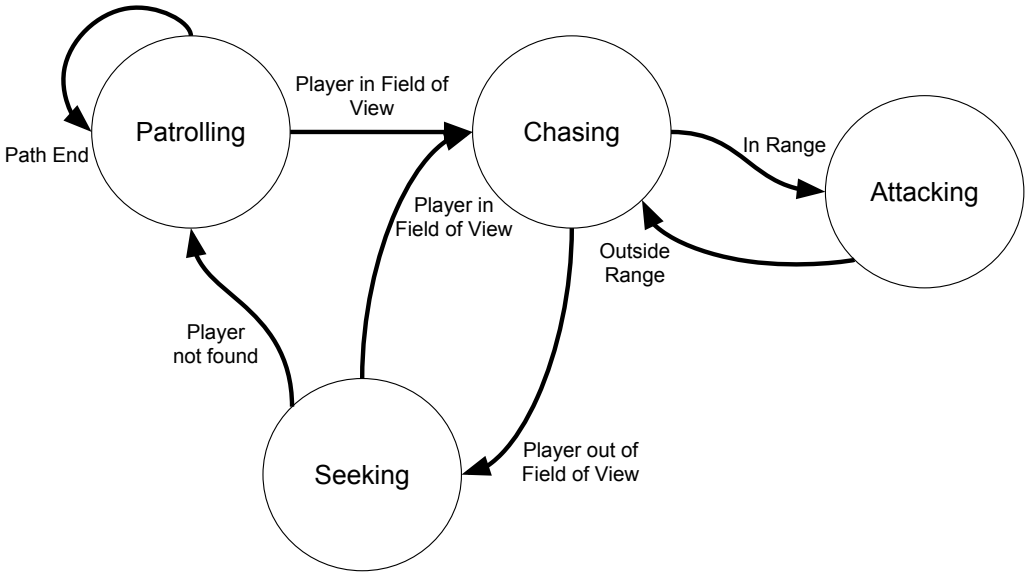


Figure 7.7: The Monster Behaviour Finite State Machine.

walking specific routes within their respective room. The chasing behaviour initiates and remains active if the player is within the monsters field of view. This behaviour triggers the run animation and creates a path between the monster and the player. If the player leaves the monster's field of view, the seeking behaviour is engaged. This behaviour prompts the monster to idle on the point of the last known player location (i.e. when the player left the field of view). If no player is found, i.e. 4 seconds without the player crossing his field of view during this behaviour, the monster re-engages the patrolling behaviour and returns to his respective room. If the monster gets within attacking range of the player, it will engage the attacking behaviour damaging the player. If the player is hit more than 5 times, the player dies and re-locates to the level starting position. Monsters also emanate different sound types, such as footsteps and growls. These sounds are spatial, and depend on both the monster and player positioning within the level. The intensity of positional sounds increases if the distance between the two objects is short, and it will slowly fade out as the distance becomes larger. Monsters can produce three different types of sounds: the patrolling growl, which emanates at a randomly selected intervals during the patrolling behaviour; the chasing growl triggers when the player traverses the monsters line of sight when in the patrolling behaviour; footstep sounds emanate during monster movement.

The instant play starts the gameplay recording and data logging process commences, stopping only when the player reaches the objective or the total time limit is reached. For the purposes of this experiment each level variant included a time limit of 4 minutes. Once the limit is reached the participant is informed of the fact and guided towards the video annotation process.

7.2.4 Video Annotation

At the end of each level variation, the game prompts the participant to switch to the video annotation software. The annotation tool will ask the participant to load the most

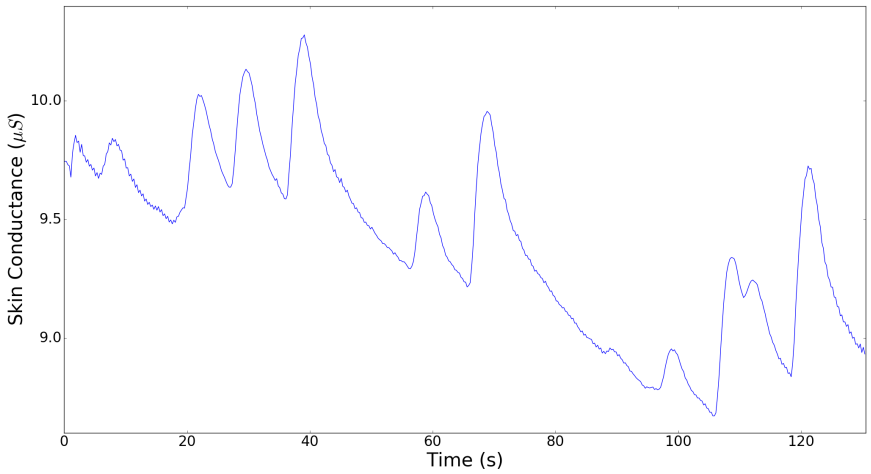


Figure 7.8: Example of a Skin Conductance signal obtained through the Empatica device. The y-axis consists of the value of conductance measured in μS and the x-axis represents the time in seconds.

recent playthrough recording to begin annotation. Annotation will only start once the participant presses inwards on the wheel controller interface. Once pressed, the participant's last playthrough will begin playing along with the annotation timeline. Similarly to the game's logging system, the annotation logs are written on a separate thread so as to not interfere with video playback. Additionally, it is important to note that annotation logs are only kept if the video is playing, if the participant pauses the video for any reason, the logging system will also halt. Once video playback finishes the software will indicate the participant to return to the game in order to play the remaining level variants.

7.3 Feature Extraction

Given the different types and the vast amount of data collected from participants, the ability to condense data into statistical features is possibly one of the most important steps of statistical analysis. This section will describe the different features that were extracted from skin conductance signals and the participant annotations.

7.3.1 Skin Conductance

Figure 7.8 shows an example of a raw skin conductance signal obtained from the Empatica device. SC signals are traditionally characterized by two different types of activity: *tonic* and *phasic*. Tonic skin conductance refers to the phenomenom of slow changing variation of the signal through time, considered to be the level of skin conductance in the absence of external events or stimuli. Phasic skin activity is the abrupt increase of conductance levels occurring within short-term event intervals. These typically occur after the presence of an environmental event or stimuli. Similarly to audio feature extraction, for the purposes of statistical analysis it is advantageous to apply some form of pre-processing (i.e. smoothing) and statistical feature extraction methods. Continuous Decomposition Analysis (CDA) proposed by Benedek and Kaernbach (2010), can decompose SC signals into continuous tonic and phasic activity (see figure 7.9). By sampling the signal at defined intervals an estimation

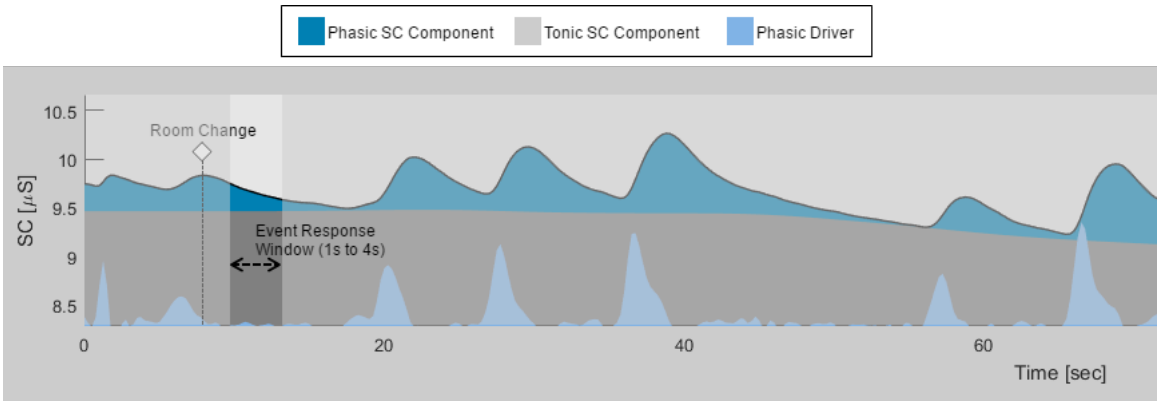
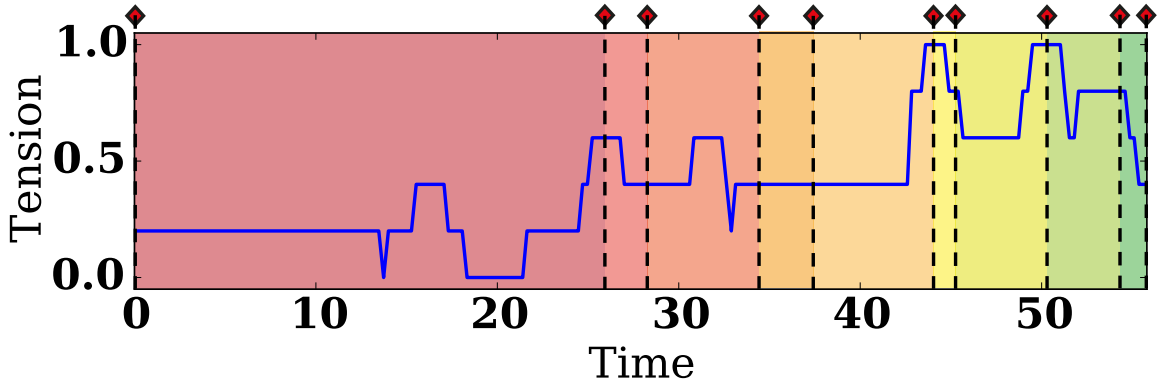


Figure 7.9: Continuous Decomposition Analysis (CDA) of a sub-section of figure 7.8. Three components are extracted from the raw signal data: Phasic Activity (Dark Blue), Tonic Activity (Grey) and Phasic Driver (Light Blue). These features are extracted within the event window response.

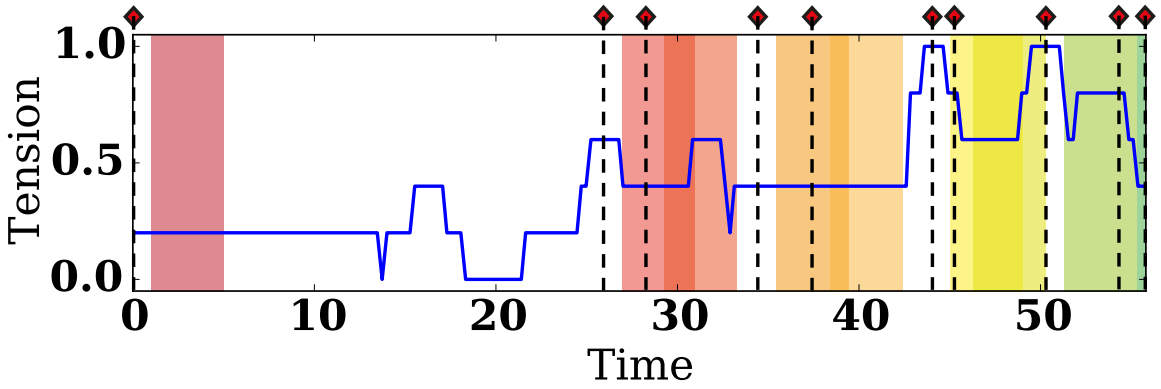
of the tonic activity can be derived, presuming the SC signal is stable. Phasic activity can then be extracted by simply subtracting the tonic activity, resulting in what is called a *phasic driver* expressed in μS. As such the phasic driver consists of a baseline corrected measure, capable of capturing the affect of a given stimulus. The stimulus-response window for SC typically ranges around intervals of $[1, 3]$ or $[1, 5]$ seconds after a stimulus event (Benedek and Kaernbach, 2010). For the purposes of this thesis, SC statistical features are extracted within a defined response window of $[1, 4]$ seconds, after the occurrence of a stimulus event. Additionally a minimum phasic threshold of $0.01\mu S$ is set, meaning that only events whose phasic values are above this threshold are taken into consideration. The Matlab tool *Ledalab* was used for all SC signal processing, which includes all features previously described with the addition of SC signal preprocessing algorithms. According to Holmgård et al. (2015); Bach and Friston (2013); Benedek and Kaernbach (2010); Boucsein (2012) the following statistical features have been considered appropriate for the detection of elicited stress:

- The phasic driver mean within the response window ($\bar{P}_d$);
- The phasic driver integral within the response window ($\int P_d$);
- The tonic mean within the response window ($\bar{T}$);
- The global mean within the response window ($\bar{G}$).

To reduce the noise of the raw SC signals, a gaussian smoothing function is applied on each SC signal before the application of CDA. It is also important to note that only valid SC signals, which presented a stable continuous signal, were taken into consideration. However, a repairing function was applied on partially stable signals, where it specifically presented a stable signal from the beginning or at the end of a gameplay segment. For these situations the noisy section of the signal was cut, and the time-frame of the stable signal was used exclusively for analysis.



(a) Example of the Continuous Windowing Method for extracting features of participant annotations. Windows consist of timeframes in-between two diverging events.



(b) Example of the Reactive Windowing Method for extracting features of participant annotations. Windows consist of extracting singnals 1 second after the event and ends after 5 seconds.

Figure 7.10: Example of a gameplay annotation (blue-trace) split by the two windowing methods utilised within this thesis. The *Continuous Event Window* (see Figure 7.10a) extracts a partial signal between two events, while the *Reactive Event Window* (see Figure 7.10b) extracts a partial signal 1 second after the event occurs for a period of 5 seconds. Dotted lines mark the exact time of an event, while the coloured areas define the exact window extracted after each event.

7.3.2 Gameplay Annotation

Once each level variant was complete, participants were tasked in annotating data using the methodology described in section 7.1.2. This annotation consists of a continuous representation of participant perceived tension, during a segment of game-play. For the sake of simplicity, this annotation can be considered a signal, where x is time and y is perceived tension. It is also worth reminding that annotations are normalized between $[0, 1]$, where 0 is considered a low tension state, and 1 a high tension state.

Annotation feature extraction commences by parsing the signal into several framing windows, which from these statistical features are extracted. A frame consists of a partial subset of the entire signal, representing the perceived emotion elicited through a gameplay event. Within the context of this thesis, an example of a gameplay event may be the point in time the player enters a new room. Figure 7.10, showcases the two window framing methods utilised for the extraction of features in annotation signals. The first method is referred to as

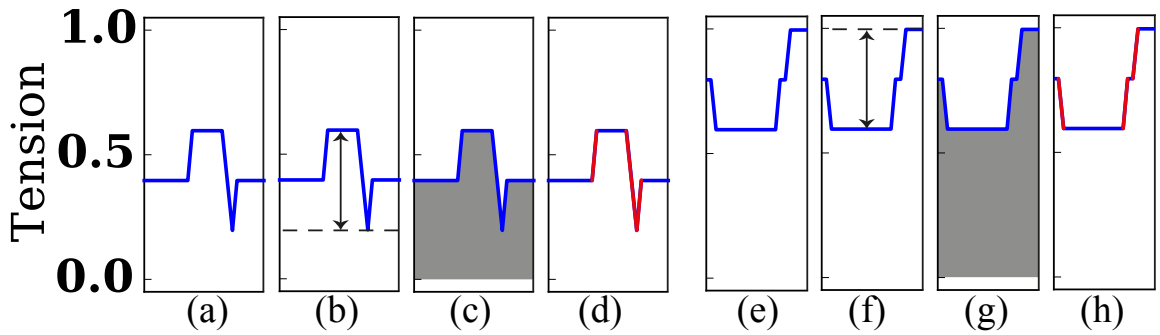


Figure 7.11: Two indicative time windows from Figure 7.10a being processed by feature extraction. The $\bar{W}$ is 0.45 in 3a and 0.7 in 3e, as the average value of the approximately 30 data points in these windows. Calculation of $\hat{W}$ is shown in 3b and 3f based on the amplitude of the partial signal in that window (0.4 in both cases). The integral is calculated based on trapezoidal integral of the area under the trace in Figure 3c and 3g. The average gradient calculates the difference of adjacent data points, which is non-zero for the red parts of Figure 3d and 3h; note that ΔW is 0 for 3d as there are equal positive and negative gradients which cancel each other out.

a *Continuous Event Window* (W_c), where signal parsing starts immediately after an event occurs and continues until another event re-occurs. More specifically, the entire window consists of the section between two types of gameplay events. The second methodology is referred to as a *Reactive Event Window* (W_r), which is inspired by the windowing function from the feature extraction method of skin conductance. With this methodology signal parsing occurs 1 second after an event occurs and continues for an additional 5 seconds. This methodology hypothesizes that the majority of participant annotations are reactive to occurring events. Additionally, it takes into consideration that the annotation might have a delay due to the annotators reaction, similarly to how skin conductance reacts to stimuli.

Once the signal is parsed through either framing technique, statistical features are extracted from each available window (see Figure 7.11). For the purposes of experimentation several statistics are derived including:

- The total mean of the window ($\bar{W}$);
- The peak-to-peak difference of the window ($\hat{W}$)
- The mean of the sum of gradients of a window (ΔW)
- The composite trapezoidal integral ($\int W$)

As the name suggests the $\bar{W}$ consists of calculating the mean of all tension values within the window frame, while $\hat{W}$ consists of calculating the difference between minimum and maximum tension values within the window. ΔW is obtained by calculating the mean of the sum of gradients, where T_t is the actual recording time of t, and n is the total number of points within that window, such that:

$$\Delta W = \sum_{t=1}^{n} \frac{x_{t-1} - x_t}{T_{t-1} - T_t} \tag{7.1}$$

Table 7.1: The mean, maximum and minimum time (in seconds) of the participant playthroughs, according to: the total levels played; the first level played; the second level played; the third level played; level of figure 7.6a, ignoring ordering; level of figure 7.6b, ignoring ordering.

	Mean	Max	Min
Total	100.34 (s)	258.143 (s)	31.586 (s)
First Level of Order	122.87 (s)	258.143 (s)	39.202 (s)
Second Level of Order	87.272 (s)	185 (s)	37.623 (s)
Third Level of Order	89.34 (s)	192.454 (s)	31.586 (s)
Total Level 1	84.23 (s)	185.204 (s)	31.586 (s)
Total Level 2	117.408 (s)	258.143 (s)	37.623 (s)

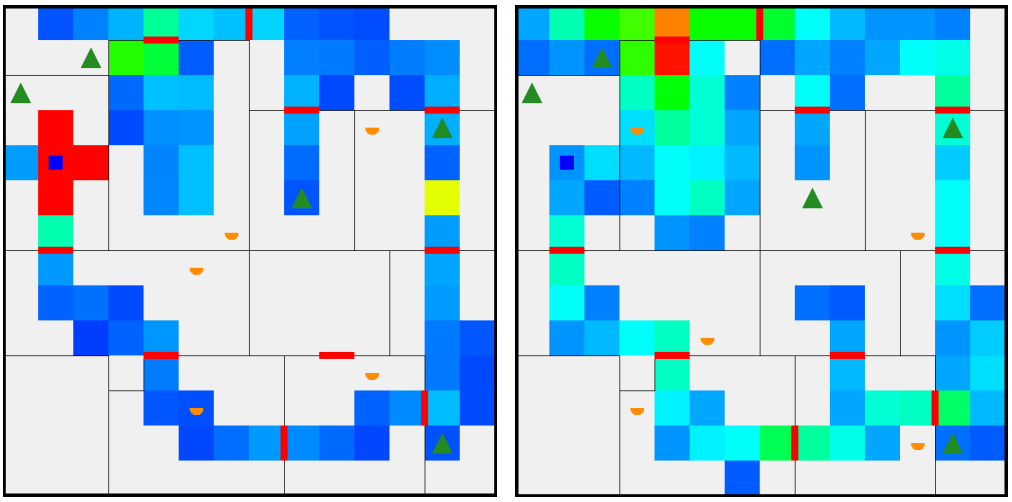
Finally, $\int W$ consists of the integral calculation of the time window using the composite trapezoidal rule.

7.4 Results

The majority of participants were recruited from the university campus, where both students and lecturers took part in the experiment. A minor subset of participants from outside the university also participated, thanks to several advertisements sent through social media applications such as Facebook and Twitter. In total 41 participants were recruited for the experiment, although several complications arose with the Empatica E4 device, leading to several incomplete or unusable experiments. Out of the 41 participants, 28 individuals were male, 12 female and 1 preferred not to specify. The most represented age group with a total of 24 participants were between the ages of 25 to 34, with the second largest being between the ages of 18 to 24 with 14 participants in total, while 3 participants stated that they were between 45 to 54 years of age. Additionally, out of all the 41 participants, no one stated that horror was their favourite genre, however 13 participants did state that they do enjoy it, while the remaining 11 and 17 stated that they did not enjoy the genre or were indifferent to it, respectively. Interestingly, participants were slightly divisive about their feelings on playing a horror game, where 17 individuals stated that the prospect of playing a horror game frightened them, while the remaining participants stated that it did not scare them, or were indifferent. The majority of participants were passionate about games, with only 5 individuals in total stating that they rarely played any type of game, while 28 individuals stated that they played games quite frequently and 8 stated that they casually played games.

7.4.1 Level Playthrough Analysis

Table 7.4.1 presents a statistical analysis of level play duration. The average time played for all levels was around 100.26 seconds ($\simeq$ 1.67 minutes), suggesting that the majority of participants were able to reach the objective. The maximum time obtained out of all level playthrough's was 258.143 seconds ($\simeq$ 4.3 minutes). The 30 second delay after the level timer expired, consists of the time the participant took to confirm that play had terminated. It is important to note that these extra length sections are cut from both skin



(a) Heatmap of all runs of level 1.

(b) Heatmap of the first run of level 1.

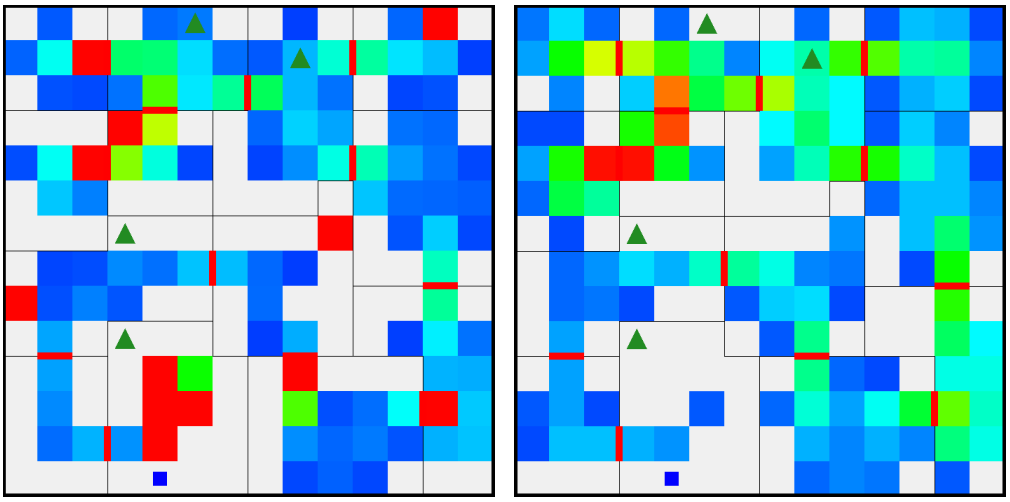
(c) Heat Gradient

Figure 7.12: A heatmap of the most visited tiles of Level 1 (see figure 7.6a). Figure 7.12 consists of a heatmap of the total number of runs of level 1. Figure 7.12b consists of a heatmap of the first run of level 1, only.

conductance and video annotation, where only gameplay data is kept. As expected, due to the memorization of the level layout, participants tended to be faster on the second and third runs of the level, with a few minor exceptions. Interestingly, the second level took longer to complete in comparison to level 1. Even though both levels contain the same number of monsters within the progression, level 2 presented more alternative paths, which often confused players even on subsequent playthroughs, leading to a longer playing time.

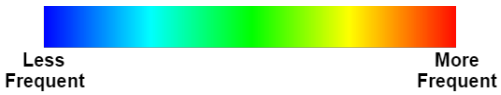
Figure 7.12 showcases two diverging heatmaps of the most visited tiles of Level 1, where figure 7.12a consists of the entire set of runs for each participant, while figure 7.12b consists of the first run of each participant, only. As expected during the first run, participants tended to get lost more frequently, which subsequently allowed them to explore the initial layout of the level more. Once participants started to memorize the level a less exploratory pattern emerges, where the majority of participants tended to go straight towards the objective. This also confirms our previous suspicions, where participants tended to spend the most amount of time within the first run of a level. Interestingly, the first two rooms near the door commonly shows a concentration of activity, this is due to players hiding from the monster patrolling within the adjacent room, which often walked along the narrow corridor leading to the door. To avoid the monster, players often took advantage of the walls in the starting room in order to hide from it, hence it was common pattern among the majority of players to go back and forth within that general area.

Figure 7.13 similarly shows the most visited tiles of all 3 runs (see figure 7.13a), and the first run (see figure 7.13b) of level 2. Alike the previous analysis, the participant tendency on the first run is still more exploratory then subsequent runs, although for this particular level it can be seen that participants had a much higher difficulty reaching the



(a) Heatmap of all runs of level 2.

(b) Heatmap of the first run of level 2.



(c) Heat Gradient

Figure 7.13: A heatmap of the most visited tiles of Level 2 (see figure 7.6b). Figure 7.13 consists of a heatmap of the total number of runs of level 2. Figure 7.12b consists of a heatmap of the first run of level 2, only.

objective in comparison to level 1. We hypothesize that this is due to this particular level presenting more maze-like properties than the previous level. For example the starting room immediately branches out, while the second room and third room of the progression does so as well, while in the previous level room branching was much less consistent and presented more linearity in comparison. This also stays consistent with our previous assumption, as this particular level presented higher completion times, and was also the only level where participants reached the maximum time limit. Additionally, due to the probability of getting lost being higher, the chances of being chased by a monster also increased, which also explains why more space was explored through all three runs. Particularly, walls and doors were used substantially more, than the previous level, to hide from enemy sights, even from adjacent rooms as monsters could chase players from another room if caught within their line of sight.

The importance of maze-like properties, and room branching were verified in this particular analysis. Even though level memorization was still present in the subsequent runs of level 2, participants did have a higher difficulty in beelining towards the objective, contrasting the results obtained from level 1. For future reference, a new methodology might be worth investigating, where the emphasis would be specifically on the creation of branching levels, instead of the main progression from start-to-finish. A methodology that could utilise the same methods expressed within this thesis, such as the incorporation of designer intent would need to be rethought, however. On the other hand, this analysis also suggests that emphasizing exclusively on one specific progression for an entire level, might not be the best solution for the creation of engaging levels for the players themselves. Even though the level generation method presented in this thesis takes into consideration the designer's intended

experience, it does so at a cost, where resulting levels are often restricted to mostly linear progressions with small branching paths in-between. One solution could allow designers to specify more than just a singular path, but multiple paths, instead. Although PCG algorithms can offer gameplay replayibility through the consistent construction of new levels, it is interesting to think about PCG algorithms with the capability of optimizing towards a singular level's replayibility, instead of level divergence. Even though this was not the objective of this thesis, it still poses an interesting question for future work.

7.4.2 Rank Correlations

In order to investigate the relationship between perceived tension, skin conductance and the pre-defined values within the level progression, the binomially-distributed pairwise correlation methodology defined by Yannakakis and Hallam (2011), and previously described in section 6.4 of this thesis, was used for the creation of a global ranking of diverging data types. To study if there is a correlated tendency between two diverging signals, features are extracted at the exact same instances using the previously described windowing methods (see Section 7.3). Each window is subsequently compared with the remaining windows from the same signal such that: z is $+1$ if both signals are concordant, or -1 if both signals are discordant. Concordance consists of analysing the divergences between two windows (i.e. is the signal growing or decreasing), and comparing with the window divergence of the opposing signal. If the divergences agree between both signals then it is concordant, if there is disagreement, then it is discordant.

For the purposes of this study, noisy and incomplete experiment runs were discarded. Only complete runs with stable skin conductance signals, gameplay annotations and logs were considered. Additionally, in order to create feasible event windows for signal feature extraction of both skin conductance and video annotations, a filtering technique is utilised. If the difference between two window pairs is below a threshold parameter, that comparison is discarded for both signals. Furthermore, if the same exact event occurred repeatedly within 2 seconds after the first instance of that event, the repeated occurrences are removed. An example of reoccurring events is the situation when players repeatedly enter and leave the same room within a small time frame. Unfortunately about half of the experiments were discarded due to noisy and unreliable skin conductance signals, as the device was unable to effectively measure the conductance of participants skin. The cause of this was due to the Empatica's inability to capture conductance measures of participants with particularly dry skin. Thus, the final total of reliable runs is 40, which was utilised in the following experiments.

In order to thoroughly study the relationship between two diverging signals, different pairwise comparison experiments were conducted. The first experiment assumes that there is a global signal relationship, and thus compares each possible window pairing of the entire signal. The second experiment assumes that the relationship of a signal is local, meaning that pairs are only compared in sequential order. $T-1$ compares between two directly adjacent pairs, while the $T-2$ analysis compares between three directly adjacent pairs.

Skin Conductance and Gameplay Annotations

This section will investigate the relationship between the skin conductance signals and gameplay annotations. The statistical features extracted from both skin conductance and video annotation signals, utilised the methodologies defined in sections 7.3.1 and 7.3.2,

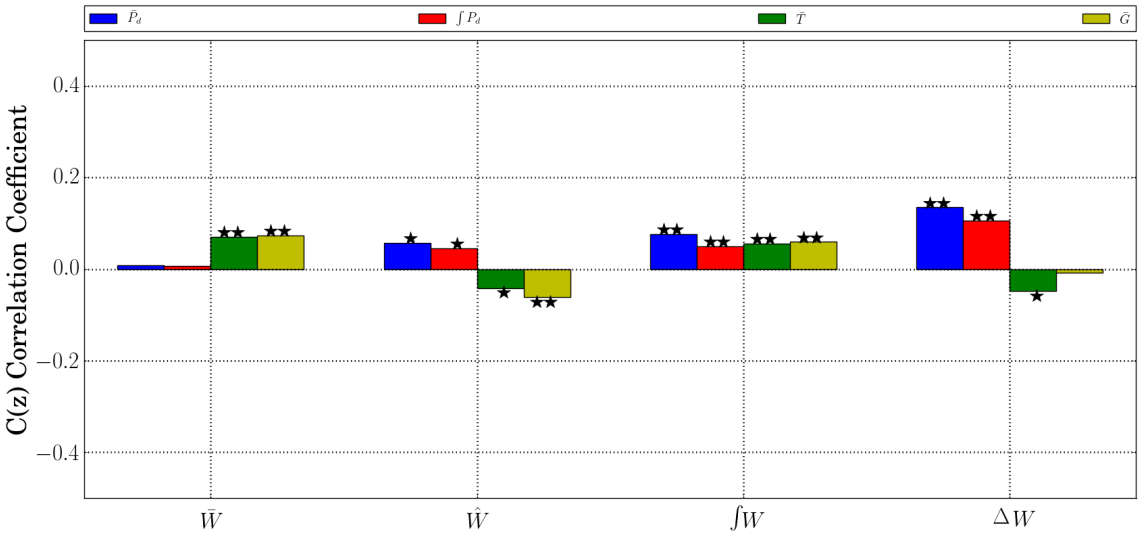


Figure 7.14: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **continuous** window type and annotator memory (all windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

respectively. The first case study consists of splitting the different signals based on a change room event. These events represent the exact time a player has changed room, and the current soundscape that is playing in the background. This study will investigate the agreement between ground-truth and perceived tension, within the specific situation of soundscape change. However, it is also important to note that, because it is an interactive experience other factors, e.g. monsters, can arise during play, which can also contribute to a player's overall affective state.

The following experiments also utilised a threshold filtering method, where the difference of each pairs is calculated in order to filter ambiguous comparisons, i.e. their difference is below a certain α are discarded.

The threshold α for all video annotation features in addition to the $\bar{P}_d$ and $\int P_d$ skin conductance features is set to 0.02. The Tonic and Global skin conductance features utilised an α of 0.03, as these signals presented much more stability and thus a more aggressive threshold was necessary to compare pairs with more divergence.

Figure 7.14 showcases the binomially-distributed correlations of global pairings between the all video annotations and skin conductance features, over 40 runs of both level 1 and 2, using the continuous windowing method. The majority of values obtain high significance due simply to the amount of comparisons obtained when comparing each signal globally. The average total number of pair-wise comparisons is ≈ 2741 between each feature correlation analysis. Interestingly, when comparing the signal globally a slight correlation exists between the mean of annotation features, and both tonic and global skin conductance features. This suggests that relatively to the entire signal there is a tendency for the average participant annotation value to follow the tonic and global averages. However, given the small correlation value we also suggest that this tendency might be due to some degree of positive bias within the tonic and global features, which was a tendency often observed in the experimental runs. In fact it is quite common for these features to present a continuous rising signal, where certain participant's presented $\bar{T}$ and $\bar{G}$ values that were persistently

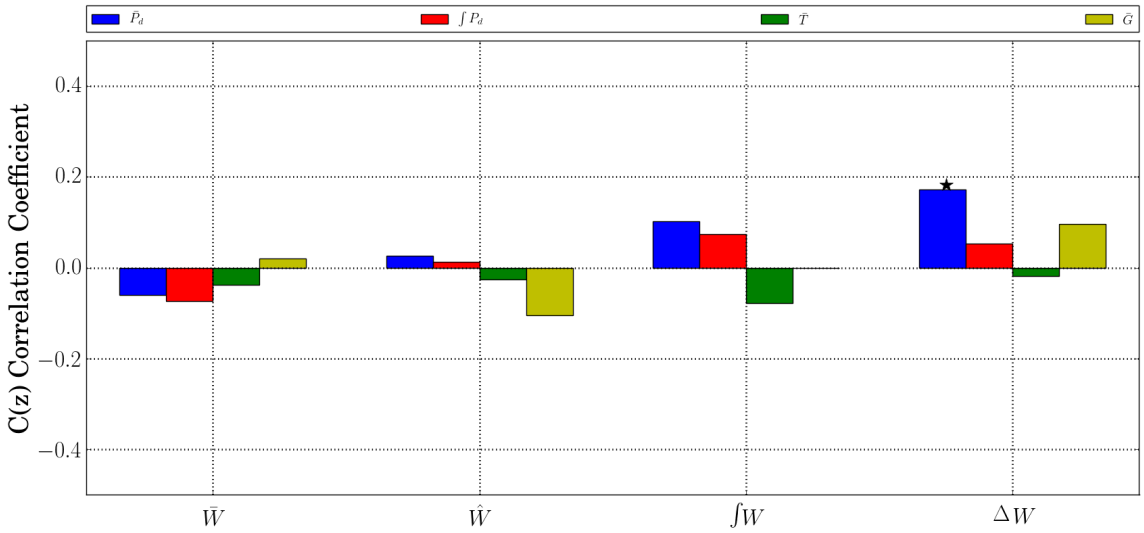


Figure 7.15: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **continuous** window type and annotator memory ($T-1$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

rising throughout the experiment. Thus when comparing between annotators who favour rising tension throughout the annotation phase will cause a high degree of correlation to take place, which can potentially swing the correlation towards positive values. Contrarily both $\hat{W}$ and $\Delta 1$ presented significant negative correlations with similar features, however the correlation coefficients are quite minor (< 0.1), which suggests that there is no particularly high tendency either towards positive or negative correlation. The slight tendency observed is due to the sensitivity these features have towards fluctuating windows, where a sudden decrease or rise can heavily impact both features as one deals with the min-max value of a window, while the other on slope tendency. Thus when comparing between an entire signal, particularly with windows that can have varying sizes, can cause these features to be erratic depending on the annotator. The highest significant positive correlation was between the $\Delta 1$ annotation feature and the $\bar{P}_d$ and the $\int P_d$. Although the correlation is below 0.2 the high significance does suggest that some degree of agreement exists between the obtained annotations and the phasic driver. Thus, we can hypothesize that an annotator can potentially perceive and then annotate the stimulus-response-pattern of an event, through a relative annotation value of rising or lowering tension instead of computing the actual annotation value.

Expectedly when comparing the immediate adjacent windows a lower number of comparisons is collected comparatively to a global comparison. Figure 7.15 showcases the obtained correlation coefficients for directly adjacent continuous windows, where the average number of total comparisons collected between each feature correlation was ≈ 249. Unlike the previous analysis several of the positive correlations observed are not present when analysing annotations locally. This is particularly true for both $\bar{T}$ and $\bar{G}$ GSR features, which further suggests that our previous assumption was correct. It is also important to remind that the tonic feature tends to be independent to stimuli-response-patterns, often presenting a slower and more consistent signal along time. Thus, it is sensible to assume that when comparing the annotation and the tonic globally there is a larger probability that both signals align,

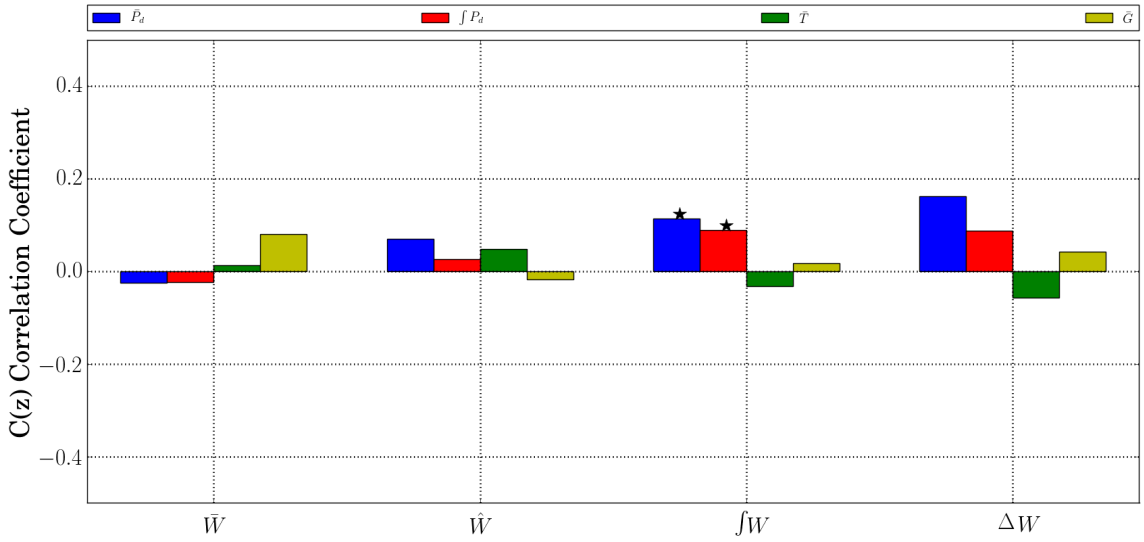


Figure 7.16: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **continuous** window type and annotator memory ($T-2$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

which is less probable to happen when comparing each window sequentially. In fact, the reactive nature of participant annotation is reinforced by the significant correlation obtained between ΔW and $\bar{P}_d$, which increased relatively to the global comparisons. These results suggest two diverging annotation tendencies. The first is an annotators tendency of locality, where results suggest that the annotations are not relative to the entire annotation cycle, instead the ground-truth is closer to the direct previous annotation value. Thus, reinforcing our second hypothesis, which consists that the actual annotation value is not particularly important to the annotator, what matters is the actual direction of annotation, i.e. is it rising or falling. Suggesting that this is the reason why indirect features such as $\int W$ and ΔW performed better, than other annotation features when comparing the windows locally. Furthermore, the increase in correlation for ΔW and $\bar{P}_d$ also confirms our intial assumption within the global analysis, that some participants are able to effectively annotate the stimulus-response-pattern, where ΔW does tend to predict the ground truth closer than the other features.

By analysing Figure 7.16, which showcases the pairwise comparison of the $T-2$ windows, it can be observed that a similar tendency occurs to that of the previous $T-1$ correlation analysis. The total average of pairwise comparisons for each correlation in the $T-2$ correlation analysis is ≈ 374. By applying an additional window each pairwise comparison is subjected to a more rigorous pruning method, as certain sequences may not present substantial variance capable of surpassing the defined threshold values, where ΔW was the most effected feature. Although the positive correlation between the ΔW and $\bar{P}_d$ persisted in the $T-2$ analysis, no significance was obtained due specifically to a lack of available comparisons. However, given a higher dataset a similar trend would have been observed, as the p-value obtained was substantially close to significance. On the other hand, a significant correlation is obtained with $\int W$ and both the $\bar{P}_d$ and $\int P_d$ GSR features. Unlike the ΔW, $\int W$ obtained a higher number of comparisons reinforcing the previously suggested values. This also confirms the potential of utilising relative measures in comparison to the mean

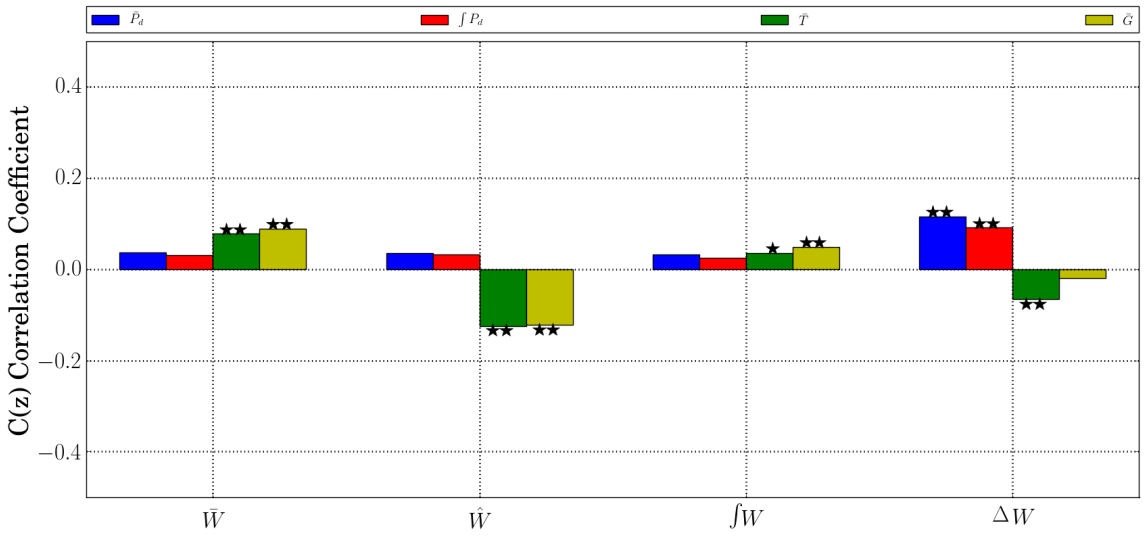


Figure 7.17: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **reactive** window type and annotator memory (all windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

and peak-to-peak feature types, as both correlations did present some degree of robustness despite the loss of data features.

In order to compare between the diverging windowing methods, the same previous pairwise comparison correlation analysis was realised for the reactive windowing methods. Figure 7.17 showcases the obtained pairwise correlations over the same number of runs (i.e. 40 runs) using the reactive window method. A similar number of comparisons to the previous global correlation analysis was obtained, where the total average of pairwise comparisons between each feature correlation consists of ≈ 2655. A very similar trend is observed between both the reactive and continuous global correlation analyses, with some slight differences. In particular the negative correlation trend between the peak-to-peak with both the tonic and global means becomes more pronounced with reactive windowing surpassing the -0.1 correlation value. Expectedly the varying window size of continuous windows, which can be either longer or shorter than reactive windows depending on the sequence of events, directly influences measures dependent on value such as mean and peak-to-peak. Although the mean was not particularly affected, it is not unexpected for peak-to-peak measures to be heavily influenced considering that it consists of the maximum amplitude obtained from an entire window, which is substantially more sensitive towards the fluctuation of annotation values. Given the highly significant negative correlation found between peak-to-peak and both tonic and global GSR features, does suggest that there is an opposite tendency between these features. However, the mean did not suffer significant changes, which is likely due to the inherent nature of the annotation system consisting of square shaped signals. If a participant is excessively idle one particular value will dominate the average, while in the situation of peak-to-peak a slight change in value will have a larger impact. Thus it is difficult to determine exactly why these correlations exist, although we maintain our previous theory that the positive bias present within the tonic and global features can significantly influence these correlations, hence why they are more prominent within the global comparisons. Furthermore, it is important to point out that these correlations are less apparent in

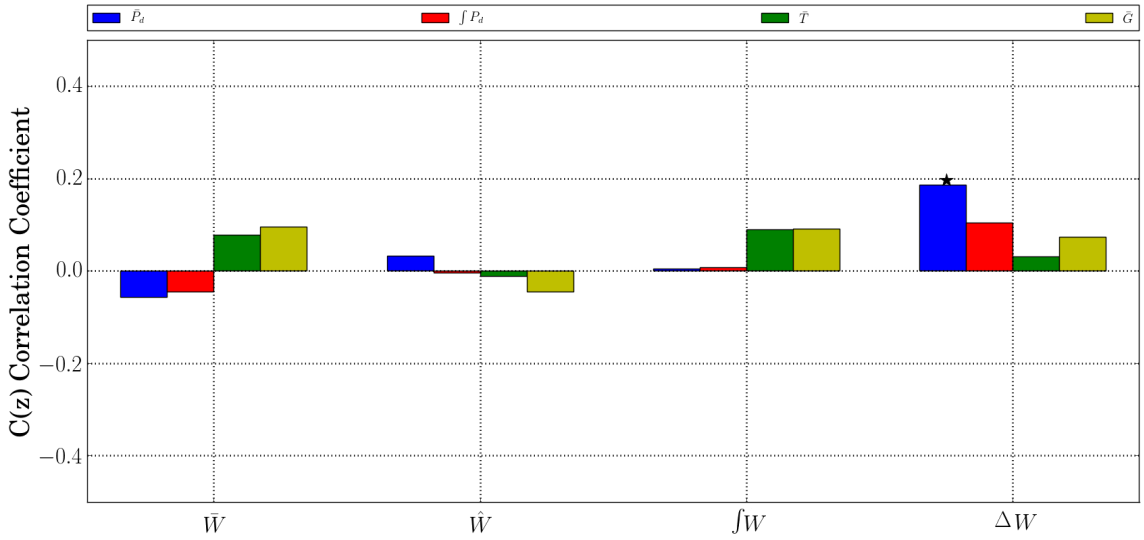


Figure 7.18: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **reactive** window type and annotator memory ($T-1$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

the relative features, which also suggests that these features in particular are likely more robust to noise present in the data. The $\int W$ feature obtained similar correlation coefficients, where the main divergence was the significance of $\bar{P}_d$ and $\int P_d$. The ΔW annotation feature was the less effected by the reactive windowing method, where the correlation coefficients obtained relatively similar values comparatively to the continuous windowing method.

Continuing the reactive windowing analysis, Figure 7.18 presents the direct adjacent window ($T-1$) correlation coefficients between the GSR and video annotation features. Although fewer number of comparisons were observed in comparison to the previous $T-1$ correlation analysis, the difference was not particularly substantial. The average total of pairwise comparisons between all different correlation analyses was ≈ 225. Results observed follow a similar trend to the continuous $T-1$ analysis, where the reactive windowing did in fact increase the correlation between ΔW and $\bar{P}_d$, where the exact value obtained was 0.186. Given our assumption that player annotation is also reactive, it does make sense that there would be a closer relationship between the phasic driver and the gradients of player annotation specifically. This is also aided by the windowing mode utilised, which intends to extract the specific reactionary post-event subsections of the annotation signal. No other significant correlations were observed when analysing the signal locally, in particular the correlations with tonic and global GSR features. This further suggests that these tendencies are not reflected when comparing the signal locally, and thus this pattern emerges once the analysis drifts away from directly adjacent window observations, and moves towards a more "global" observation.

Figure 7.19 reduces the locality and presents the correlation coefficient results between the pairwise comparison of two directly adjacent reactive windows. An average total of ≈ 331 pairwise comparisons between all diverging correlations analysed was obtained after applying the threshold filtering. As expected by reducing the locality emerging patterns observed in the global correlation analysis start to appear. Specifically the tonic and global GSR features become more prominent when comparing between the mean and, in particu-

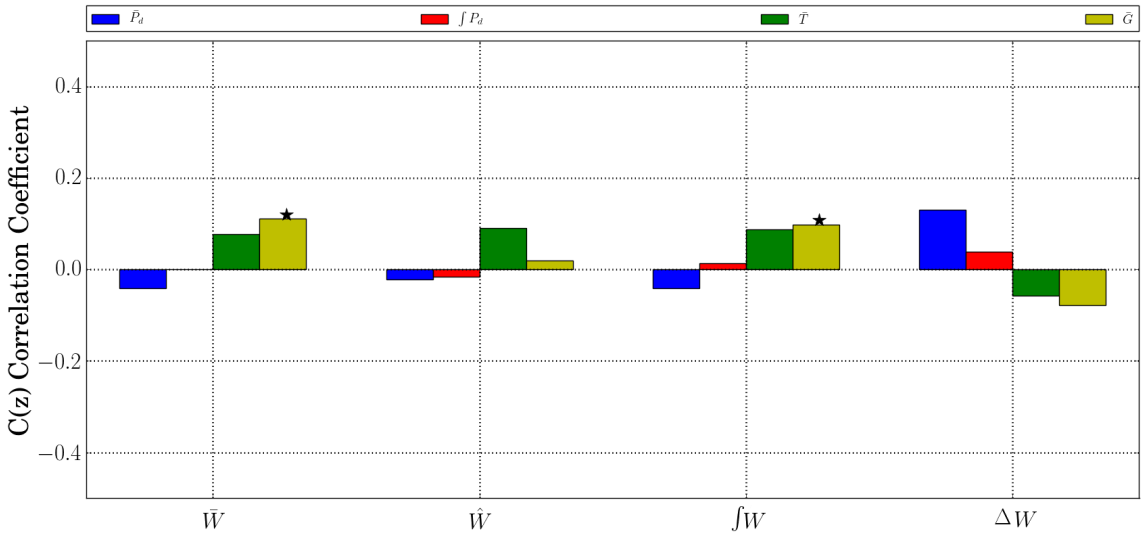


Figure 7.19: Rank correlation values between metrics of the normalized annotation values and the features of SC, computed across the **reactive** window type and annotator memory ($T-2$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

lar the integral features of video annotation. Surprisingly the integral feature suffered the most divergences relatively to the previous continuous $T-2$ correlation analysis. By observing both the global and $T-1$ reactive correlations the trend of $\int W$ is distinguishably apparent, while comparatively to the continuous observations more correlation emphasis emerges from the phasic driver features specifically. Although it is difficult to explain why this phenomenon emerges exactly, it is apparent that the windowing method does have a substantial effect on certain features. Given the close relation of $\int W$ to the $\bar{W}$ annotation feature, we theorize that in this particular situation the integral acts like a mean because it is incapable of capturing the entire curvature due to the small windowing. Furthermore, due to the square shape of annotation signals, it is very likely that only the constant value of participant idling is extracted from such a small window. Thus, it is possible that a larger window could produce more consistent results. Apart from the previous feature, the remaining annotation features obtained similar non-significant correlation coefficients.

Skin Conductance and Level Progression

This section investigates the relationship between each level's progression values, obtained through *Sonancia*'s level generation methodology, and the skin conductance signals obtained from the participants playing the game. In order to create a relationship between the level progression and actual gameplay events, a player trace is created based on the room change event from the game state logs. This player trace consists of the visited room sequence (i.e. the path) of each player during gameplay. Additionally, a filtering process was applied on the events, in order to remove short sequences of alternating room changes. Once a player trace is obtained, a progression based on the player's trace can be computed similarly to how the level progression is obtained for fitness evaluations (see section 4.3), which will serve as the level progression. This method allows for the effective comparison of actual values utilised for level generation and a player's gameplay run. Similarly to the previous section, only stable skin conductance signals were utilised for the purposes of this study,

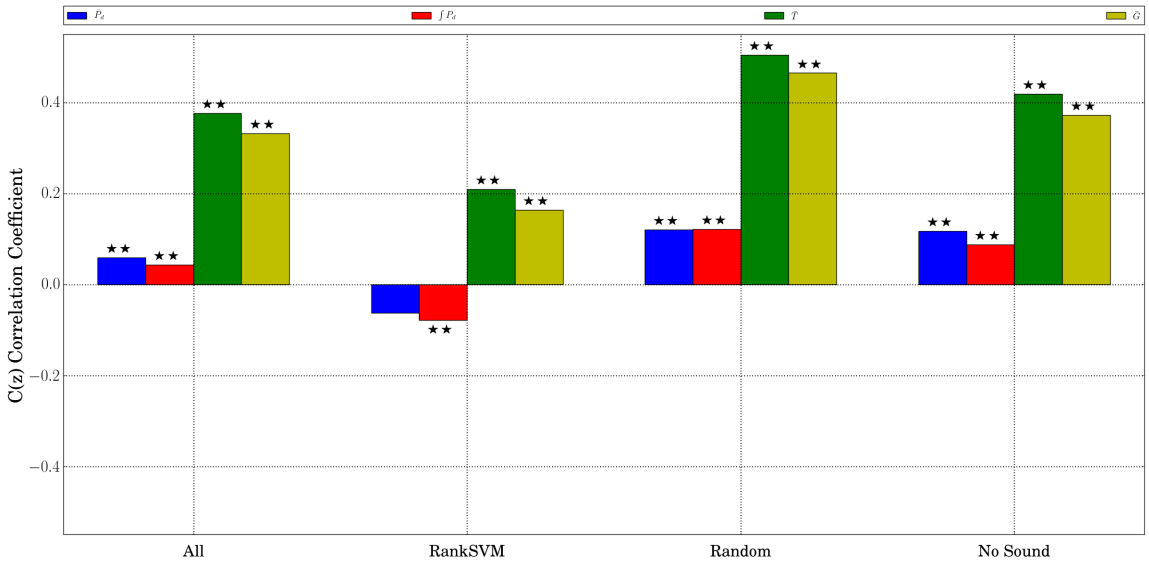


Figure 7.20: Rank correlation values between metrics of the level progression values and the features of SC, computed for different audio model variants and all windows. *All* analyses the combination of all audio models, while *RankSVM*, *Random* and *No Sound* analyses the correlation of levels where their models were exclusively utilised. Significant values are in bold [0.05 (*) and 0.01 (**)]

and the same threshold method was used in order to reduce the noise from GSR features.

In order to further investigate how each audio model influenced a participant's physiology, three additional studies were also conducted. Correlation between the player trace and physiology was also analysed restrictively on runs where the RankSVM or Random models were used, in addition to the level where no diegetic sound was used exclusively (i.e. No Sound). Each model was tested equally in different orderings with the same number of participants as in the previous analysis; the only exception was one run of the RankSVM variant which had to be discarded due to a noisy GSR signal. In total 14 runs of the Random and No Sound, with the 13 runs of Rank SVM levels were individually analysed against the intended player trace and the actual physiology of players. For both Random and No Sound all runs are equally split between both levels, while the additional run for RankSVM was in the first level.

Figure 7.20 presents the binomial-distributed correlation values of the player trace and the skin conductance features. The most apparent significant positive correlation consists of both the tonic and global means, which persists throughout all level variants. This heavily implies that the intended player trace values does indeed fluctuate similarly to at least the tonic and global features. However, it is important to note that the tonic and global GSR features tend to be positively biased, as the signal tends to consistently increase and rarely decrease during play. This is also true for the player trace, as the increase of tension tends to be more aggressive. In fact monsters tend to increase tension immediately by 1, while to decrease tension a decay effect must occur, or a light source must be present. This makes tension decrease at a slower pace as the values are lower, requiring that more rooms be visited in order to effectively decrease tension. Thus it makes sense that a high degree of correlation exists between the two when comparing the signal globally, as there

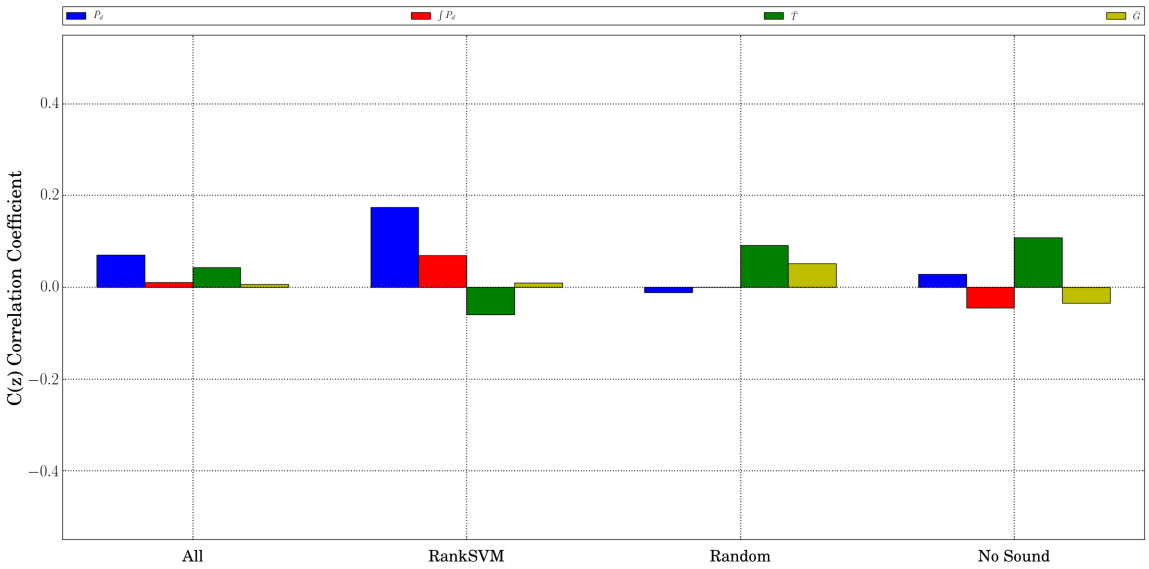


Figure 7.21: Rank correlation values between metrics of the level progression values and the features of SC, computed for different audio model variants and $T-1$ adjacent windows. *All* analyses the combination of all audio models, while *RankSVM*, *Random* and *No Sound* analyses the correlation of levels where their models were exclusively utilised. Significant values are in bold [0.05 (*) and 0.01 (**)]

is a heavy bias towards positive increasing values. Interestingly, RankSVM was the only variant to present a lower correlation value for both the tonic and global features, and negative correlations for the phasic features. These values suggest that some divergences do exist between the different level types, where audio models did in fact have some influence on the players physiology.

Based on the $T-1$ correlation analysis (see Figure 7.21), we can observe that comparison locality heavily influenced this correlation, we theorize that this potentially relates to the discrepancy between the actual player trace values and the combination of play and audio. In fact, the latter makes sense considering that the player trace changes significantly based on the amount of times a player revisits a room, which will gradually increase the discrepancy between the actual value of the room, i.e. the value assigned for sonification, and the value of the player's trace. Thus it is possible to have a room with low tension, but due to the player trace be considered high tension due to previous experiences. Both Random and No Sound obtained significant correlations with both phasic drive features, suggesting that by using either a random audio predictor or just diegetic sounds the system was able to closer relate the level tension to the player's physiology. Although it is difficult to exactly pin-point why RankSVMs were the only variant to obtain negative correlations, it is a curious phenomenon. Even though these results seem to suggest that the RankSVM level variants performed worse overall, the $T-1$ results offer a different perspective.

Figure 7.21 showcases the correlations of $T-1$ adjacent pairings for all level variants. It is immediately apparent that locality of pairwise comparisons significantly changes the coefficients. Comparatively to the global comparison no correlation obtained significance, which was caused particularly by an insufficient number of comparisons. However, it is interesting to note the higher correlation value of the RankSVM variant with $\bar{P}_d$. Although

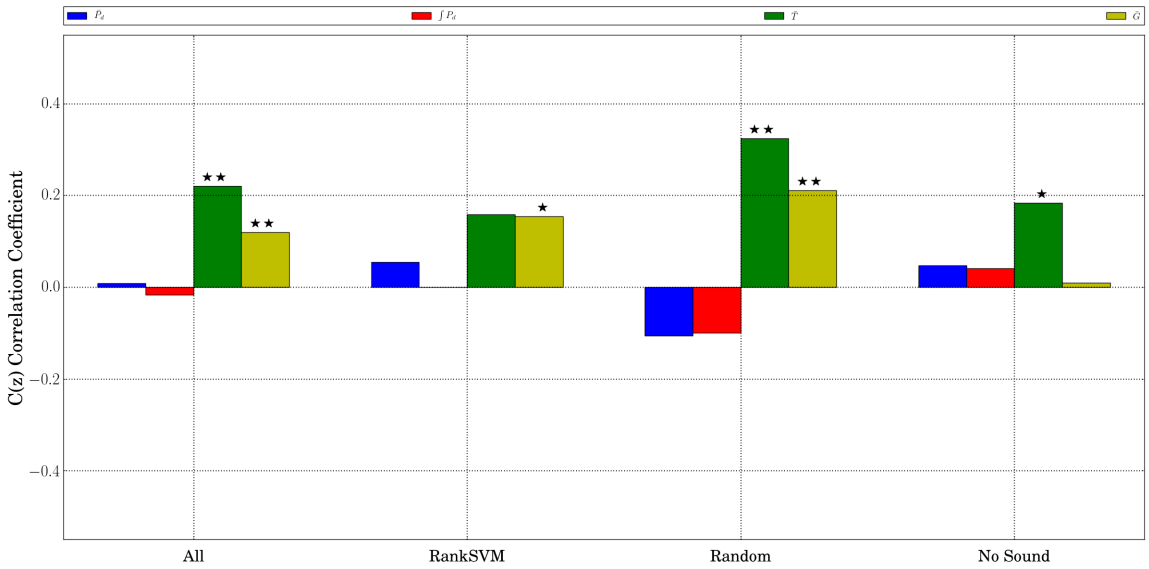


Figure 7.22: Rank correlation values between metrics of the level progression values and the features of SC, computed for different audio model variants and $T-2$ adjacent windows. *All* analyses the combination of all audio models, while *RankSVM*, *Random* and *No Sound* analyses the correlation of levels where their models were exclusively utilised. Significant values are in bold [0.05 (*) and 0.01 (**)]

it is not significant, this particular correlation still obtained a p-value of 0.08, which suggests some degree of agreement between both feature types. It also points towards the importance of locality, while the other variants obtained high correlations with global pairwise comparisons, when restricting the comparison to a sequential order the previously high correlations diminish. As previously suggested, this is due to a positive bias of both the player trace and some features of skin conductance. More precisely, situations where tension decay occurs are less frequent and aggressive than rising tension events. Thus, when comparing globally an overwhelming amount of positive rising bias occurs, hence why tonic and global features often perform well. By analysing pairs in direct sequence, the amount of bias is substantially cut as the focus is on instant trends, which give more emphasis to the decaying subsections. Thus, we argue that this method more clearly demonstrates the relationship between player physiology and the intended level tension due to the lower positive bias. These results also suggest that using a machine learned predictor such as the RankSVM, the system was capable of stimulating physiological responses from players that more closely resembled the intended tension value. It is important to note that due to the stochastic nature of the game and the lack of physiological data, it is difficult to claim these results as anything but preliminary. However, it does infer that there is some viability of multi-faceted generators.

Finally, Figure 7.22 presents the correlation coefficients of the $T-2$ adjacent pairwise comparisons of each level variant. By increasing the locality by one additional sequential pairwise comparison, patterns from the global correlation analysis start to emerge. The most prominent being the correlations between tonic and global features. This further suggests the aggressiveness of player traces and both GSR features in its ability to dominate trends such as tension decay. Apart from these patterns no other significant correlation is

observed. Despite the increase of locality the RankSVM and $\bar{P}_d$ feature still presents a positive correlation, although quite less substantial. Interestingly, the Random model also obtains a negative correlation albeit not significant, it does contrast with results obtained with the global pairwise comparison. This does point towards a lack of robustness with this particular correlation, reinforcing our previous claim that a direct sequential comparison is better suited for this specific experiment.

The emergent behaviour inherent in games can lead the player to be subjected to a wide variety of different stimuli, not foreseen beforehand. This does reinforce the argument that games are in fact multi faceted experiences, and that constructing levels by looking exclusively at one point of view (i.e. the designer), might not be the ideal solution. However, the idea of defining intended experience can still be a viable solution, although it would need to take into account both the stochastic nature of digital games and their interactivity. Ideally, level generation could include both points of view, where the level generation's intended experience moulds itself according to a player experience model, similar to the work of Pedersen et al. (2009, 2010). Furthermore, if an online solution is utilised, this experience could build more personalised content, where sounds are chosen utilising the audio models presented in this thesis, in conjunction with context sensitive gameplay data, for example. However, constructing online procedural content generators is not a trivial task, especially if utilising readily available game engines, which are strict about the rendering and construction of the 3D environments. An additional problem also presents itself when attempting to use non-playable characters, such as the monsters present in the *Sonancia* game. If NPCs are required, navmeshes must be built on top of the level architecture, so that they can effectively traverse the level through a path-finding algorithm.

Gameplay Annotations and Level Progression

The final analysis consists of investigating the relationship between perceived tension obtained from video annotations and the level progression. This study provides an insight on how the intended tension derived from the system is perceived by the players. Furthermore this study, similarly to the previous one, will investigate if the different audio model variants affect the perception of tension and if by using an audio model can this perception get closer to the system defined tension.

Similarly to the previous two sections, both the video annotations and the level progression values were derived utilising the same methodologies. Both the continuous and reactive windowing methods are analysed, in addition to comparing the signal globally and sequentially ($T-1$ and $T-2$). A compatible level progression to the video annotations is derived by analysing the player's gameplay trace exactly like the previous analysis. The same dataset was used due to its balance between the different level variants, and a threshold of 0.02 was used to filter non-variant window sequences.

Figure 7.23 showcases the correlations between level progressions utilising different models and the video annotation features. The highest correlations obtained was with the RankSVM and No Sound runs, obtaining ≈ 0.21 significant correlation with the $\bar{W}$ and $\int W$ annotation features, respectively. The No Sound level variant was the most consistent, further obtaining a significant positive correlation with the $\bar{W}$ feature. Contrarily RankSVMs were less consistent, where a significant correlation was only obtained by correlating the average feature. Interestingly the random level variant consistently obtained negative coefficients, although they are significant the exact correlation value is not particularly low (≈ -0.08). The observed results again suggest that the model variants did

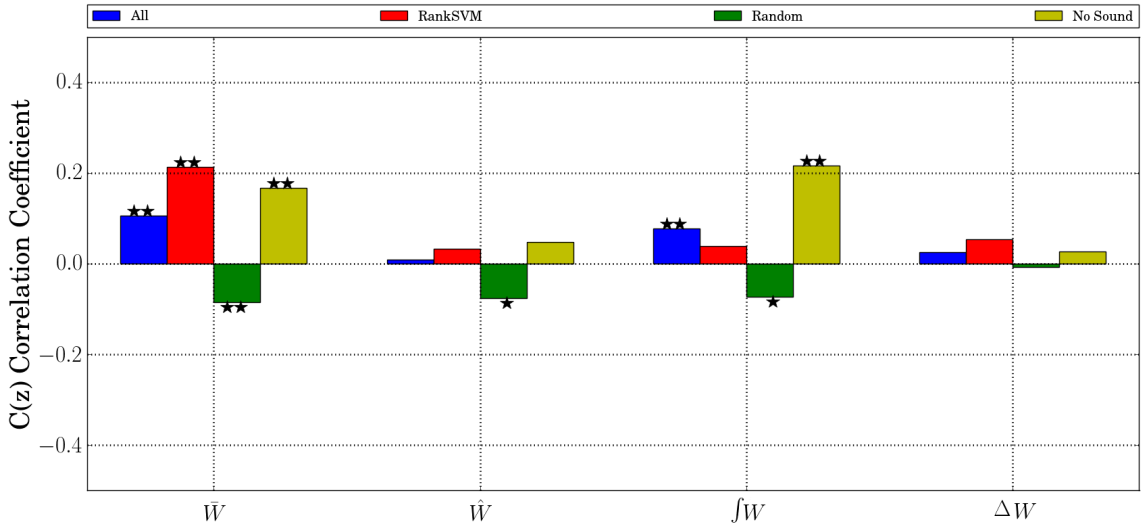


Figure 7.23: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **continuous** window type and annotator memory (all windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

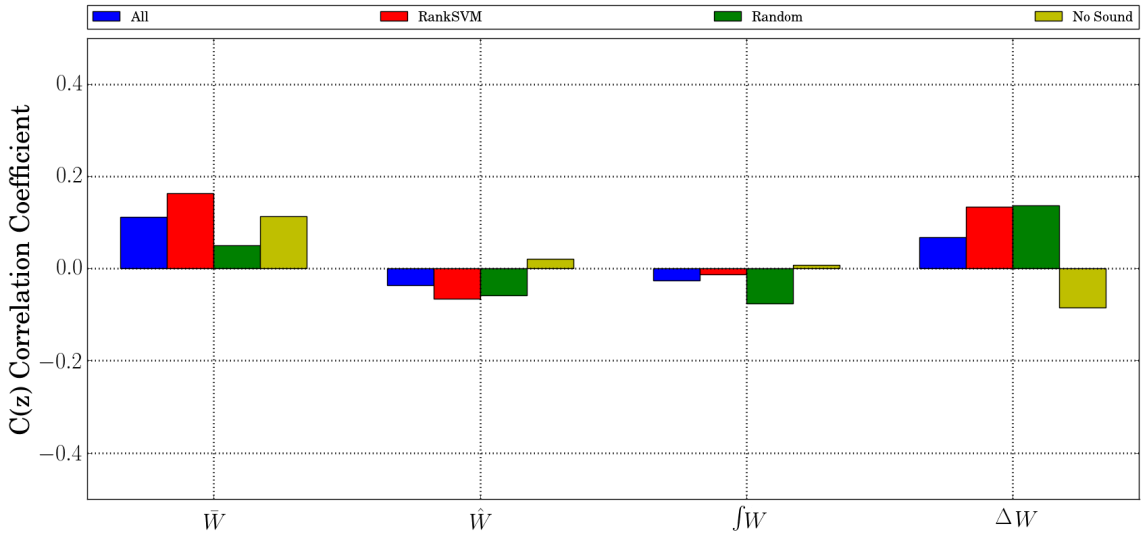


Figure 7.24: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **continuous** window type and annotator memory ($T-1$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

affect the participants perception, where both the RankSVM and No Sound playthroughs did present a closer relationship to the intent of the level in comparison to using Random.

In order to investigate the impact of locality on the correlations of both level progression and participant annotations Figure 7.24 showcases the direct sequential $T-1$ pairwise comparisons. Although some correlations do persist, albeit without significance, the $\int W$ annotation feature in particular ceases to present any type of correlation with the differ-

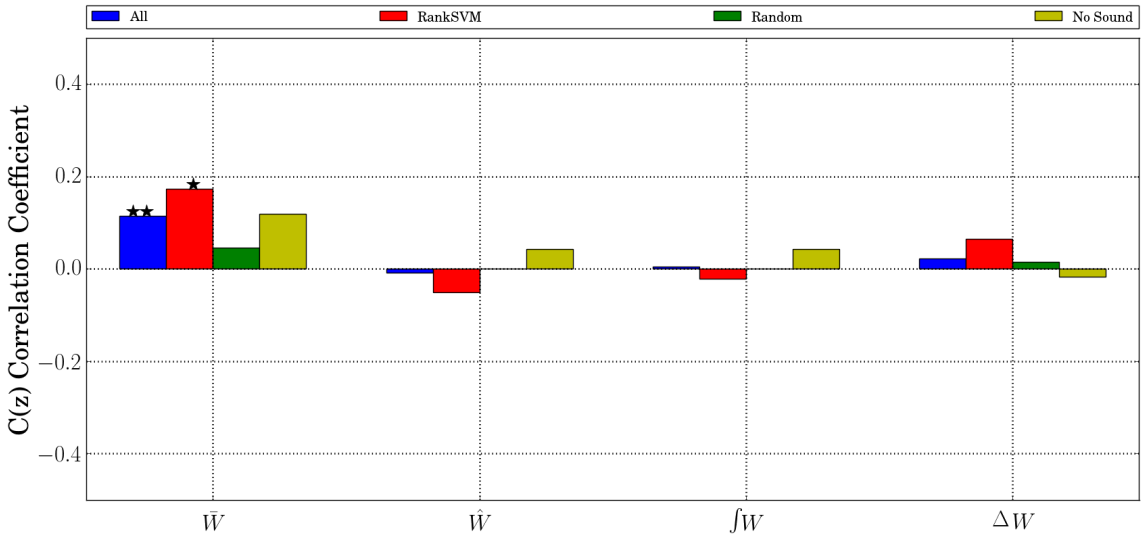


Figure 7.25: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **continuous** window type and annotator memory ($T-2$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

ent level progressions. On the other hand, $\bar{W}$ does present a similar pattern to the global pairwise comparisons, although due to a lack of data does not obtain significance. Unlike previous results the Random level variant does present a positive correlation with the gradient feature, achieving a comparable result with RankSVM. Similarly to previous results, the correlation locality plays a substantial role for certain features, we continually suspect that this is due to some degree of positive bias in the player trace specifically, which is reduced as pairs are not compared among diverging pairs in the entire signal.

Figure 7.25 presents the results obtained by calculating the correlation of pairwise $T-2$ adjacent windows of video annotation features and the player trace. Among all results obtained the most consistent feature was $\bar{W}$, where despite the diverging pairwise comparison methods, the correlations between each level variant show very similar tendencies throughout. In particular the RankSVM levels consistently obtained higher correlations with video annotations, suggesting that the perceived emotion from annotators was closer to the intended level progression than other model variants. Alternative features such as $\hat{W}$, $\int W$ and ΔW were substantially more erratic, causing correlations to consistently diverge.

In order to further investigate how the windowing methodologies can directly effect the correlations between video annotation and level progressions, the same analysis was conducted with reactive windows for extracting video annotation features. Figure 7.26 showcases the obtained correlations calculated from global pairwise comparisons. The $\bar{W}$ video annotation feature, where despite minor discrepancies in the RankSVM and No Sound models, achieved very similar correlation coefficient values to the ones observed in the previous continuous window analyses. Interestingly, the $\int W$ also obtained a very similar pattern, almost mimicking each correlation of $\bar{W}$ with the RankSVM obtaining a slightly higher value. Furthermore, unlike the previous global experiment, the reactive windows did in fact significantly improve the correlation between RankSVM levels. For $\hat{W}$ both No Sound and Random obtained a significant negative correlations, although previously the

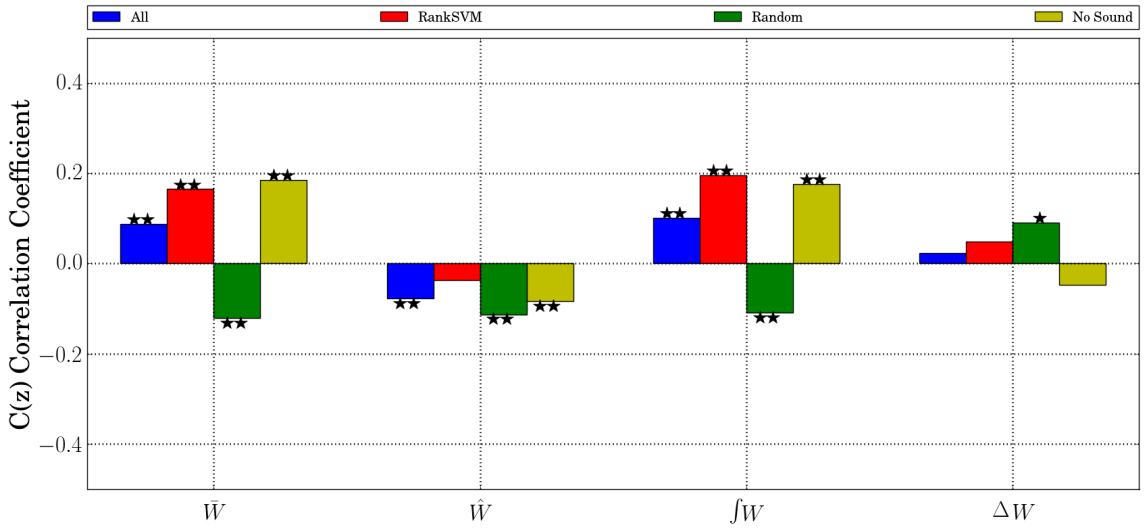


Figure 7.26: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **reactive** window type and annotator memory (all windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

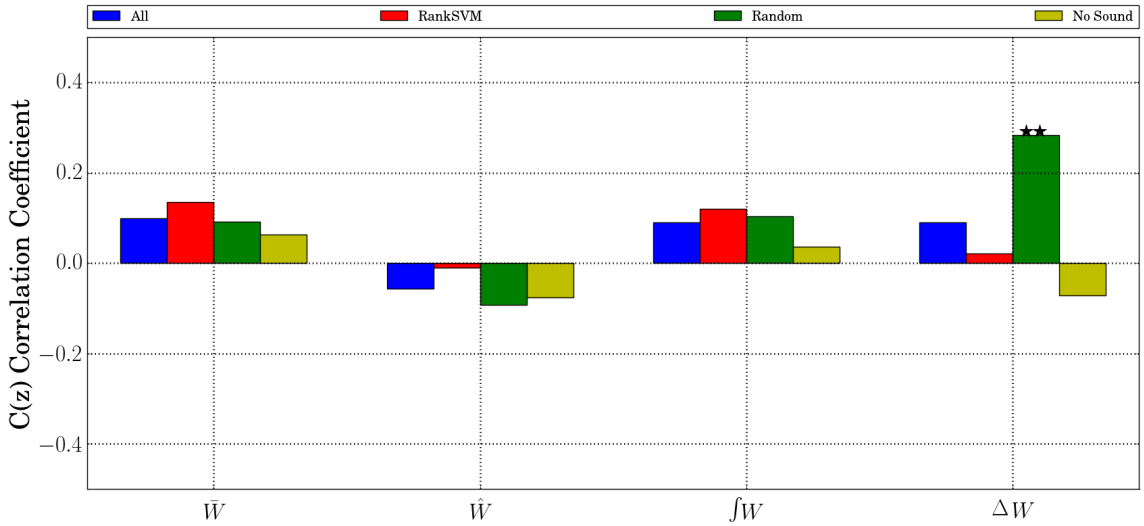


Figure 7.27: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **reactive** window type and annotator memory ($T-1$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

No Sound level variant showed no sign of correlating with this particular feature. However, given the erratic nature of $\hat{W}$ and its dependency on window size it is not surprising that these results can alter substantially in comparison to the continuous windowing feature. Surprisingly the random level obtained a significant positive correlation with the ΔW video annotation feature, although there was a positive correlation between random and ΔW in the previous $T-1$ analysis, this is the first instance of a significant correlation between both features.

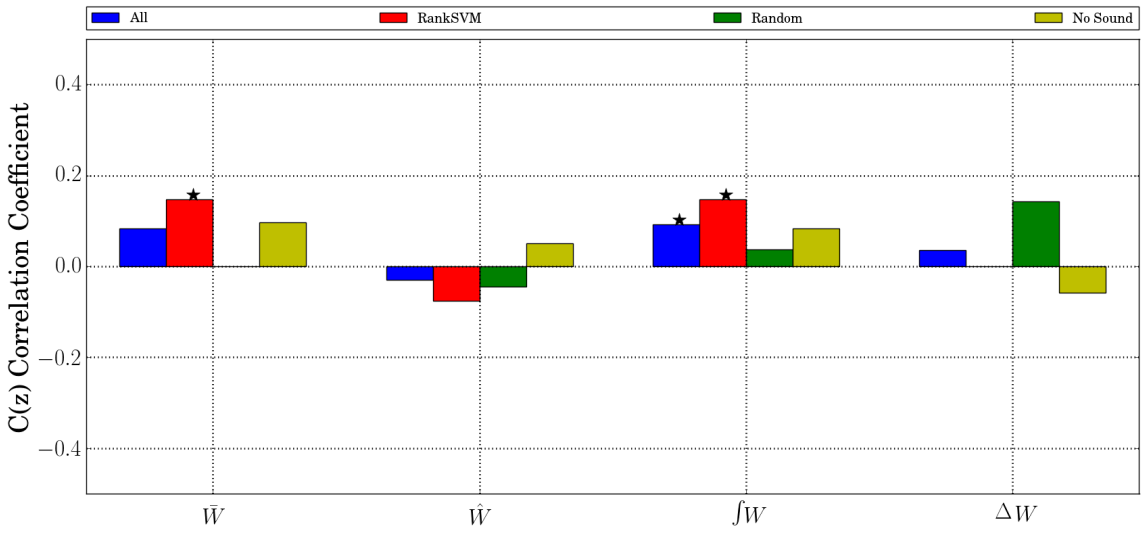


Figure 7.28: Rank correlation values between metrics of the normalized annotation values and the level progression values, computed across the **reactive** window type and annotator memory ($T-2$ adjacent windows). Significant values are in bold [0.05 (*) and 0.01 (**)]

Figure 7.27 presents the correlations obtained by comparing the direct adjacent reactive windows of video annotation with the player trace. The most glaring discrepancy compared to the previous correlations studies, is the significant positive correlation between the ΔW annotation feature with the Random level player traces. This particular correlation achieved a coefficient value of 0.28, suggesting that the slope of annotations did frequently align with the player traces of Random levels. Although this discrepancy does suggest particular fault with the data, a thorough analysis suggests that no particular outliers are present, although some runs did in fact present some degree of overlapping windows, their removal however did not significantly alter the presented results. Furthermore, some runs of the random level did present some degree of interaction with monsters, which can influence the perceived tension annotation. Although it is difficult to exactly pin-point the reason why this phenomenon occurs, these results does suggest that the random model did in fact influence the perceived tension of participants. Furthermore, it was particularly effective with the relative gradient feature, where the direction of the annotation slope has a similar tendency to that of the player trace. Interestingly, once this particular correlation's locality is increased the coefficient value diminishes, suggesting that there is a particular sensitivity to these values which can relate to some degree of noise in the annotation data. Similarly to the previous $T-1$ correlation analysis RankSVMs did not obtain significant correlations, although the coefficients for both $\bar{W}$ and $\int W$ did show resembling values. Another similar trend to previous analyses was the lower correlations of the No Sound level variants by using the $T-1$ pairwise comparisons. We theorize that this is due to how No Sound levels are perceived as situational. More precisely, because monsters are placed at certain locations within the level it makes sense that a monster sighted in room 1 and 6 are correlated, as the player annotation increases due to the sighting and so does the player trace. This might also explain why No Sound levels tend to have higher correlations on global comparisons.

The final analysis presented in Figure 7.28, consists of the correlations of adjacent $T-2$ pairings of the player trace and annotation features extracted with the reactive windowing

method. This final analysis reinforces the robustness of RankSVM correlations, which despite locality and windowing methods, presented a similar correlation trend throughout. Although the most consistent annotation feature was $\bar{W}$, where coefficient values ranged from 0.15 to 0.21 for both annotation windowing methods, the $\int W$ feature in particular presented similar correlations once the reactive window was used. The fixed size of reactive windows presented a more reliable feature, as some situations the length of the continuous window can potentially be detrimental for the integral feature. However, for No Sound the continuous window was not particularly detrimental for the integral feature, although it is important to note that the window length can substantially change from run to run and depends on the player. As expected by increasing the locality of pairwise comparisons the Random and ΔW correlations diminish, although the value still reaches 0.14 albeit without significance.

7.5 Discussion

This thesis proposes a multi-faceted procedural content generator by combining the capabilities of search-based and machine learning methodologies. Levels are generated through defined designer guidelines, which within the context of this thesis are framed as progressions of tension. These guidelines inform the level generation system in the construction of a procedural level which is then subsequently sonified. Through the aid of a machined learned model capable of ranking audio based on perceived affect, the level generation system allocates sounds based on both the desired tension of each subsection of the level and the predicted audio rankings. In order to validate the quality of our proposed system, a validation study using human participants was conducted. The purpose of this study was to verify the claim given at the beginning of this thesis: "Can our system, i.e. *Sonancia*, construct a gameplay experience that follows a desired progression of tension?". Furthermore this thesis suggests that a more accurate experience can be met through the combination of both search-based methodologies and machine learned predictors. Thus in order to validate our claims human physiology data, gameplay logs and participant annotations were collected from two diverging generated levels, where different models of sonification were applied.

This chapter investigated the correlation between the players trace and the extracted skin conductance features at the moment of room traversal. This is defined as such due to how sound works within the game, which is local to each room, thus changing rooms changes sounds. Furthermore, due to the interactive nature of digital games it is difficult to constrain players to the exact progressions of the level (i.e. the main path), thus a compromise was made where a level progression is derived through the player trace (i.e. the sequential ordering of rooms visited by the player). The derived level progression consists of the exact same methodology used for generating levels, effectively emulating this concept. Although it is not exactly a 1 to 1 comparison, it offers insight on the effectiveness of different sonification variants in their ability of tightening the gap between a progression and actual player physiology, through the sounds these models suggested prior to play. Three types of correlation analyses were conducted in order to study the effect of locality, where one analysis assumes that the tendency of the signal is constant throughout, while the others assume that the tendency is merely local either directly adjacent $(T-1)$ or by a sequential subsection $(T-2)$. All different assumptions provided a wide range of results, where the consistently highest correlations were obtained by the tonic and global skin conductance

features, with the exception of the $T-1$ correlation analysis. Although this suggests that some degree of agreement exists between these features and the player trace, there is reason to believe that this is specifically caused by how the player trace is calculated, in addition to the own inherent tendencies of the tonic signal. Throughout other correlation studies presented in this chapter, such as the annotation and skin conductance correlations, both tonic and global features often presented high or low coefficient values. This trend tends to be more prominent in global comparisons as these features present slow response times to stimulus, and thus have a tendency to slowly increase over time. In fact, it makes sense that these features outperform within the global perspective as they have a continuous tendency, while phasic features tend to appear as a direct response to stimuli (presented as a peak in the signal). For this particular reason the work of Holmgard et al. (2013), Benedek and Kaernbach (2010), Bach and Friston (2013) and Boucsein (2012), tend to commonly agree that phasic features are more accurate representations of the manifestations of stress, due to its instant response to stimulus. When analysed from a global perspective the player traces of RankSVM tend to negatively correlate with the phasic features, which implies that there is a tendency for disagreement. However, once the correlation is viewed from the $T-1$ perspective it obtains the most positive coefficient value out of the different variants. Even though this correlation is not significant, it does point towards a discrepancy between global and local comparisons. This divergence is due to the fact that the phasic driver is an almost instant manifestation of stress in relation to a stimulus, which is represented by a numerical value expressing its intensity. Thus in the global correlation analysis we assume that the intensity of phasic manifestations is partially influenced by all of the previous manifestations, while in the $T-1$ analysis assumes that intensity is only partially influenced by the previous manifestation. Thus, what these results suggest is that the Random and No Sound variants obtained intensity manifestations that were consistent throughout an entire run, meaning that if room 1 is more tense than room 5 in the progression, this was slightly reflected by the intensity of phasic features. On the other hand RankSVM obtained direct tendencies, where the overall intensity value only matters when comparing to the directly adjacent value. Thus minor phasic increases can be reflected as correlated with a major increase of tension in the player trace. It is important to remember that the threshold filter does remove minimal comparison differences, thus only non-ambiguous phasic values are compared. Given this, it is difficult to exactly pin-point which model was more consistent in providing the closest level progression experience. What can be affirmed however, is that the No Sound and the Random model runs provided a consistent correlation on the overall level progression, while the RankSVM model provided a more consistent relation between the actual trend of the rise and fall of a level's tension.

In order to further investigate the relationship between level progression and player affect, an additional correlation analysis was conducted between the annotations and player trace. The objective of this particular analysis was to investigate how the perceived tension, annotated by the players, is compared to the level progression. This study followed the same pattern of previous video annotation analyses, where windows were defined based on the continuous and reactive windowing methods, where video annotation features were subsequently extracted from these windows. The overall results show that the RankSVM levels were often more consistent in correlating with the player trace. Furthermore, the pattern was quite consistent despite locality and windowing method. This consistency and the correlations themselves do reinforce that the perceived tension was at least somewhat influenced by the RankSVM variant, as the annotations on these particular runs tend to follow the values of level progression more closely. Interestingly, it is important to remem-

ber that these models were also trained on audio annotations of perceived effect, thus these results do align with the training method utilised for the construction of this model. The No Sound variant comparatively performed better with the global comparisons, which in retrospect is logical given that monsters were the principal stimulus of these particular levels. More specifically, there is a tendency for participants to rise tension at the sight of a monster, similarly to how the player trace is derived. In the areas where no monsters are present, it is quite common for participants to stay idle during the annotation process rather than decrease the tension value. These values are consequently removed by the threshold filter. Thus, the obtained tension annotation has a higher probability of staying consistent to monster locations due to the way annotation progresses through time, even if the annotator is idle the position of the annotation cursor is still carried over once idling terminates. Although the Random level variants did not obtain strong correlations in this analysis. One particular discrepancy was present in the $T-1$ reactive window analysis, where a significant positive correlation was obtained. Although it is difficult to determine exactly why the correlation between the player trace and ΔW annotation feature was observed, these results do suggest that Random levels did indeed achieve closer correlations to the player trace. Given that positive random level correlations were only observed in the reactive windowing method, it is safe to assume that the varying window sizes were detrimental to the gradient values of this correlation, as such results were not observed with continuous windowing which consist of a substantially larger set of data points. Furthermore, the majority of correlations observed for Random was with the gradient feature, meaning that the direct fluctuation of annotations often correlated with the rise and fall of progressions. Similarly to the previous analysis between level progression and skin conductance, no specific model severally outperformed others. Although it is true that the Random variant obtained the highest significant correlation, it was particularly less consistent than either RankSVM and No Sound. Random is also particularly sensitive to locality, where an additional comparison ($T-2$) severally reduces the correlation coefficient. In terms of consistency RankSVM outperformed the other variants, where the best results obtained was with the global comparisons.

An additional contribution offered in this thesis through the realization of these experiments was the development and analysis of a new annotation methodology. Emotion annotation is a fundamental aspect of affective computing, where the annotation of emotion is critical for labelling and subsequently constructing machine learned predictors of affect. Thus it is necessary to build reliable annotating measures capable of minimizing the discrepancy between annotation and the ground-truth. Unlike previous continuous annotation methods such as FeelTrace (Cowie et al., 2000) or Gtrace (Cowie and Sawey, 2011) which are bounded to an interval of $[0, 1]$, our proposed methodology is unbounded allowing participants to increase and decrease their annotation without constraints. In order to study the reliability of the proposed annotation system, an extensive correlation analysis between the participant's skin conductance signal and their tension annotation was realized. Similarly to the other studies both the continuous and reactive windowing methods were utilised in order to study the divergences between varying windowing methods. The most consistent correlation observed within each of the analyses realized was between the $\bar{P}_d$ skin conductance feature and the ΔW annotation feature. Although this correlation was non significant in the $T-2$ pairwise comparison study, among the remaining correlation experiments these particular features consistently obtained significant correlations, where the highest coefficient value observed was 0.186. Thus it is safe to assume that the linear correlation between ΔW and the phasic features show increasing potential, where

future studies could eventually utilise such features in the construction of machine learned predictors.

7.6 Summary

This chapter presented an in-depth user experiment where several participants were tasked in playing 2 diverging levels generated by the *Sonancia* and sonified by three different audio models: RankSVM, Random and No Sound. The first two sections of the chapter presented the experimental methodology, where the data collection process was defined, from skin conductance measurements, game state logging and gameplay annotation. The following section then presented the feature extraction methodology, where detailed descriptions of how features were extracted from skin conductance and the real-time annotation system. The final section of the chapter presented several studies from the obtained user experimentation. The first study presented an in-depth statistical analysis of the different player runs, where the average completion time of each level was analysed and all the player traces of each level. Results showed that players took longer to complete level 2 as it presented more alternating paths than level 1. Furthermore this chapter introduced a new annotation tool, for a continuous yet unbounded and relative annotation of affect. The tool's own interface promotes relative-based annotation as it relies on a wheel-like interface; where the annotator can reference his previous annotation trace without bounding limits. To test the efficiency of this system, each participant was tasked in annotating the perceived tension of their own playthrough after play. The correlation between annotation and skin conductance was subsequently investigated. The relative metric of average gradient of the annotation traces obtains the most consistent and robust correlation with the phasic driver of skin conductance, regardless of type of comparisons and windowing methods. The following study subsequently explored the correlation between a derived level progression with both the skin conductance and the player annotations, in order to investigate how close an intended level progression can be to actual human data. Furthermore, each derived level progression was compared to the skin conductance and video annotations of experiments whose levels were sonified by the diverging audio models. These results showed that audio model predictors did in fact perform effectively well compared to the different variants, obtaining the most consistent and robust results. Although some RankSVM correlations did prove worse than the other variants in some experiments, the preliminary results suggest that some degree of potential exists for multi-faceted system capable of combining both search-based approaches combined with machine learned predictors.

Chapter 8

Discussion and Conclusions

This thesis proposed a methodology for the procedural generation of multi-faceted game content. Machine learned predictors are used to derive a global ranking of the perceived tension of different audio pieces, while evolutionary algorithms are used to construct levels based on a predefined progression of tension. A rule-based method is subsequently applied on both the generated level and the predictive global ranking in order to effectively orchestrate sound within the level. A proof-of-concept system was developed in order to showcase a methodology capable of addressing the following questions: 1) *how can we blend different artistic artefacts for the autonomous construction of a playable experience?* 2) *how can we orchestrate this content towards a common goal?* 3) *does the player experience match with the intended goal of the content generated?*

In order to answer the first question this thesis proposed a new type of procedural content generation system, one that combines the capabilities of search based algorithms, allowing generators to construct and orchestrate content towards a specific intended theme or progression; and the capabilities of machine learned prediction, autonomously informing the generator about key characteristics such as theme or the emotional impact of different artistic assets. The proof-of-concept system called *Sonancia* was developed for the realization of this thesis. It showcases the proposed methodology on two diverging facets: Level Architecture and Audio. Given the ambitiousness of such a system, these two facets were chosen to simplify the problem, while still offering some practical solutions on multi-faceted generation. For the autonomous construction of level architecture an evolutionary algorithm was developed for this thesis capable of evolving levels towards a specific goal. The level generator was extensively tested through different parametric experiments, which allowed us to observe the flexibility and limitations of the proposed method. For the construction of audio affect models a user data collection experiment was first conducted, where the perceived emotion of several audio pieces were annotated by human participants. This data was subsequently used to train different machine learning models capable of ranking audio pieces based on the tension, arousal and valence affect dimensions.

The horror genre was chosen as it provides an interesting case study for both question 2 and 3. This genre in particular relies heavily on audio (Ekman and Lankoski, 2009), and an archetypical narrative progression of suspense. Typically the genre of horror follows an intended progression of suspense, emotionally leading the audience towards anxious and frightful situations. Thus to answer question 2, this progression serves as the common goal between the diverging facets, where the generator attempts to emulate a defined experience through the construction and orchestration of different faceted content.

This genre in particular also preys on strong fearful human emotions, allowing us to measure human player physiology such as stress, anxiety and tension (Holmgard et al., 2013), and then subsequently compare it to the intended experience of the generated level. Thus in order to answer question 3, a user study was conducted on two diverging *Sonancia* level architectures, whose soundscapes were additionally sonified by three different audio orchestration methods. During play human participant physiology and gameplay annotations were collected, in order to analyse the players physiological and perceived gameplay experience against the intended progression. Furthermore, to investigate the relationship between physiological and the participant annotations, an additional study was conducted in order to analyse the relationship between both data types.

8.1 Contributions

This section summarizes the different contributions of this thesis in the advancement of the field of procedural content generation. The main contributions of this thesis also extend into the area of audio affect modelling. Although the models constructed in this thesis were used exclusively for the construction of digital game content, the solutions presented can easily be extended to other multimedia domains such as film, audio production tools, or even music creation. Finally, contributions have also been made in the area of affective computing, where a new methodology of affect annotation was explored. The following list details the main contributions of this thesis:

- The proposal of a multi-faceted procedural content generator pipeline by combining different artistic artefacts through a hybrid system that combines search-based PCG algorithm and preference learning algorithms capable of predicting the affect of audio.
- The development of a fully playable multi-faceted generator (*Sonancia*), capable of generating levels and orchestrating sounds based on a designer defined guideline of gameplay.
- The construction and experimentation of two diverging preference learning algorithms capable of learning-to-rank the low-level features of audio against the affective dimensions of tension, arousal and valence.
- The proposal and construction of affect models capable of ranking the same audio artefact influenced by several digital signal processing effects, in order to analyse how their perceived affective state varies.
- An extensive user study for the evaluation of our proposed methodology was conducted. Each participant's physiology and annotation of perceived affect was compared against the intended experience, with the intent on investigating how our proposed methodology was able to closely resemble the defined gameplay experience.
- The user study also explored a new method of annotating affect in real-time. We introduced an annotation tool which was subsequently analysed against the actual physiological data of each participant. This experiment allowed us to investigate how close the annotation labels approximated to the ground truth of affect (i.e. arousal).

8.2 Limitations

This section describes the different limitations and drawbacks of the proposed methodologies. One of the principal difficulties of this work is formalizing intent for a combination of different digital game domain types, such as formalizing the specific theme or gameplay experience of a genre outside of horror. This problem also stems from the fact that level construction is based on predefined values, which ideally would be derived through extensive user experimentation. Furthermore, the different types of affective models can be difficult to construct, as training tends to require a high volume of data for accurate predictions.

8.2.1 Limitations of Multi-Faceted Procedural Content Generation

Visual Themes and Inter-Domain Level Generation

While the proposed level generation system presented robustness and a degree of flexibility, it is difficult to argue against its ability to present the same effectiveness within other domains or formalizations. Although the narrative progression can be more easily transcribed into other genres through the exploitation of its tropes, similarly to what was done in this thesis, it is potentially more complicated when the level generator attempts to further take into account a theme with distinctive visuals. This is particularly important for the concept of world building, where most games tend to follow a similar theme throughout a game (e.g. fantasy, cyberpunk, sci-fi). Although it can be tempting to suggest that all these limitations could be eliminated through the application of machine learning algorithms – where the problem of computer vision has been extensively investigated – such a approach cannot be considered trivial. The core difficulty in such a system, is the reliance on high volumes of human annotated data and resources which can significantly complicate their construction. However, if constrained to a particular domain, such as proposed in this thesis, the possibility of applying both visual themes and narrative progressions for the construction of multi-faceted levels becomes a more realistic goal.

Adding Complexity to Level Generation

The current implementation of level generation focuses exclusively on the layout and object placement. However, games often present more complicated activities, such as solving a puzzle or finding a particular item, instigating the player to explore the level completely. Thus, for such levels the formalization defined in this thesis could be considered insufficient as it focuses exclusively on a direct progression from start to finish, ignoring backtracking.

One potential solution is to extend the designer formalization so that it includes a way to define the wanted characteristics, e.g. such as the number of puzzle. Another potential solution is for the system to derive a progression with backtracking from the defined tension curve, where the generator forces players to backtrack by using a lock-and-key method, for example. However, such a system would require a new genetic representation and fitness function so as to take into account the different puzzles and backtracking. Interestingly, the first method suggested could be defined as a more designer controlled formalization, whereas the latter as a system interpretive formalization, where the generator makes autonomous decisions based on the designer defined progression.

Predefined Values of Tension Progression

Another limitation consists of the predefined values of fitness in the proposed implementation in this thesis. Although these values are derived and simplified from the literature, ideally inferences such as tension increase and decay would be obtained through extensive user analysis, and potentially even learned. It is important to note that the annotation method proposed within this thesis in addition to the user data collected could potentially be used to construct models capable of learning such progressions.

8.2.2 Limitations of Audio Affect Models

Preference Ambiguity in Annotation Data

One of the biggest difficulties of constructing machine learned models of affect is the reliance on participant annotation. For preference learning in particular the amount of annotations required increases substantially depending on the number of items within the corpus. Although this thesis was successful in obtaining substantial data from the crowdsourcing experiment realized, the data obtained did not cover the substantial amount of pairs required. Furthermore, given the ambiguous nature of emotional perception within audio complicates the problem further as conflicting pairings will often arise within the data. Thus, such models will always present a substantial dependency on the volume of user annotated data in order to reduce the model ambiguity and theoretically improve the prediction accuracy.

Contextualizing Audio Soundscapes

It is often said that context is everything, and the emotional perception of audio is no different. Audio played in varying contexts might suggest very diverse meanings, due to culture, symbolism or past perceptions (Fahlenbrach, 2008). Given that digital games are multi-faceted experiences where players are often put in different contextual locations, there is a potential discrepancy between the predicted affect of the audio model, and the actual affect with the additional contextual layer. Due to the models being trained devoid of context – as participants simply listened and compared between two pieces of sound, the actual emotion might deviate substantially from the in-game felt emotion. As such, models intended for multi-faceted experiences could provide stronger predictions if context information is included in addition to audio low-level descriptors.

Domain Focused – The Horror Genre

The models constructed for this thesis were exclusively trained on an audio library of horror soundscapes. Thus, it is safe to assume that a strong bias exists towards the genre of horror. Although this limitation was beneficial within the context of this thesis, we are uncertain about its application beyond this particular genre. It is important to state however, that the methodologies for model construction presented in this thesis are in fact applicable to other genres and different types of audio libraries as well.

8.2.3 Limitations of The *Sonancia* System

Towards an Experience-Driven Approach

In this thesis we proposed an offline solution, where levels adapt to an a priori defined formalization of designer intent. However, one of the limitations of our proposed solution

is that intent is specified solely in accordance to the designer, but not personalized to the players' actual experience. More precisely, although the intended player experience is defined and levels adapt to it, each individual player has her own perceptions and reactions to what is considered scary or tense. Although, the focus of this work was not specifically tied to experience-driven procedural content generation (Yannakakis and Togelius, 2011), such a system could prove a better and more accurate representation of the player experience. In fact, a lot of the methods proposed in this thesis could be applied for the construction of experience-driven PCG systems, such as audio orchestration using our proposed models, or keeping track of the players affective state and comparing it to the intended experience, thus influencing the gameplay environment such that the distance between both measures is minimized.

Improving the Orchestration of Audio

For the application of sonification and orchestration of audio this thesis opted for a rule-based system, where sounds are placed exclusively to one room based on its tension value, given by the level generator, and the audio's ranking, labelled by the audio model predictors. Although this particular system was capable of obtaining good results, as showcased in Chapter 7, it becomes less viable when the audio complexity increases. More precisely, when audio becomes diegetic and is perceived within the virtual world. Such as suggested in the work of Ekman (2005); Ekman and Kajastila (2009), the perception of sound and its location within the virtual environment can influence the emotional impact of audio. Thus the orchestration of sound can in fact become even more complex, by allowing the system to orchestrate between both non-diegetic and diegetic sounds, and the perceived direction of diegetic audio. The latter would require audio assets to be placed within different points of the environment in order for players to perceive the sound cues at different angles. This could be achieved through a similar methodology as the one described by Tremblay and Verbrugge (2015). In this work the diverging paths of a level are extracted through an A* algorithm, allowing the system to predict the most favourable locations for environmental asset placement. Such a system could similarly be extended to the usage of audio placement.

User Evaluation and Annotation Limitations

Due to the extensive number of variables that emerge during real-time gameplay, it is difficult to derive what exactly influenced a player's affective state. Even though experiments can be extensively controlled and parametrized forcing each participant to a specific path, this does not guarantee that the participant will act accordingly to what is intended. This is simply due to the interactive aspect of digital games. Unlike films or audio, players can backtrack, hide, stay still or even die to the monster within the game. This particularity accounts for an extensive amount of variables that can substantially alter the experience between each participant. Due to this limitation it was difficult to directly compare between the physiology and the annotations of two different level playthroughs. Furthermore, a more robust skin conductance sensor could have given more insight and better results, as the one utilised within this thesis was unable to effectively capture the majority of participants' skin conductance. Lastly, while the annotation system showcased capabilities of approximating (linearly at least) the ground truth, annotations still presented a lot of ambiguity and to some extent annotation fatigue. It is reasonable to argue that an entire gameplay session, followed by an annotation period of the same length can become monotonous after

3 different playthroughs.

8.3 Extensibility

This section reviews additional domains that could potentially benefit from the intertwining systems specifically developed for the realization of this thesis.

8.3.1 Extending Multi-Faceted Procedural Content Generation

Developing Mixed-Initiative Solutions

Mixed-Initiative systems consist of proactive helper tools capable of suggesting alternative solutions based on the designer's original creation and several defined parameters, which consist of the desired characteristics of the final artefact. Systems such as the Sentient Sketchbook by Liapis et al. (2013b), explores this concept for the development of real-time strategy maps. In that work parameters consist of map balancing values, such as the number of player bases and resource nodes, providing the system an optimization goal which complies with the designers needs. The level generation process suggested in this thesis could be beneficial for such a mixed-initiative PCG system, where a designer defines the intended progression and obtains several suggestions of levels with that progression realized. Designers could then subsequently apply these suggestions, or continue with their design process while the system continues to suggest new alternatives built from the designers current version of the map.

Improving Audio Soundscape Development Tools

The capabilities of game development tools have significantly improved over the last 5 years, with many of these tools being widely available for free. Game engines such as *Unity* (Unity Technologies, 2005) and the *Unreal Engine* (Epic Games, 1998) greatly facilitate the development and production of modern digital games. These engines provide a diverging set of toolboxes that allow for the easy integration of graphics, animations, levels and game logic. The majority of engines also offer robust audio software methods assisting audio designers in meticulously placing their hand-crafted sounds directly onto a constructed level (Stevens and Raybould, 2013), while controlling the triggers and other parameters that define how sound should react to players in-game. This process can be quite tedious, particularly for large levels with long progressions containing a high volume of different audio cues. Orchestration capabilities in combination with music information retrieval techniques such as the ones suggested in this thesis, could potentially be used to aid audio designers in populating audio within a specific level. Furthermore, the system could suggest a list of relevant audio assets to pick based on the defined parameters of the audio designer, e.g. "I need a tense sound", "I need a calming sound". Given that audio designers tend to work with large audio libraries, a tool that could filter audio based on the designers exact requirements could be beneficial.

8.3.2 Extending Audio Affect Models

Music Emotion Retrieval Systems

The construction of models capable of ranking audio pieces based on perceived emotion can be a valuable tool outside of the digital game space. One notable example is the autonomous categorization of large audio libraries based on their perceived affect. It is quite common for multimedia industries such as film and even digital games to utilise large audio repositories containing thousands of audio pieces. The autonomous cataloguing of such content could prove to be extremely valuable to audio designers whose job is to convey specific emotions through sound. In modern systems this cataloguing is often done by hand, where the sounds creator manually describes the general concept of the sound for later use. While the latter is particularly helpful for designers who wish to find a specific sound, such as that of an object, the more synthesized and mixed sonorities could be suggested through the prior system, by providing a short list of sounds and how they rank based on the different affective dimensions.

Artistic Music Production Tools

Within this thesis initial investigations were conducted on models capable of ranking different digital signal processing effects based on how they influence the perceived audio affect. Such a model could provide viable solutions for audio production systems. Trained models can suggest the ideal parametrization of different audio pieces so that it increases or decreases a particular affective state. A model could potentially be trained to effectively detect clipping that arises from faulty effect parametrizations. Models could even suggest different parametric adjustments and types of effects so that the tone of a particular instrument can be adjusted to convey a particular emotion. Although the latter models are significantly more complex to construct and require a larger volume of data, such a system could provide powerful tools for musicians with a large degree of flexibility.

8.3.3 Extending The *Sonancia* System

Real-Time Annotations Beyond Digital Games

The real-time annotation system presented in this thesis could easily be extended towards other domains besides digital games. Particularly the majority of time-based media could utilise the proposed method, where participants can be tasked of annotating video, film, or audio artefacts within different genres, for example. Thus, its application could be particularly relevant as an alternative to the traditional questionnaire method within the field of affective computing.

Experience-Driven Audio Orchestration

Yannakakis and Togelius (2011) defines experience-drive procedural content generation, as the process that is capable of personalizing generated digital game content towards a player's affective state. The horror game *Nevermind* (Flying Mollusk, 2014) in particular was quite successful in providing stress-based gameplay adaptation. Thus, it is safe to assume that models trained within this thesis could potentially be adapted to such systems, where audio can be dynamically selected and placed within a virtual world based on either the tension,

arousal or valence affect of a player. Similarly to how this thesis placed audio, an experience-driven PCG system could orchestrate audio in such a way that it calms or stresses a player. Furthermore, such models could be extended to include diegetic-type sounds, as previously suggested. This however would require a more complex orchestration method from the one proposed. Although triggers would still be used, the positioning of audio sources would greatly differ as the direction where sound is emanated could potentially be used as another layer for influencing affect. A particular example can be to place sounds in the players line-of-sight, behind the player or even in adjacent rooms.

Generalizing *Sonancia* Across Game Genres

Given the focus of this thesis on the survival horror genre from the level structures to the audio library used, it is important to note that some methods explored in this thesis could potentially be applicable towards other digital game genres. Heavy narrative based single-player games could use similar level framing and sonification methods for the construction of multi-faceted content, although in certain genres the reliance on one single affective dimension might not be feasible (e.g. action games) and thus might require re-framing content towards different emotional progression types. Notable examples such as the Doom (id Software, 1993–2016) or Call of Duty (Activision, 2003–2017) series, could use framing to inform level generation on the ideal placement of enemy encounters and relaxing moments (i.e. unwinding after intense action sequences). Sonification could act as a form of foreshadowing encounters, and add flavour to sequences that are more exploratory and less action packed. Another notable example could consist of framing the progressions of dungeons in Massive Multiplayer Online (MMO) or "Rogue" style games, allowing designers to control certain aspects of the progression, while retaining the procedural component of the dungeons. It is important to note that the approach proposed in this thesis may not be applicable for some types of game genres, such as genres that do not specifically rely on level progressions (i.e. fighting games).

8.4 Summary

This thesis proposed a pipeline for the autonomous construction of multi-faceted content, by combining the capabilities of evolutionary computation and preference learning. In order to test our proposed methodology, a system was built with the capabilities of blending both level architecture and audio. Although the system proved capable of constructing thematically focused horror levels, a number of limitations do exist in the proposed methodology. Primarily different orchestration methods would need to be developed as the tension progressions were derived specifically from the horror genre, thus such orchestrations might differ for other genres where emotional engagement diverges from those found in horror. Furthermore, the construction of models capable of ranking all the different facets of digital games could prove to be substantially difficult, specifically due to the abundant data required for training such models. The promising results obtained in this thesis suggest that the procedural generation of two different domains (levels and sounds) is possible and can be successful; however a general solution capable of integrating all possible facets in all different genres is potentially unrealistic to achieve with the current implementation. Apart from these limitations, several extensions from the various components of the *Sonancia* system were suggested, in particular within the domain of mixed-initiative co-creative systems, music information retrieval and experience-driven procedural content generation.

Bibliography

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015. — Cited on page 20.

John L Andreassi. *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2013. — Cited on page 100.

Dominik R Bach and Karl J Friston. Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, 50(1):15–22, 2013. — Cited on pages 110 and 132.

Laura-Lee Balkwill and William Forde Thompson. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception: An Interdisciplinary Journal*, 17(1):43–64, 1999. — Cited on page 19.

Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010. — Cited on pages 109, 110, and 132.

Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995. — Cited on page 27.

Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012. — Cited on pages 110 and 132.

Cameron Browne. *Automatic generation and evaluation of recombination games*. PhD thesis, Queensland University of Technology, 2008. — Cited on page 16.

AE Bryson and Yu-Chi Ho. Applied optimal control. *Blaisdell, Waltham, Mass*, 8:72, 1969. — Cited on page 29.

Luigi Cardamone, Georgios N Yannakakis, Julian Togelius, and Pier Luca Lanzi. Evolving interesting maps for a first person shooter. In *Applications of Evolutionary Computation*, pages 63–72. Springer, 2011. — Cited on pages 2 and 13.

Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, pages 13–17. ACM, 2008. — Cited on page 16.

Yun-Gyung Cheong and R Michael Young. Narrative generation for suspense: Modeling and evaluation. In *Interactive Storytelling*, pages 144–155. Springer, 2008. — Cited on pages 4 and 41.

Andrea Clerico, Cindy Chamberland, Mark Parent, Pierre-Emmanuel Michon, Sebastien Tremblay, Tiago Falk, Jean-Christophe Gagnon, and Philip Jackson. Biometrics and classifier fusion to predict the fun-factor in video gaming. In *IEEE Conference on Computational Intelligence and Games*, pages 233–240. IEEE, 2016. — Cited on pages 100 and 101.

Karen Collins. *Playing with sound: a theory of interacting with sound and music in video games*. MIT Press, 2013. — Cited on pages 1, 2, 10, 19, and 67.

S. Colton. Creativity vs. the perception of creativity in computational systems. In *Papers from the AAAI Spring Symposium on Creative Intelligent Systems*, 2008. — Cited on page 47.

Simon Colton. The painting fool: Stories from building an automated painter. In *Computers and creativity*, pages 3–38. Springer, 2012. — Cited on pages 2 and 18.

Simon Colton, John Charnley, and Alison Pease. Computational creativity theory: The face and idea descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*, 2011. — Cited on pages 2, 37, and 47.

Simon Colton, Goodwin Jacob, and Tony Veale. Full-face poetry generation. In *Proceedings of the International Conference on Computational Creativity*, 2012. — Cited on page 18.

Michael Cook and Simon Colton. A rogue dream: Automatically generating meaningful content for games. In *Proceedings of the AIIDE workshop on Experimental AI & Games*, 2014. — Cited on page 18.

Michael Cook, Simon Colton, and Alison Pease. Aesthetic considerations for automated platformer design. In *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment conference*, 2012. — Cited on pages 2 and 18.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995. — Cited on page 25.

Roddy Cowie and Martin Sawey. Gtrace - general trace program from queens, belfast, 2011. — Cited on page 133.

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000. — Cited on page 133.

Mihaly Csikszentmihalyi. Toward a psychology of optimal experience. In *Flow and the foundations of positive psychology*, pages 209–226. Springer, 2014. — Cited on page 17.

Ian Daly, Etienne B Roesch, James Weaver, and Slawomir J Nasuto. Machine learning to identify neural correlates of music and emotions. In *Guide to Brain-Computer Music Interfacing*, pages 89–103. Springer, 2014. — Cited on page 19.

Joris Dormans. Adventures in level design: generating missions and spaces for action adventure games. In *Proceedings of the 2010 workshop on procedural content generation in games*, page 1. ACM, 2010. — Cited on page 15.

Joris Dormans and Stefan Leijnen. Combinatorial and exploratory creativity in procedural content generation. In *Proceedings of the 4th International Workshop on Procedural Content Generation in Games*, 2013. — Cited on page 15.

Anders Drachen, Lennart E Nacke, Georgios Yannakakis, and Anja Lee Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, pages 49–54. ACM, 2010. — Cited on page 16.

Tuomas 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*, 40(4):349–366, 2011. — Cited on page 20.

Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 2010. — Cited on pages 19 and 20.

Arne Eigenfeldt and Philippe Pasquier. Evolving structures for electronic dance music. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, pages 319–326. ACM, 2013. — Cited on page 18.

Arne Eigenfeldt, Adam Burnett, and Philippe Pasquier. Evaluating musical metacreation in a live performance context. In *Proceedings of the Third International Conference on Computational Creativity*, pages 140–144, 2012. — Cited on page 18.

Inger Ekman. Meaningful noise: Understanding sound effects in computer games. *Proceedings of the Digital Arts and Cultures*, 2005. — Cited on pages 12, 13, and 139.

Inger Ekman and Raine Kajastila. Localization cues affect emotional judgments–results from a user study on scary sound. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009. — Cited on pages 3, 12, 13, 17, and 139.

Inger Ekman and Petri 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*, pages 181–199, 2009. — Cited on pages 1, 2, 3, 11, 12, 19, 35, 37, 47, and 135.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. — Cited on page 19.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3): 572–587, 2011. — Cited on page 74.

Florian Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, 2016. — Cited on pages 72 and 73.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013. — Cited on page 72.

Kathrin Fahlenbrach. Emotions in sound: Audiovisual metaphors in the sound design of narrative films. *Projections: The Journal for Movies and Mind*, 2(2):85–103, 2008. — Cited on pages 1, 11, 12, and 138.

Vincent E Farrugia, Héctor P Martínez, and Georgios N Yannakakis. The preference learning toolbox. 2015. — Cited on page 25.

Gilles Fauconnier and Mark Turner. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books, 2008. — Cited on page 18.

Johannes Fürnkranz and Eyke Hüllermeier. *Preference learning*. Springer, 2011. — Cited on pages 21 and 25.

Alf Gabrielsson and Patrik N Juslin. Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of music*, 24(1):68–91, 1996. — Cited on page 19.

Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica e3a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 39–42. IEEE, 2014. — Cited on page 101.

Tom Garner and Mark Grimshaw. A climate of fear: considerations for designing a virtual acoustic ecology of fear. In *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, pages 31–38. ACM, 2011. — Cited on page 12.

Tom Garner and Mark Grimshaw. Sonic virtuality: Understanding audio in a virtual world. *The Oxford Handbook of Virtuality*, page 364, 2014. — Cited on pages 10 and 13.

Tom Garner, Mark Grimshaw, and Debbie Abdel 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*, page 10. ACM, 2010. — Cited on pages 12, 19, 69, 77, 78, 88, and 95.

Hans-Peter Gasselseder. Re-scoring the games score: Dynamic music and immersion in the ludonarrative. *Proceedings of the Intelligent Human Computer Interaction conference*, pages 1–8, 2014. — Cited on pages 1 and 10.

Jeremy Gow and Joseph Corneli. Towards generating novel games using conceptual blending. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015. — Cited on page 19.

Kazjon Grace, John Gero, and Rob Saunders. Representational affordances and creativity in association-based systems. In *Proceedings of the International Conference on Computational Creativity*, 2012. — Cited on page 47.

Erin Jonathan Hastings, Ratan K Guha, and Kenneth O Stanley. Automatic content generation in the galactic arms race video game. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(4):245–263, 2009. — Cited on page 16.

Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6 (2):156–166, 2005. — Cited on page 100.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 97–102. IET, 1999. — Cited on page 26.

Javier Hernandez, Rob R Morris, and Rosalind W Picard. Call center stress recognition with person-specific models. In *International Conference on Affective Computing and Intelligent Interaction*, pages 125–134. Springer, 2011. — Cited on page 100.

John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975. — Cited on page 30.

Christoffer Holmgard, Georgios N Yannakakis, Karen-Inge Karstoft, and Henrik Steen Andersen. Stress detection for ptsd via the startlemart game. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 523–528. IEEE, 2013. — Cited on pages 16, 132, and 136.

Christoffer Holmgård, Georgios N Yannakakis, Héctor P Martínez, Karen-Inge Karstoft, and Henrik Steen Andersen. Multimodal ptsd characterization via the startlemart game. *Journal on Multimodal User Interfaces*, 9(1):3–15, 2015. — Cited on pages 98, 99, 100, and 110.

Vincent Hom and Joe Marks. Automatic design of balanced board games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 25–30, 2007. — Cited on pages 15 and 16.

A. K. Hoover, P. A. Szerlip, M. E. Norton, T. A. Brindle, Z. Merritt, and K. O. Stanley. Generating a complete multipart musical composition from a single monophonic melody with functional scaffolding. In *Proceedings of the International Conference on Computational Creativity*, 2012. — Cited on page 18.

Amy K Hoover, William Cachia, Antonios Liapis, and Georgios N Yannakakis. AudioInSpace: Exploring the creative fusion of generative audio, visuals and gameplay. In *Proceedings of the EvoMusArt conference*, pages 101–112. Springer, 2015. — Cited on pages 14 and 19.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. — Cited on page 27.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002. — Cited on pages 21 and 26.

Daniel Johnson and Dan Ventura. Musical motif discovery in non-musical media. In *Proceedings of the International Conference on Computational Creativity*, 2014. — Cited on page 18.

Lawrence Johnson, Georgios N Yannakakis, and Julian Togelius. Cellular automata for real-time generation of infinite cave levels. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*, page 10. ACM, 2010. — Cited on page 15.

Rilla Khaled and Georgios N Yannakakis. Village voices: An adaptive game for conflict resolution. In *Proceedings of the 8th Conference on the Foundations of Digital Games*, pages 425–426, 2013. — Cited on page 98.

Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 255–266, 2010. — Cited on page 20.

Daniel Kromand. Sound and the diegesis in survival-horror games. *Proceedings of the 3rd Audio Mostly Conference*, 2008. — Cited on page 12.

Anna Krzeczkowska, Jad El-Hage, Simon Colton, and Stephen Clark. Automated collage generation – with intent. In *Proceedings of the International Conference on Computational Creativity*, 2010. — Cited on page 18.

Boyang Li, Alexander Zook, Nicholas Davis, and Mark O Riedl. Goal-driven conceptual blending: A computational approach for creativity. In *ICCC*, volume 10, 2012. — Cited on page 18.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. Story generation with crowdsourced plot graphs. In *The AAAI Conference on Artificial Intelligence*, 2013. — Cited on pages 21 and 70.

Antonios Liapis. 3.5 creativity facet orchestration: the whys and the hows. *Artificial and Computational Intelligence in Games: Integration*, page 217, 2014. — Cited on page 18.

Antonios Liapis, Georgios N. Yannakakis, and Julian Togelius. Generating map sketches for strategy games. In *Applications of Evolutionary Computation*, pages 264–273. Springer Berlin Heidelberg, 2013a. — Cited on pages 2, 13, and 15.

Antonios Liapis, Georgios N Yannakakis, and Julian Togelius. Sentient sketchbook: Computer-aided game level authoring. In *Proceedings of the 8th Conference on the Foundations of Digital Games*, pages 213–220, 2013b. — Cited on pages 15 and 140.

Antonios Liapis, Georgios N Yannakakis, and Julian Togelius. Computational game creativity. In *Proceedings of the International Conference on Computational Creativity*, pages 285–292, 2014. — Cited on pages 2, 18, and 47.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. — Cited on pages 20 and 100.

Phil Lopes and Georgios N Yannakakis. Investigating collaborative creativity via machine-mediated game blending. In *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment conference*, 2014. — Cited on page 18.

Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. The c2create authoring tool: Fostering creativity via game asset creation. In *Proceedings of the Conference on Computational Intelligence and Games*, pages 1–2. IEEE, 2014. — Cited on page 98.

Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. Sonancia: Sonification of procedurally generated game levels. In *Proceedings of the 1st Computational Creativity and Games Workshop*, 2015. — Cited on page 12.

Reza Lotfian and Carlos Busso. Practical considerations on the use of preference learning for ranking emotional speech. 2016. — Cited on page 21.

O.L. Mangasarian. *Nonlinear Programming*. McGraw-Hill, 1969. — Cited on page 26.

Héctor P Martınez. Advancing affect modeling via preference learning and unsupervised feature extraction. 2013. — Cited on page 29.

Hector P Martinez, Georgios N Yannakakis, and John Hallam. Dont classify ratings of affect; rank them! *IEEE Transactions on Affective Computing*, 5(3):314–326, 2014. — Cited on pages 20, 21, 25, and 70.

Héctor Perez Martínez, Maurizio Garbarino, and Georgios N Yannakakis. Generic physiological features as predictors of player experience. In *International Conference on Affective Computing and Intelligent Interaction*, pages 267–276. Springer, 2011. — Cited on pages 21, 30, 68, and 99.

Martin F McKinney, Jeroen Breebaart, et al. Features for audio and music classification. In *The International Society on Music Information Retrieval Conference*, volume 3, pages 151–158, 2003. — Cited on page 72.

Zbigniew Michalewicz. A survey of constraint handling techniques in evolutionary computation methods. In *Proceedings of the 4th Annual Conference on Evolutionary Programming*, pages 135–155. 1995. — Cited on page 63.

Eduardo Reck Miranda and Julien Castet. *Guide to Brain-Computer Music Interfacing*. Springer, 2014. — Cited on pages 79 and 87.

Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998. — Cited on page 30.

Tom Mitchell. *Machine Learning*. McGraw Hill, 1997. — Cited on page 24.

Pedro A Nogueira, Vasco Torres, Rui Rodrigues, Eugénio Oliveira, and Lennart E Nacke. Vanishing scares: biofeedback modulation of affective player experiences in a procedural horror game. *Journal on Multimodal User Interfaces*, 10(1):31–62, 2016. — Cited on page 17.

Pedro Alves Nogueira, Rúben Aguiar, Rui Amaral Rodrigues, Eugénio C Oliveira, and Lennart Nacke. Fuzzy affective player models: A physiology-based hierarchical clustering method. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2014. — Cited on page 17.

Chris Pedersen, Julian Togelius, and Georgios N Yannakakis. Modeling player experience in super mario bros. In *IEEE Symposium on Computational Intelligence and Games*, pages 132–139. IEEE, 2009. — Cited on page 126.

Christopher Pedersen, Julian Togelius, and Georgios N Yannakakis. Modeling player experience for content creation. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(1):54–67, 2010. — Cited on pages 30 and 126.

Geoffroy Peeters. A large set of audio features for sound description (similarity and classification). *UIDADO I.S.T. Project Report*, 2004. — Cited on page 72.

Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985. — Cited on page 15.

Bernard Perron. Sign of a threat: The effects of warning systems in survival horror games. In *Proceedings of COSIGN*, pages 132–141, 2004. — Cited on pages 3 and 12.

Bernard Perron. *Horror video games: Essays on the fusion of fear and play.* McFarland, 2009. — Cited on pages 2, 3, 4, 11, and 12.

David Plans and Davide Morelli. Experience-driven procedural music generation for games. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(3):192–198, 2012. — Cited on page 17.

Elise Plans, Davide Morelli, and David Plans. Audionode: prototypical affective modelling in experience-driven procedural music generation. In *Proceedings of the First Computational Creativity and Games Worshop*, 2015. — Cited on page 17.

Pramila Rani, Nilanjan Sarkar, and Changchun Liu. Maintaining optimal challenge in computer games through real-time physiological feedback. In *Proceedings of the 11th international conference on human computer interaction*, volume 58, 2005. — Cited on page 16.

Graeme Ritchie. Current directions in computational humour. *Artificial Intelligence Review*, 16(2):119–135, 2001. — Cited on page 18.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. — Cited on pages 17, 19, and 21.

Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice Hall, 2003. — Cited on page 27.

Pasi Saari, Gyorgy Fazekas, Tuomas Eerola, Mathieu Barthet, Olivier Lartillot, and Mark Sandler. Genre-adaptive semantic computing and audio-based modelling for music mood annotation. *IEEE Transactions on Affective Computing*, 7(2):122–165, 2016. — Cited on page 20.

Ulrich Schimmack and Alexander Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000. — Cited on pages 19, 20, 25, 42, and 96.

Erik M Schmidt and Youngmoo E Kim. Learning emotion-based acoustic features with deep belief networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 65–68. IEEE, 2011. — Cited on page 97.

Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 312–315, 2009. — Cited on page 72.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011. — Cited on page 72.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012. — Cited on page 72.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013. — Cited on page 72.

Marco Scirea, Yun-Gyung Cheong, Mark J Nelson, and Byung-Chull Bae. Evaluating musical foreshadowing of videogame narrative experiences. In *Proceedings of the Audio Mostly: A Conference on Interaction With Sound*. ACM, 2014. — Cited on pages 14 and 17.

Marco Scirea, Gabriella A. B. Barros, Noor Shaker, and Julian Togelius. SMUG: Scientific music generator. In *Proceedings of the International Conference on Computational Creativity*, 2015. — Cited on page 18.

Stefania Serafin and Giovanni Serafin. Sound design to enhance presence in photorealistic virtual reality. In *Proceedings of the International Conference on Auditory Display*, 2004. — Cited on page 1.

Noor Shaker, Miguel Nicolau, Georgios N Yannakakis, Julian Togelius, and Michael O'Neill. Evolving levels for super mario bros using grammatical evolution. In *Conference on Computational Intelligence and Games*, pages 304–311. IEEE, 2012. — Cited on pages 2, 13, and 15.

Noor Shaker, Georgios Yannakakis, and Julian Togelius. Crowd-sourcing the aesthetics of platform games. *IEEE Transactions on Computational Intelligence and AI in Games*, 5 (3), 2013. — Cited on pages 21 and 70.

Noor Shaker, Julian Togelius, and Mark J. Nelson. *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*. Springer, 2015. — Cited on pages 1, 2, 3, and 13.

Noor Shaker, Antonios Liapis, Julian Togelius, Ricardo Lopes, and Rafael Bidarra. Constructive generation methods for dungeons and levels. In *Procedural Content Generation in Games*, pages 31–55. Springer, 2016a. — Cited on page 15.

Noor Shaker, Julian Togelius, and Mark J Nelson. Fractals, noise and agents with applications to landscapes. In *Procedural Content Generation in Games*, pages 57–72. Springer, 2016b. — Cited on page 15.

David Sonnenschein. *Sound design*. Michael Wiese Productions, 2001. — Cited on page 1.

Nathan Sorenson, Philippe Pasquier, and Steve DiPaola. A generic approach to challenge modeling for the procedural creation of video game levels. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(3):229–244, 2011. — Cited on page 17.

Richard Stevens and Dave Raybould. *The Game Audio Tutorial: A Practical Guide to Creating and Implementing Sound and Music for Interactive Games*. Taylor & Francis, 2013. — Cited on pages 1, 23, 36, 69, and 140.

Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1989. — Cited on page 19.

Julian Togelius and Jurgen Schmidhuber. An experiment in automatic game design. In *Computational Intelligence and Games, 2008. CIG'08. IEEE Symposium On*, pages 111–118. IEEE, 2008. — Cited on pages 15 and 16.

Julian Togelius, Renzo De Nardi, and Simon M Lucas. Towards automatic personalised content creation for racing games. In *IEEE Symposium on Computational Intelligence and Games.*, pages 252–259. IEEE, 2007. — Cited on pages 2, 13, and 15.

Julian Togelius, Mike Preuss, Nicola Beume, Simon Wessing, Johan Hagelbäck, and Georgios N Yannakakis. Multiobjective exploration of the starcraft map space. In *IEEE Symposium on Computational Intelligence and Games (CIG)*, pages 265–272. IEEE, 2010a. — Cited on page 15.

Julian Togelius, Mike Preuss, and Georgios N Yannakakis. Towards multiobjective procedural map generation. In *Proceedings of the workshop on procedural content generation in games*, page 3. ACM, 2010b. — Cited on page 15.

Julian Togelius, Georgios N Yannakakis, Kenneth O Stanley, and Cameron Browne. Search-based procedural content generation: A taxonomy and survey. *Proceedings of the Computational Intelligence and Games Conference*, 3(3):172–186, 2011. — Cited on pages v, 1, 2, 13, 14, 15, 24, 35, and 39.

Simone Tognetti, Maurizio Garbarino, Andrea Bonarini, and Matteo Matteucci. Modeling enjoyment preference from physiological responses in a car racing game. In *IEEE Symposium on Computational Intelligence and Games (CIG)*, pages 321–328. IEEE, 2010. — Cited on page 16.

Mike Treanor, Bryan Blackford, Michael Mateas, and Ian Bogost. Game-o-matic: Generating videogames that represent ideas. In *Proceedings of the FDG workshop on Procedural Content Generation in Games*, page 11. ACM, 2012. — Cited on page 18.

Jonathan Tremblay and Clark Verbrugge. An algorithmic approach to decorative content placement. In *The 11th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015. — Cited on page 139.

Konstantinos Trochidis and Emmanuel Bigand. Emotional response during music listening. In *Guide to Brain-Computer Music Interfacing*. Springer, 2014. — Cited on page 19.

Tony Veale. From conceptual mash-ups to bad-ass blends: A robust computational model of conceptual blending. In *Proceedings of the Third International Conference on Computational Creativity*, pages 1–8, 2012. — Cited on page 18.

Tony Veale. Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of the International Conference on Computational Creativity*, pages 152–159, 2013. — Cited on page 18.

D Wilkie. Pictorial representation of kendalls rank correlation coefficient. *Teaching Statistics*, 2(3):76–78, 1980. — Cited on page 78.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. — Cited on page 29.

Yi-Hsuan Yang and Homer H Chen. *Music emotion recognition*. CRC Press, 2011a. — Cited on pages 21 and 84.

Yi-Hsuan Yang and Homer H Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (4):762–774, 2011b. — Cited on pages 20, 21, 25, 79, and 84.

Georgios N Yannakakis and John Hallam. Game and player feature selection for entertainment capture. In *IEEE Symposium on Computational Intelligence and Games*, pages 244–251. IEEE, 2007. — Cited on page 30.

Georgios N Yannakakis and John Hallam. Ranking vs. preference: a comparative study of self-reporting. In *Affective computing and intelligent interaction*, pages 437–446. Springer, 2011. — Cited on pages 20, 25, 74, 77, 100, and 116.

Georgios N Yannakakis and Hector P Martinez. Grounding truth via ordinal annotation. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 574–580. IEEE, 2015. — Cited on pages 25 and 70.

Georgios N Yannakakis and Héctor P Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015. — Cited on pages 20, 21, 25, and 70.

Georgios N Yannakakis and Julian Togelius. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing*, 2(3):147–161, 2011. — Cited on pages 13, 16, 17, 98, 139, and 141.

Georgios N Yannakakis, Manolis Maragoudakis, and John Hallam. Preference learning for cognitive modeling: a case study on entertainment preferences. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 39(6):1165–1175, 2009. — Cited on page 21.

Georgios N Yannakakis, Héctor P Martínez, and Arnav Jhala. Towards affective camera control in games. *User Modeling and User-Adapted Interaction*, 20(4):313–340, 2010. — Cited on pages 16, 21, 99, and 100.

Georgios N Yannakakis, Antonios Liapis, and Constantine Alexopoulos. Mixed-initiative co-creativity. In *Proceedings of the 9th Conference on the Foundations of Digital Games*, 2014. — Cited on page 98.

S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, D Povey, V Valtchev, and P Woodland. The htk-book 3.4. *Cambridge University, Cambridge, England*, 2006. — Cited on page 72.

Marcel Zentner, Didier Grandjean, and Klaus R Scherer. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4):494, 2008. — Cited on page 19.