VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wananga o te Upoko o te Ika a Maui

School of Engineering and Computer Science

Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Internet: office@ecs.vuw.ac.nz

Personalised Prediction of Self-Reported Emotion Responses to Music Stimuli

Kameron Christopher

Supervisor(s): Professor Dale A. Carnegie, Associate Professor Gina M. Grimshaw

A Thesis Submitted to the Victoria University of Wellington in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

2019

Abstract

In this thesis I develop a robust system and method for predicting individuals' emotional responses to musical stimuli. Music has a powerful effect on human emotion, however the factors that create this emotional experience are poorly understood. Some of these factors are characteristics of the music itself, for example musical tempo, mode, harmony, and timbre are known to affect people's emotional responses. However, the same piece of music can produce different emotional responses in different people, so the ability to use music to induce emotion also depends on predicting the effect of individual differences. These individual differences might include factors such as people's moods, personalities, culture, and musical background amongst others. While many of the factors that contribute to emotional experience have been examined, it is understood that the research in this domain is far from both a) identifying and understanding the many factors that affect an individual's emotional response to music, and b) using this understanding of factors to inform the selection of stimuli for emotion induction. This unfortunately results in wide variance in emotion induction results, inability to replicate emotional studies, and the inability to control for variables in research.

The approach of this thesis is to therefore model the latent variable contributions to an individual's emotional experience of music through the application of deep learning and modern recommender system techniques. With each study in this work, I iteratively develop a more reliable and effective system for predicting personalised emotion responses to music, while simultaneously adopting and developing strong and standardised methodology for stimulus selection. The work sees the introduction and validation of a) electronic and loop-based music as reliable stimuli for inducing emotional responses, b) modern recommender systems and deep learning as methods of more reliably predicting

individuals' emotion responses, and c) novel understandings of how musical features map to individuals' emotional responses.

The culmination of this research is the development of a personalised emotion prediction system that can better predict individuals emotional responses to music, and can select musical stimuli that are better catered to individual difference. This will allow researchers and practitioners to both more reliably and effectively a) select music stimuli for emotion induction, and b) induce and manipulate target emotional responses in individuals.

Acknowledgments

I would like to give a tremendous thank you to my supervisors, Professor Dale A. Carnegie and Associate Professor Gina M. Grimshaw. You two have been both excellent advisors and great role models to me throughout my studies. From the first day I met each of you, I knew we would develop something quite special. Dale, on top what you have contributed as an advisor to my Doctoral Thesis, you have made a significant contribution to my abilities to plan, lead, manage, and execute in life. Gina, in addition to all the above, you have also given me the gift of communication, and this has opened opportunities that I would have never imagined. I will carry the lessons you two have taught me everywhere I go in the future.

Thank you to Victoria University of Wellington's School of Engineering and Computer Science, School of Psychology, and Faculty of Graduate Research for providing me with the support, scholarships, and awards that allowed me achieve my life long goal of completing doctoral studies. Thank you, Dr. Will Browne. You have been ever present throughout my doctoral studies and I appreciate the many great and inspiring conversations we have had over the years. Dr. Michael Tooley, thank you also, you have always been very generous with your time, and you embraced me as a member of the CANLAB from day one.

I would also like to thank my colleagues and friends. Diana Siwiak-Dean and Christopher Dean, I could not be more appreciative of your support throughout these years. Thank you Laura Kranz, Rosie Moody, and the entire CANLAB for the support you have given me through my doctoral studies. Thank you as well to Jon He, Dijana Sneath, Mo. Zareei, and the entire SELCT lab for your support in my studies. You have been a true family to me here in New Zealand.

Finally, I would like to thank my family. You saw something you believed in and you helped it to flourish. You have never told me too high or too far. I have also had some tremendous teachers throughout my life who have joined that family: Fernando Pullum, Patsy Payne, Dr. Stephen-Wolf Foster, and Tibor Pusztai. To my sisters, Kandyse, Kristia, and Taneisha, thank you for believing in me and continuing to support me through all of these years. To my parents, Desiree and James, you are beyond words. What you have done for me, my siblings, my friends, and our communities shows a level of love and selflessness that I can only aspire to. All that I have been able to accomplish is because of you.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Music and affect are deeply entwined. Musicians and composers have used music for centuries to both communicate and manipulate emotional states through the organisation of sound in time; evolving deep bonds between musical components and human emotional responses (Cook & Dibben, 2010). As early as the 1930s, psychologists such as Hevner and Gundlach began studying the relationships between music structure and emotion, identifying components such as modality, tempo, and pitch as key contributors to music's affective expressions (Gundlach, 1935; Hevner, 1935, 1936). Music is considered one of the most powerful tools available to researchers who seek to manipulate emotion (Baumgartner, Esslen, & Jäncke, 2006; Juslin & Laukka, 2004; Kenealy, 1988; Zentner, Grandjean, & Scherer, 2008; Zhang, Hui, & Barrett, 2014), but an understanding of the emotional correlates of music also yields benefits in many other applied fields. For example, music therapy, marketing, film, and video game development, all rely on music to achieve affective goals (Juslin & Sloboda, 2013).

However, while it is widely accepted that music can induce emotional responses in people, very little research has been conducted to predict and control for how music affects the emotional responses of individuals. Researchers have repeatedly warned that the failure to account for how individual differences affects emotional responses can lead to (a) inconsistent results and failures to replicate psychological studies (Frieler et al., 2013; Juslin & Västfjäll, 2008), and (b) the inability to systematically control for experimental variables (Juslin & Sloboda, 2011). Some of the factors of individual differences are known; including mechanisms such as:

- peoples' personality (Vuoskoski & Eerola, 2011a, 2011b; Vuoskoski, Thompson, McIlwain, & Eerola, 2012),
- episodic memory (i.e. a piece of music may trigger the memory of a specific event) and social contagion (i.e. an individual's experience of an excerpt may be subject to social influences) (Evans & Schubert, 2008; Juslin & Västfjäll, 2008), and
- a person's musical background including music listening history and preferences (Belcher & Haridakis, 2013; Juslin & Laukka, 2004).

However, many of the factors are still unknown. Of even greater importance, all individual differences affect how a person experiences a piece of music, but they are rarely accounted for in emotion induction procedures.

There are two key areas that researchers must improve on in order to (a) better account for individual differences in emotional responses, and (b) develop more effective and consistent music emotion induction. The first is in developing methods to better predict the emotional responses of specific individuals to any given musical stimulus – until now, researchers have primarily relied on the average responses to stimuli as their selection criteria, if at all (Eerola & Vuoskoski, 2013). The second is in adapting a strong and standardised methodology for stimulus selection. Much of the research in music affect induction has relied on stimulus sets (usually of classical music) that have been selected based on unevaluated or relatively arbitrary assumptions (Frieler et al., 2013; Juslin & Västfjäll, 2008).

The shortcomings of research in this domain to date is understandable, as manually mapping each individual's emotional response to music to individual differences and musical features would be a very challenging and time-consuming task. However, modern recommender system and deep learning techniques are able to learn models of individual differences as latent variables and apply them to create personalised predictions. Thus, the objective of this thesis is to use machine learning techniques to develop a personalised

emotion prediction system that can (1) predict individuals' emotional responses to music, and (2) select musical stimuli that are catered to individual differences. This is a vital contribution to the literature as it will allow researchers and practitioners to more reliably and effectively (a) select music stimuli for emotion induction, and (b) induce and manipulate the target emotional responses in individuals.

1.2 Thesis Specifications

In this thesis, I develop a novel approach to predicting emotional responses to music in individuals. Although an individual's emotional responses are influenced by mechanisms that are not yet well understood in research (Juslin & Västfjäll, 2008), this need not hinder our ability to model and account for these mechanisms' effects. I show that it is possible to use modern machine learning, deep learning, and recommender techniques to predict how latent factors will affect individual emotional responses. These methods leverage the emotion that was reported both (a) in response to other musically similar items, and (b) by cohorts of similar individuals, to more precisely forecast an individual's emotional response to any given music stimulus.

The two primary goals of this thesis were to develop a system that can (1) predict individuals' emotional responses to music, and (2) enable the selection of musical stimuli to be catered to individual differences. Therefore, with each study in this work, I iteratively develop a more reliable and robust system for predicting personalised emotion responses to music, while simultaneously adopting and developing strong and standardised methodology for stimulus selection. Figure 1.1 provides an overview of the thesis progression across chapters, and indicates how each chapter contributes towards achieving each of these two defined research goals.

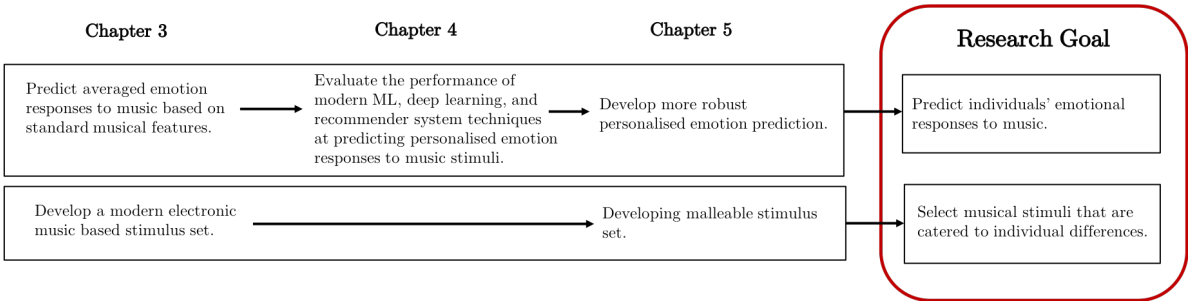

Figure 1.1 Flow diagram showing progressive development of the thesis goals from chapters 3 to 5. For each of the two research goals, each chapter iteratively builds on the research of the previous chapter: (1) the emotion prediction system is iteratively developed to become more reliable and robust in each chapter, and (2) an electronic stimulus set is established and validated in Chapter 3, which forms the groundwork for the development of a malleable electronic loop-based stimulus in Chapter 5.

Ultimately, while it is important to understand what mechanisms contribute to a listener's response, not all mechanisms for every individual-stimulus combination can be thoroughly described, and building a stimulus set to account for all possible combinations would be an exhaustive, time-consuming, and likely over-fitted process for researchers to implement. The modern machine learning, deep learning, and recommender techniques suggested allow for the development of a personalised emotion prediction system that can not only be used to predict individuals' responses, but also select musical stimuli that are catered to a specific individual's characteristics. The proposed personalised emotion prediction system allows researchers and practitioners to control for individual differences (i.e. the underlying factors specific to individual experience), and makes it possible for them to more reliably manipulate emotion in research and applied fields.

1.3 Novel Contributions

Within this thesis, I offer several novel contributions to the field of musical emotion induction research. Namely, this thesis:

1. Validates, for the first time, that modern electronic-based music stimuli can be just as effective in inducing emotion as traditional orchestral stimuli. This allows researchers and practitioners to explore forms of musical stimuli

that are more amenable to manipulation, and will thus more easily facilitate the manipulation of emotional responses in emotion studies (Eerola, Friberg, & Bresin, 2013).

2. Maps how certain musical features affect emotional responses, and as a result explores and validates, for the first time, the use of loops to create musical stimuli that are amenable enough to facilitate personalised manipulation of emotional responses.
3. Provides a strong and robust method for selecting music stimuli for emotion induction in individuals. This is an area where the research has traditionally suffered (Eerola & Vuoskoski, 2013).
4. Introduces novel approaches to predicting individuals' emotional responses to music stimuli. This development allows researchers and practitioners to more precisely control for individual differences and improves their ability to achieve intended emotional outcomes.
5. Introduces deep learning and recommender techniques as a solution to account for the latent variables that affect individuals' emotional responses to music. Researchers are still far from understanding all the factors that contribute to an individual's emotion response to music stimuli, and mapping these factors to specific individuals and their emotional responses would be an exhaustive process. However, the methods introduced in this thesis can address this issue and more precisely predict individuals' emotional responses.

1.4 Thesis Structure

This thesis consists of a comprehensive review of the existing literature, followed by three novel studies. Overall, these studies iteratively build to the final development of

a personalised emotion prediction system for more reliable and precise music emotion induction.

Specifically, in Chapter 2, I provide an overview of the relevant literature to give the reader a background on emotion, the emotional components of music, personalised emotion prediction, and approaches toward developing malleable stimulus sets for more personalised emotion induction. I begin, in Section 2.1, by introducing several psychological models of emotion, and discussing specifically why the 'dimensional' model of emotion is chosen for this thesis. In Section 2.2, I discuss how music has traditionally been understood to affect emotional responses, and the limitations that stem from not accounting for individual differences. In Section 2.3, I introduce literature pertaining to the development of emotion-based personalised music recommender systems, illustrating (a) some advantages of personalised approaches over the traditional averaging approach, and (b) how this research could be further developed to create more rigorous personalised music emotion prediction systems. Furthermore, in Section 2.4, I highlight Affective Algorithmic Composition (AAC) as a potential solution to creating malleable music stimulus sets, the limitations of certain approaches pertaining to AAC, and why I deem the sequence-based AAC approach to be the optimal solution. I conclude Chapter 2 with a discussion of how the literature informs the development of a personalised emotion prediction system for more reliable and precise music emotion induction, and explain what approaches are developed upon further in the thesis.

Chapter 3 looks at the selection of musical stimuli, and explores which musical features may contribute to people's emotional responses to music. In Section 3.1, I establish a method for selecting music stimuli and introducing modern electronic-based music as emotional stimulus. The method consists of (a) asking a committee of music experts to select stimuli which they hypothesise will have the targeted emotional effect on people, and then (b) asking a group of randomly selected people to provide emotional ratings on a Likert-scale for each music excerpt. In Section 3.2, I examine the results of

the ratings task and compare participants' emotional responses between genres, finding that modern electronic-based music can be just as effective at inducing emotion as the more traditional orchestral music. In Section 3.4 and Section 3.5, I examine the musical features that contribute to participants' averaged emotional responses to musical excerpts, and show the features that appear to be generally predictive of emotional responses. I perform factor analysis to link low-level musical features to high-level concepts such as rhythm, tone, and timbre, and use several standard machine learning algorithms to show the features' predictive capabilities.

In Chapter 4, I develop the first iteration of a personalised music emotion prediction system, which is based on the stimulus set and ratings collected in Chapter 3. In Section 4.2, I use clustering to reveal cohorts of individuals with similar and contrasting emotional responses to musical stimuli. I then introduce collaborative and content-based filtering techniques in Section 4.3, and use them to predict individuals' emotional responses to the music stimuli. In Section 4.4, I introduce and evaluate a novel content-based filtering approach based on Convolutional-Recurrent Neural Networks (CB-CRNN). I then compare its ability to predict personalised emotional responses to music with other personalised recommender techniques and non-personalised (based only on averaged ratings for a music excerpt) emotion prediction.

In Chapter 5, I extend the research from Chapter 4, using modern recommender system techniques to better predict personalised emotion responses, and I introduce a loop-based method for developing malleable stimulus sets. With the loop-based stimulus set, the musical voices of stimuli can be changed to induce emotional responses. In Section 5.3, I discuss the process of selecting and preparing a loop-based stimulus set to evaluate whether it could be used to manipulate emotional responses. In Section 5.4, I repeat the techniques from Chapter 4 with more participants and more musical excerpts (i.e. I administer a ratings tasks to 1,943 randomly selected people, asking them to provided emotional ratings to each of 1,307 music excerpts generated from the loop-based stimulus

set). This larger data set of ratings is then used in Section 5.5 to re-evaluate the recommender techniques that were originally introduced in Chapter 4, to confirm that they are able to better predict individuals' emotional responses to music stimuli. Section 5.6 is focused on (a) giving the reader an understanding of which musical features contribute to individuals' emotional responses to music, and (b) demonstrating how those features, represented in the voices of different loops, can be used to manipulate emotional responses.

A summary of the research and my concluding remarks are provided in Chapter 6, delineating my conclusions from this research, my novel contributions to the field, and exploring avenues of investigation for future research. Finally, several appendices are also provided with supplementary information (e.g. a catalogue of the music that was used in various studies, and evidence of ethical approval for this research).

Chapter 2 Background and Literature Review

Emotions are psychophysiological states characterized by an individual's subjective experiences, physiological responses, and responsive and adaptive behaviours (Grimshaw, 2017; Kleinginna Jr & Kleinginna, 1981). Listening to music has a measurable effect on self-reported moods such as happiness, exhilaration, despondency, and sadness (Eerola & Vuoskoski, 2013; Kenealy, 1988; Västfjäll, 2002). The effects of listening to music have also been measured as physiological responses (Baumgartner et al., 2006). Music has long been known to have a marked effect on emotion, and as such, it has been used as a tool to induce emotion in applications across many domains (e.g. psychological research, consumer marketing, music therapy, film, and videogames).

Little research has been developed to explore how specific components (or features) of music affect emotional responses. Additionally, individual differences such as peoples' personalities (Vuoskoski & Eerola, 2011a, 2011b; Vuoskoski et al., 2012), episodic memories and social contagion (Evans & Schubert, 2008; Juslin & Västfjäll, 2008), and musical backgrounds (Belcher & Haridakis, 2013; Juslin & Laukka, 2004) can all greatly affect an individual's emotional experience while listening to music. These individual differences and other latent factors (e.g. situational factors in certain contexts) appear to affect emotion and are triggered by the interaction of many high- and low-level components that make up a musical composition (Bai et al., 2016; Kim et al., 2010a; Panda, Rocha, & Paiva, 2015; Yang & Chen, 2012). Failure to account for individual differences and latent factors, can have negative implications in emotion induction procedures, including for example, inconsistency in results, inability to systematically

control for underlying mechanisms, and failure to replicate in psychological studies (Frieler et al., 2013; Juslin & Sloboda, 2011; Juslin & Västfjäll, 2008).

2.1 Chapter Goals/Objectives

The literature review in this chapter discusses (a) the models of emotion that are typically studied in music emotion induction research, (b) the musical components that are traditionally studied, (c) a brief overview of machine learning and musical use cases, (d) background research in affective recommender systems related to creating a more personalised system to predict emotional response to music, and finally, (e) research related to creating reliable music stimuli for emotion induction in individuals.

2.2 Discrete and Dimensional Emotion

Two primary models of emotion are used across studies that induce emotion through music; namely, *discrete* and *dimensional*. Discrete emotional models premise that a set of basic emotions or emotional categories exist, from which all other emotions are derived (Ekman, 1992; Ortony, 1990; Plutchik, 1980). For example, Ekman (1992) suggests that happiness, anger, surprise, fear, disgust, and sadness are the six basic human emotions, and any other emotions must therefore stem from a combination of these. Dimensional models on the other hand, suggest that emotional experiences are described by their location on dimensional planes. For example, Russell's circumplex model of emotion, presents a two-dimensional scale based on arousal (i.e. sleepy to activated) and valence (i.e. pleasure to displeasure; Russell, 1980).

Researchers have long debated whether discrete or dimensional models are best suited to represent emotional states. From a physiological perspective, one-to-one mappings between discrete emotions and specific regions in the brain have not been observed, lending physiological support to the dimensional representation (Hamann, 2012). However, this does not necessarily eliminate the importance of discrete

representations. In fact, the value of both discrete and dimensional representations in psychological research can be seen throughout the literature (Harmon-Jones, Harmon-Jones, & Summerell, 2017), and are not necessarily mutually exclusive. That is, the basic emotional states from the discrete model can exist variably across the dimensional scales of arousal and valence (see Figure 1.1) For example, a beautiful sunset and a smiling baby may both induce happiness; however, it is possible that the sunset could induce slightly less pleasure, and a much lower level of arousal than the smiling baby (Hamann, 2012). As such, a dimensional model of emotion affords the ability to differentiate between emotional states with high resolution, which is important in the context of developing a personalised emotion prediction system. Furthermore, the dimensional model has been shown to perform better than the discrete model in characterizing emotionally ambiguous music (Eerola & Vuoskoski, 2010).

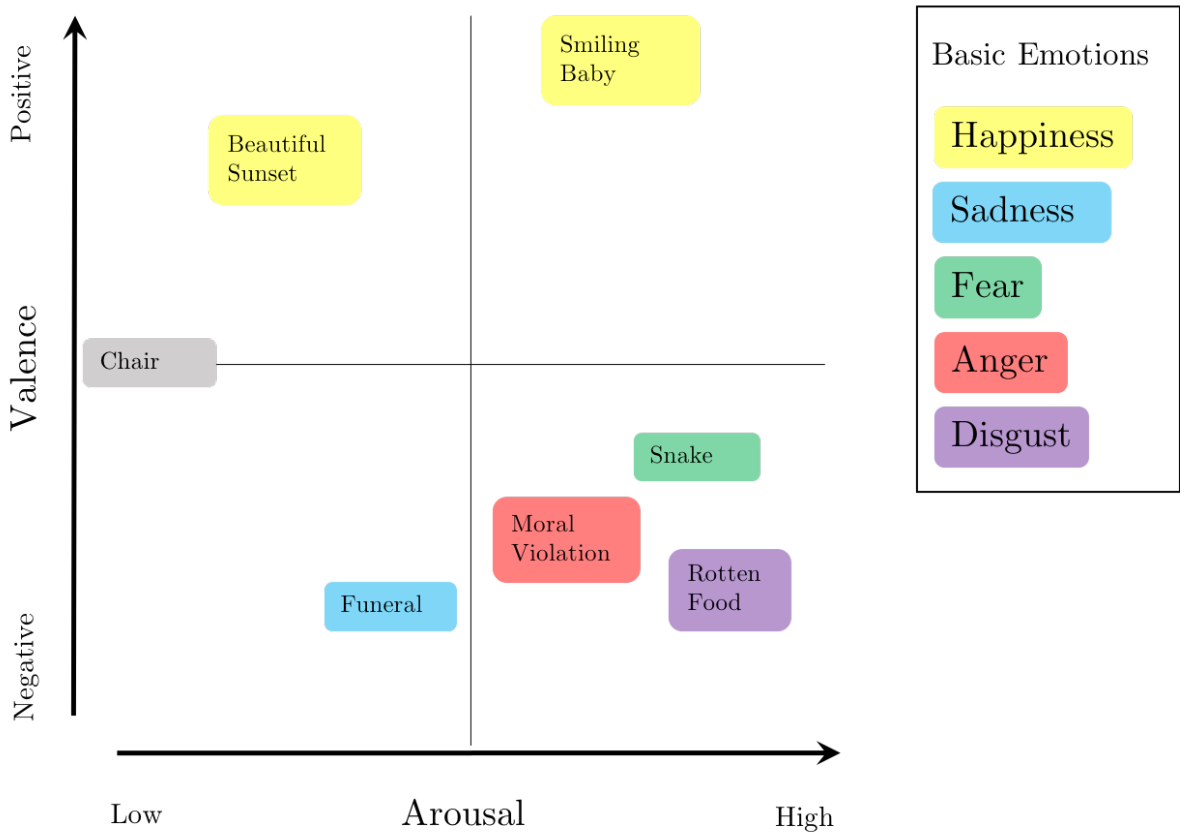

Figure 2.1 A scatterplot showing how different experiences of discrete emotions exist variably along the valence and arousal dimensional scales. For example, while sunsets and smiling babies may both induce happiness, the levels of arousal and valence they induce might differ.

2.3 Emotional Music Components

The emotional responses we experience when listening to music occur because of the interaction between high- and low- level features (or components) of music composition. For example, emotional responses such as sadness, happiness or even a feeling of serenity, may be generated by music through a combination of components such as mode, tempo and pitch. A high-level, intuitive understanding of how musical components may be combined to generate emotional responses is provided in Table 2.1. This table was produced in a study about the application of music in consumer marketing (Bruner, 1990), and was constructed from early works in understanding how music components affect

emotional responses (Gundlach, 1935; Hevner, 1935, 1936). While the research gives a high-level, intuitive understanding of how musical features can be manipulated to elicit emotional responses, further research has been conducted to determine computationally (a) which specific elements of musical composition contribute to emotional responses (Bai et al., 2016; Eerola, Lartillot, & Toiviainen, 2009; Kim et al., 2010b; Panda et al., 2015; Thammasan, Fukui, & Numao, 2017; Yang & Chen, 2012), and (b) how these elements combine additively to do so (Eerola et al., 2013). These studies identify certain low-level musical features (e.g. Mel-Frequency Ceptrum Coefficients (MFCCs), Spectral Roll-Off and Flux, Rhythms, and Tones) and the combinations of these, as key contributors to our emotional experience of music. However, the impact that such features have on the emotional experiences of individuals (i.e. the contribution the musical features make to individual differences) has rarely been explored. These features can vary significantly according to styles and genres of music, meaning that it will not be enough to hypothesize how music components contribute to emotion response if we want to be able to accurately predict emotion induction for individuals.

Table 2.1 Musical Characteristics of Emotional Expression

Musical Element	**Serious**	**Sad**	**Sentimental**	**Serene**	**Humorous**	**Happy**	**Exciting**	**Majestic**	**Frightening**
Mode	Major	Minor	Minor	Major	Major	Major	Major	Major	Minor
Tempo	Slow	Slow	Slow	Slow	Fast	Fast	Fast	Medium	Slow
Pitch	Low	Low	Medium	Medium	High	High	Medium	Medium	Low
Rhythm	Firm	Firm	Flowing	Flowing	Flowing	Flowing	Uneven	Firm	Uneven
Harmony	Consonant	Dissonant	Consonant	Consonant	Consonant	Consonant	Dissonant	Dissonant	Dissonant
Volume	Medium	Soft	Soft	Soft	Medium	Medium	Loud	Loud	Varied

Adapted from Bruner (1990)

Despite the impact of musical features on emotion being highly subjective, the musical stimulus that is used for emotion induction in psychology research remains largely unchanged and its impact at the individual level unexamined (Juslin & Västfjäll, 2008). Much more research is required to understand how specific music components affect an individual's emotional responses so as to inform the effective selection music stimuli. This issue is growing increasingly relevant with researchers lamenting the lack of replication in empirical studies in general, and more specifically in music psychology studies (Frieler et al., 2013). To address this issue, it is no longer sufficient to study the components of music and their interaction with emotion by averaging responses across individuals and across music excerpts (Cronbach, 1957). A novel approach to music emotion induction is required; one that takes in to consideration the many latent factors that contribute to the emotional experience of an individual when listening to music. This would allow researchers to more reliably induce and manipulate emotion in basic emotion and applied research.

2.4 Accounting for individual difference in emotion responses

2.4.1 A brief overview of Machine Learning

Machine learning is a paradigm in computer science, whereby algorithms or statistical models are designed to detect patterns and structures in input data in order to perform tasks such as discrimination, ranking, and inference. In the context of this thesis, machine learning is used to form models that predict personalised emotional responses to music. However, given the interdisciplinary nature of this work, it is important to first elucidate some machine learning concepts and methods that are referred to throughout this thesis, including the types of approach, task, and some testing techniques.

There are typically three approaches to developing a machine learning model:

- Supervised Learning: Dependent Variables (labels) are used in the model "training" stage as targets for which the models learn to reproduce from independent variables

with minimal loss (optimisation). For example, you may have a dataset of musical stimuli that have been labelled with ratings on a valence scale, and the role of the algorithm would be to determine the valence rating of any given new music stimulus (Bai et al., 2016).

- Unsupervised Learning: Models use patterns and structures in input data to automatically differentiate or arrange instances - without the need of training labels to minimise loss. For example, you may have a dataset of music which you wish to automatically organise into playlists of similar items, unsupervised learning methods would be able to group music based on similar musical features (Lin & Jayarathna, 2018).
- Reinforcement Learning: Models are trained based on actions taken. As the agent explores the environment and changes states, it receives feedback in the form of rewards or penalties, and thus learns which actions are appropriate given current, previous, and potential future states. For example, you may wish to automatically create chord progression in generative music application (see section 2.5), reinforcement learning would be able to progressively generate the appropriate musical chords based on position in musical context (Shukla & Banka, 2018).

2.4.1.1 Classification, Regression, and Clustering

There are also three major categories of machine learning tasks to be aware of classification, regression, and clustering. Classification is concerned with determining which class(es) a set of input data belongs to - it outputs discrete values (labels). A classification algorithm, for example, might be used to predict the discrete emotion of musical excerpts (e.g. Happy, Sad, Angry, or Relaxed) – categorising the excerpts into a discrete emotional states based on the musical features extracted from them (Laurier, Grivolla, & Herrera, 2008). Alternatively, regression is concerned with mapping a set of input variables to a continuous rather than discrete values. For example, the output of a regression model could be used to predict the degree of arousal or valence induced by

musical excerpts (Eerola et al., 2009). Regression is the machine learning approach used in this thesis to predict personalised music emotion ratings on dimensional emotional scales (see Chapters 3 - 5).

Some of the more commonly used algorithms for classification and regression problems include support vector machines and random forest and k-means for unsupervised learning. Support vector machines, illustrated in Figure 2.2, are a technique for which the machine learning algorithm learns to construct an optimal hyperplane that divides a data space into subsets (e.g. classes) using, for example, the data vectors that maximise the margin between subsets as support vectors for the hyperplane (Cortes & Vapnik, 1995). Random forests, also shown in Figure 2.2, on the other hand, are ensemble methods which use a multitude of decision trees as base classifiers and outputs, as a prediction, the value that is the mode (classification) or mean (regression) of the individual decision trees' predictions (Breiman, 2001).

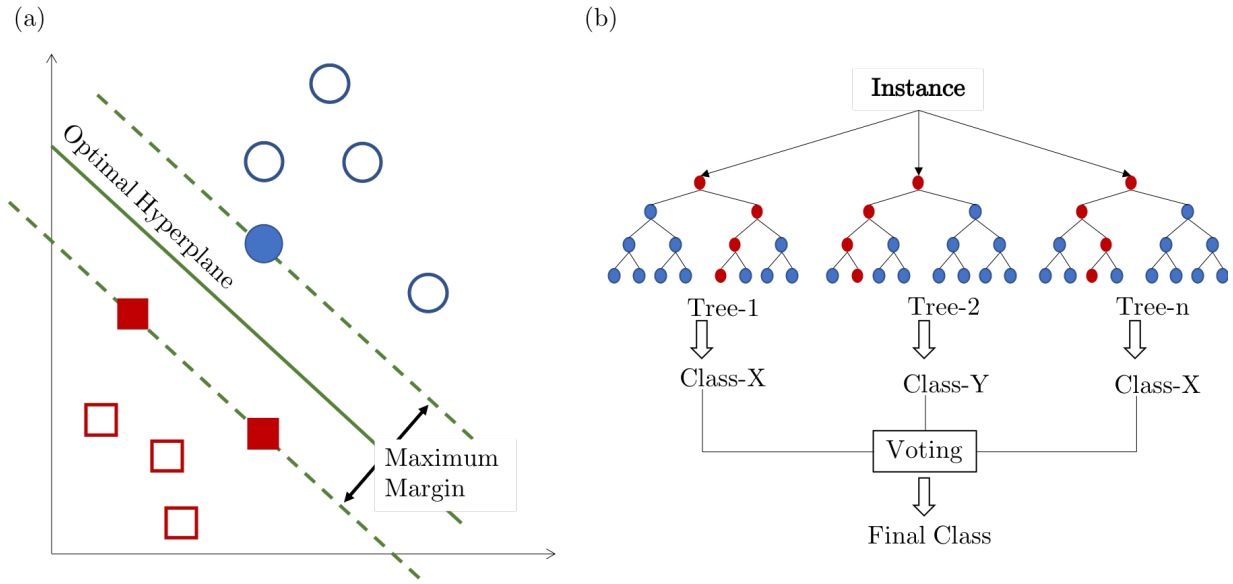

Figure 2.2 Illustration of how (a) support vector machines (b)random forests form predictive models. Support vector machines discover an optimal hyperplane to divide the space into classes. In the example above, the SVM finds the hyperplane that maximises the margin between the square and circle classes. The output of random forest is the result of many individual decision trees voting on the prediction. In the example above, the input instance passes through paths in each decision tree, with each outputting a class prediction. Voting (mode or mean) then ultimately determines the final prediction of class.

Clustering is concerned with partitioning instances into groups based on similar attributes with other instances, for example similar emotional response to music stimulus (see section 4.3). Clustering is a technique commonly associated with unsupervised learning, in which labels for data are not available and associations need to be drawn between instances. K-means is an algorithm commonly used for unsupervised clustering, the method relies on a distance function to cluster instances into "k" subsets based on minimising the within-sum-of-squares (WSS) and maximising the distance between clusters (Hartigan & Wong, 1979). Illustrations of clustering can be seen in section 4.3.2.

2.4.1.2 Feature Engineering and Selection

Critical to developing successful machine learning models is the selection of appropriate input features (Guyon & Elisseeff, 2003). In the worst case, the features captured may not be very informative about (predictive of) the output, however challenges can also include features with high collinearity, in which case those features can have an unbalanced influence in predictive modelling, or "the curse of dimensionality", in which too many feature dimensions cause the data space to become too sparse - making it easy for algorithms to overfit. Because of these challenges, several methods for reducing dimensions have been developed, including feature selection methods that attempt to identify the most relevant features and minimise redundancy (e.g. ReliefF) and feature transformation methods that attempt combine multiple features into aggregated values (e.g. Principal Component Analysis). These methods for feature selection are used and described further in Chapter 3.

2.4.1.3 Deep Learning

While traditional machine learning techniques rely appropriate feature extraction and selection as a pre-processing step, deep learning is a paradigm by which relevant feature representations can be learned directly from raw input data (e.g. images) (Goodfellow, Bengio, & Courville, 2016; LeCun, Bengio, & Hinton, 2015; Schmidhuber,

2015). The ability of deep learning models to learn features directly from the input is critical because it ensures that, given (a) best practices are followed and (b) the representations actually exist in the data, that the features used for prediction are highly informative about the output predictions. For example, the filters of the lower-layers of a multi-layer convolutional neural network (CNN) that was performing object detection could learn to recognise features such as lines and other simple patterns in the images. As the layers progress, the patterns of the lower-layers combine into more complex patterns and representations until eventually the CNN would determine if the object existed in the image. Allowing features to been engineered by deep learning algorithms means that assumptions do not need to be made about what low-level features are relevant to predictive outcomes, and could therefore allow for more predictive outcomes (see results in section 4.5).

2.4.1.4 Best Practices

As explored briefly above, there are a wide variety of basic machine learning approaches, tasks and techniques, however it is important to note that there are certain overarching elements that must be considered across all machine learning. For example, training, testing and validation is required for all machine learning, and best practice dictates that performance of an algorithm should be evaluated against novel data to ensure that model is not simply overfitting the training data (Cawley & Talbot, 2010). As such, available data is typically split into training, testing, and validation sets: the training data is used to train the models and generate features, the validation set is used for frequent evaluation of the models performance (i.e. hyperparameter tuning), and testing data is used only to evaluate the models final performance.

There are several ways to split training, testing, and validation data. If there is a sufficiently large and represented amount of sample data, it is appropriate to perform a percentage split on the sample data - forming three sets of data - testing, training, and validation. However, if there is limited sample data to train and evaluate models on, cross-

validation (a resampling procedure) is commonly used to evaluate the model (Kohavi, 1995). This procedure is used in 3.6.3, and splits the data into "k" folds, and incrementally uses one fold (of each of "k" iterations) as a test set while the other folds are used to train the model. Ultimately, the performance of the model is an aggregation of the performance of "k" runs. In each case, the groups can be formed by picking randomly, stratified (e.g. instance balances), or blocked (e.g. temporal dependence), depending on the nature of the data and thus the statistical validity. Crucially, in cross-validation, instances must remain in the same fold throughout the training and evaluation of a model. As an example, an illustration in Figure 2.3, shows how data can be split into training, validation, and testing sets across five folds.

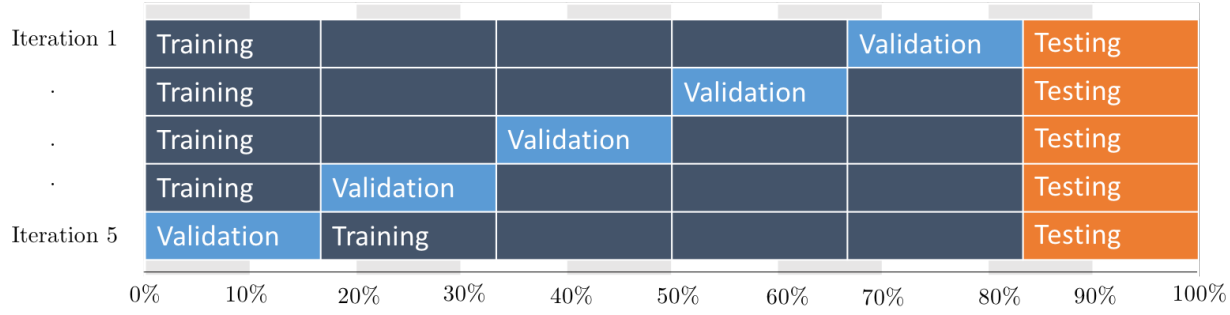

Figure 2.3 Illustration of how a dataset is split into training, validation, and testing set across 5-fold cross validation.

2.4.2 Recommender Systems

Recommendation systems are a class of algorithms that filter content based on user preferences which are formulated from the cohorts of similar groups of individuals and/or items. The two main approaches to recommendation systems are Collaborative Filtering and Content-Based Filtering. Collaborative filtering takes advantage of 'collective intelligence', that is, knowledge generated by the preferences of an entire userbase (Shardanand & Maes, 1995). In such cases an n x m matrix (i.e. users by items) is constructed with each cell containing a user's preference feedback (e.g. like or dislike) for a specific item and then a matrix-factorisation routine is performed to fill empty cells for each user based on other users with similar preferences. As a simple example: if person A

and person B both 'liked' 10 music excerpts and 'disliked' another 10 excerpts, we can assume that person A and B might have similar preferences, and thus use each of their ratings on new excerpts to predict the other person's preference. Content-based filtering on the other hand, takes advantage of knowledge about the content itself in relation to a user's preferences (Kuo & Shan, 2002). Content is analysed to determine components that contribute to its construction and these components are fit to a user's profile. The algorithm is then able to recommend content with similar construction. The most successful recommender systems typically combine these approaches in a 'hybrid' recommender system (Chen & Chen, 2001; Yoshii, Goto, Komatani, Ogata, & Okuno, 2006).

There have been several very prominent examples of music recommender systems developed on data descriptors for content and user modelling (e.g. semantic tags in Pandora[1]); however, in the last 5 years researchers have begun to explore the development of affective music recommendation systems (Song, Dixon, & Pearce, 2012). Affective recommendation systems utilise a user's affective response profile, which is created through a person's affective ratings of items or objects, to either generate or improve the quality of recommendations to a user (Tkalcic, Kosir, & Tasic, 2011). w created a generic emotion-based music recommender system that analysed the emotion of film music in order to generate recommendations. Emotional components of music were identified through a process of feature extraction on music tracks, and then the authors developed models that were used to 'discover' the association between music features and the film's emotion (e.g. a sad film). The system in this case was not personalised to a users' emotional profile, but to that of the films, so if a film was sad the relation was made to sad music components. Yoon, Lee, and Kim (2012) developed a personalised recommender system based on low-level music features that were found to trigger emotional responses. They used a database of 400 songs. In order to establish an emotional profile to fit each

[1] www.pandora.com

user and make recommendations of affective songs, they had users pre-identify a song for each discrete class of angry, happy, sad, and peaceful from the (discrete) extended Thayer's mood model (Thayer, 1989). Though not thorough, the study showed promising results for a personalised emotion recommender system based on musical features, and the authors acknowledged the need for understanding how individual differences, preferences, and components of music affect a person's emotional response to music stimuli.

Whilst the work in developing personalised affective music recommendation has shown the promise of affective music recommendation engines and the recommendation approach in general, these systems have largely been developed for entertainment purposes. Without the goal of developing reliable personalised music emotion induction for areas such as basic emotion research, psychology of music and musicology research, and applied research such as music therapy, past studies lack the methodology and rigour required to incorporate and predict individual differences in emotional responses to music. A further benefit of recommendation engines in the context of this thesis is that they can take advantage of prior knowledge to more intelligently account for a person's personal preferences and other individual differences, as opposed to a brute-force algorithm which would start from ground up building a unique understanding for every individual user.

2.5 Creating malleable music stimulus sets

In order to develop music selection systems that reliably induce emotional responses in individuals (i.e. accounting for the individual differences associated with latent factors), psychologists and researchers would either need (a) a very large database of music segments, or (b) a small, malleable database of music segments that could be manipulated to cater to individual differences in affective responses. The benefit of using malleable datasets is that they are scalable to research projects of different sizes or specifications, and they afford the ability to control for specific musical variables while manipulating others. Eerola et al. (2013) used MIDI-produced stimuli to manipulate musical parameters

such as mode, tempo, dynamics, articulation, timbre, register, and musical structure. MIDI pieces are electronically created and easily manipulable, however they require musicians or composers to produce the originals and can sound artificial to subjects if not produced with care, which may introduce other sources of variance in subjects.

A field of research dedicated to the development of affective algorithmic composition (AAC) has grown considerably over the last decade (Williams et al., 2014), and lends considerable insights on how we can form malleable music stimulus set for the induction and manipulation of individuals' emotional responses. These systems are designed to create music based on intended affective response, as opposed to the traditional algorithmic composition approaches which do not consider affect in generation. Algorithmic composition has existed since the 1950s and a large number of AI methods have been employed as solutions (Fernández & Vico, 2013; Nierhaus, 2009). Approaches generally fall into three categories: generative compositions, transformative compositions, and sequenced compositions (Rowe, 1992).

The generative approach to algorithmic composition uses rules to create music. For example, music is formed of underlying grammatical structures such as those described in the Generative Syntax Model (Rohrmeier, 2011), which is illustrated in an analysis of Bach's Chorale 'Ermuntre Dich, mein schwacher Geist' in Figure 2.4. This model describes a tree-based hierarchical structure of generative rules of western tonal music that is extendable to both historical and modern pieces of music. Algorithmic tools leverage such rules to create new music. The limitation of such systems though is that rules must be described exhaustively, however, the rules that describe how music influences emotion are not fully understood, so it is not currently possible to articulate them. A further concern with these systems is that it is very challenging to replicate the expression of music through rules – especially rules which only consider the syntactical and semantic components and structures of music composition. In fact, much of what makes music sound 'musical' is a result of what is known as performance practice, that is, a legacy of

unwritten practices that musicians follow in interpreting a musical performance in different styles (Rink, 2005). For example, the concept of "swing" has proved virtually impossible to replicate in rules, and is an interpretation that many musicians even struggle with (Friberg & Sundströöm, 2002; Lindsay & Nordquist, 2007). These sorts of performance practices have a significant effect on our psychological perception of musicality (London, 2012), and therefore failure to account for them could have unintended consequences on a person's emotional response to music.

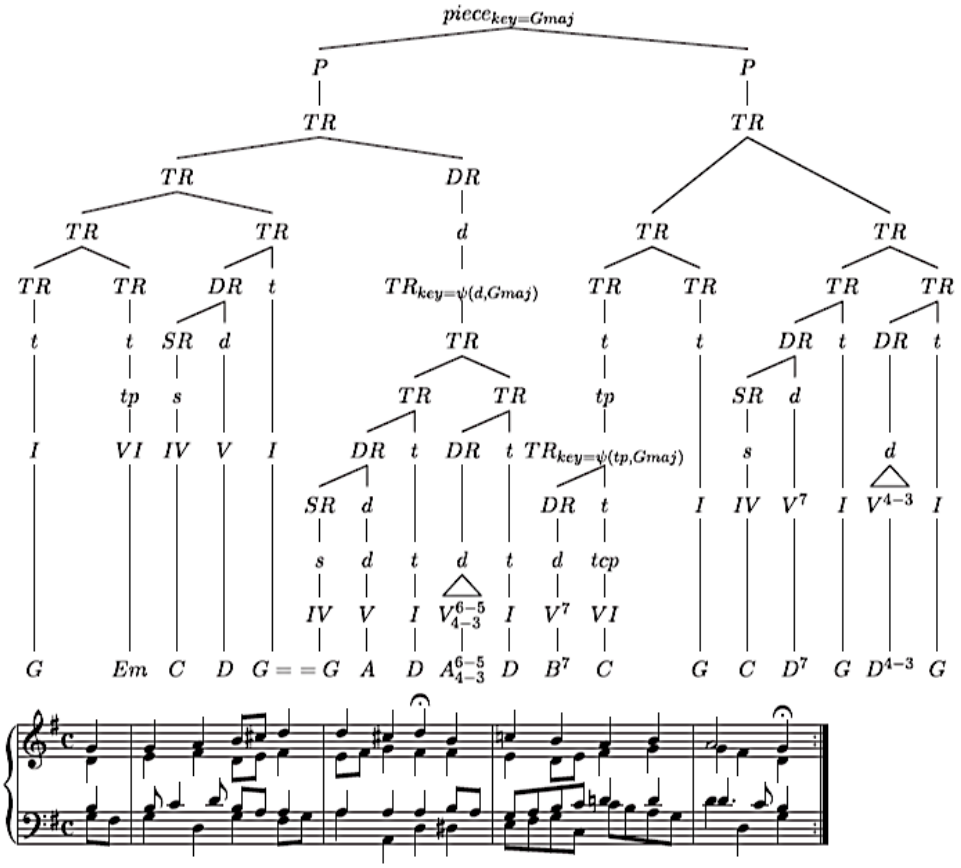

Figure 2.4. Analysis of Bach's Chorale 'Ermuntre Dich, mein schwacher Geist', mm. 1-4 using the Generative Syntax Model.

On the contrary, a transformative system takes existing music as input, and transforms it to produce a different output. This is often accomplished through some inference by the system about musical rules of transformation. For example, Weinberg, Raman, and Mallikarjuna, (2009) have used Markov Chains to learn musical input from

users, captured through MIDI, in order to develop collaborative improvisation with a musical robot. The robot generates musical responses similar to that of the human performers as a result of the learnings of the Markov Chains. A system like this accounts for the challenges of performance practice by learning directly from the input of human performers what those practices are. This could ultimately make the output of the transformative system sound more musical. However, it would be difficult to use such systems in a context where a researcher wanted to control the emotion for an individual user, because (a) there would first need to be a suitable musical input, (b) there would need to be enough flexibility in the system to change that input in order to map musical parameters to induce emotional responses, and (c) the system would need to be trained to be 'musical' by a musician – who may not be readily available at researchers' disposal.

Most promisingly, the sequence based approach works by combining pre-recorded excerpts (or loops) of music into larger musical compositions. The pre-recorded loops can be used as standalone excerpts, or combined horizontally (sequentially in time) or vertically (harmony or polyphony) to create novel excerpts. The sequencing approach is useful because it allows composers to create the music, and for the system to reorder that music in time while also allowing for segments to be pre-rated using affective rating scales by participants. An example of this type of system, in the entertainment domain, is The Affective Remixer (Chung & Vercoe, 2006; Vercoe, 2006). The authors experimented with the layering and complexity of musical loops, and recorded galvanic skin response (GSR) (used to measure arousal), foot tapping (used to measure valence), and self-reported arousal (engaged-soothing) and valence (like-dislike) scales. Based on a listener's current affective state (which was inferred using state-transition Markov Chains), the system selects music segments and re-arranges them to induce a new target affective state. However, because this system was built with the goal of entertainment and not in attempt to reliably induce emotion, no evaluation of the system's performance was reported. The authors reported inconsistencies with GSR and arousal reports, and concluded that foot

tapping correlated with valence, however they acknowledge the need of further research and the inclusion of additional listener data like music preference to improve accuracy.

The loop approach to affective music composition has been much more widely explored in the domain of interactive gaming audio. In the game scoring domain, composers are tasked with creating interactive music to accompany a user's gaming experience. The music must continue to adapt to the current state of gameplay, while keeping aligned with the theme of the game and not sounding disconnected or artificial as it develops. The sequencing approach has been used to much success in developing affective music in the context of games (Collins, 2008, 2017; Collins, Kapralos, & Tessler, 2014; Enns, 2015; Phillips, 2014). Combined with a personalised emotion prediction system, these loop-based systems could be extended to the domain of reliably inducing and manipulating emotional responses in individuals.

2.6 Chapter Conclusion

Through a review of the relevant literature, in this chapter I established how the learnings of emotional models, musical components, recommender techniques, and music generation systems can be used towards the development of more robust and reliable music emotion induction for individuals. Dimensional models of emotion offer high resolution, and will be well suited for capturing subtleties in emotional responses. As stated in Section 1.1, the failure to account for individual differences in music emotion responses can lead to poor reliability and replicability in psychological studies, and poor control for experimental variables in emotion research. Researchers have previously shown that there are psychological mechanisms that lead to individual differences in emotional responses, and the literature in Section 2.3 shows that even in terms of music, very little is known about the components that affect individuals.

However, despite these obstacles, the recommender system approaches, discussed in Section 2.4, appear to illuminate a path toward creating personalised emotion prediction

for individuals. This research has shown that collaborative and content-based filtering techniques can potentially be used in the development of a personalised emotion prediction system, to capture and account for the many latent variables that affect an individuals' emotion response to music. This would allow researchers to select stimuli that more precisely induce and manipulate emotional responses. Furthermore, the research that has been done in AAC, specifically the sequence-based (loop-based) approach has shown that this is a promising avenue to pursue in creating a stimulus set that could be catered to reliably and effectively induce emotion in individuals.

Chapter 3 Electronic Music Stimuli and Emotion Prediction

Music and emotion research continues to evolve, however the musical stimuli used for affect induction in psychology research remain largely unchanged (despite the fact that compositional techniques and elements in music have changed with modern generations). A recent review (Eerola & Vuoskoski, 2013) of 250 music and emotion studies shows that researchers still largely rely on familiar classical music excerpts (48%), that stimulus sets are often chosen either arbitrarily by the researcher (33%) or that the selection process is not identified (39%)[2].

The best music for emotion induction should effectively induce affect but also minimise the role of subjective musical preference, which is often associated with familiarity and genre preference. Eerola and Vuoskoski (2010) minimised the effect of subjective music preference by building a stimulus set from film soundtracks. They do not fall easily into genres because they are designed with the explicit goal of inducing affect within a narrative context, and are meant to do so in global audiences.

However, the film scores that were selected (Eerola & Vuoskoski, 2010) were still limited to a traditional orchestral style. Electronic music is more representative of modern musical styles that may be useful in some research domains. Furthermore, it is more

[2] Another 8% percent used a Pilot study, 9% used a previous study, 6% used an Expert panel, and 4% used Participants for selection.

amenable than orchestral music to online analysis and manipulation of musical parameters. A set of musical excerpts that can be digitally manipulated across a range of dimensions has many uses in psychological research. For example, music that can be easily manipulated can be used to determine the causal relationship between musical features and emotional responses. Such music can also be used as the output of brain-computer interfaces and other emerging technologies that produce music as a product of mental activity (Christopher, Kapur, Carnegie, & Grimshaw, 2014). Manipulation of musical features is very difficult using solely acoustic music because it requires an orchestra or ensemble to record the specific excerpts for the experimenter. Researchers often must rely on MIDI produced stimuli in order to manipulate musical parameters. These are electronically created and easily manipulable, however they require someone of musical background to produce, and often sound quite synthetic in nature (Eerola et al., 2013).

3.1 Chapter Goals/Objectives

In this study I used a similar method to that described by Eerola and Vuoskoski, (2010), to produce a stimulus set based on modern electronic film music. The vast majority of research up to this point has used orchestral music for affect induction. However, given that I wanted to develop a stimulus set that is manipuable, and electronic music is the best approach to doing that, I needed to (a) create a set of electronic music stimuli with a broad affective range, and (b) identify the specific low-level musical features that impact emotional response.

First, a panel of experts preselected music excerpts according to predefined criteria (i.e. music must be electronic and cover a range of emotions). Then the music excerpts were rated by naive participants on the dimensional emotion scales of valence and arousal; as well as degrees of liking and familiarity. These ratings were used to develop a modernised affective music stimulus set. Emotional ratings for these electronic excerpts were also compared to those elicited by orchestral musical excerpts, which are the

mainstay of music and emotion research. Orchestral and electronic music differ in spectral qualities (e.g. timbre), that are strong predictors of emotional experience. It was therefore important to determine the range of emotional experience electronic music can elicit.

Second, given that the physical features (i.e. timbre and rhythm) of music are one of the factors that drive emotional responses in individuals, I evaluated which specific musical elements of the stimulus set correspond with the emotional ratings provided by participants. I performed musical feature extraction followed by exploratory factor analysis in order to identify the musical components of the stimulus set and how they contribute to emotional responses in subjects. Furthermore, I explored the predictive qualities of the derived musical factors using regression models.

3.2 Method for Electronic Music Stimulus Selection

3.2.1 Music Excerpts

A panel of experts in electronic music composition (1 music professor, 4 PhD students in music, and 1 professional musician) were recruited in order to preselect a large set of music excerpts that would be later evaluated by non-selected participants. A total of 15 film scores using primarily electronic music, composed between 1999 and 2014, were selected to be evaluated by the experts (Table 3.1). These soundtracks encompassed 13 different film, TV, and video game genres. The 15 albums were randomly placed into three groups of five albums each. Within each group, the music tracks were anonymised and placed in random order. Each group was distributed to two experts and consisted of approximately 100 music tracks each (277 tracks total). The experts were given the following instructions:

In this task we are trying to collect a series of music samples that we can use in future experiments. For these experiments, we want a selection of electronic music that represents different emotional qualities. For example, we can distinguish music based on

whether it makes us feel positive or negative – this is a component we refer to as valence. Music at the negative end of the valence scale can make us feel unhappy, annoyed, unsatisfied, melancholic or despairing, while music at the positive end of the valence scale can make us feel happy, pleased, satisfied, contented or hopeful. We can also distinguish music based on its intensity. Music at the low end of the intensity scale can make us feel relaxed, calm, sluggish, dull, sleepy or unaroused, while music at the high end of the intensity scale can make us feel stimulated, excited, frenzied, jittery, wide-awake or aroused.

In this task, we ask you to listen to the selection of music provided and to choose excerpts from 5 music tracks that you think best represent the four categories described above (i.e. high arousal positive, high arousal negative, low arousal positive, low arousal negative; 20 excerpts in total). The excerpts should fit the following criteria:

1. No Lyrics

2. No Dialogue

3. Excerpts must be 20-30 seconds in duration (within the constraints of musical phrasing)

4. Excerpts must be predominantly electronic in nature (this excludes midi mock-ups of acoustic instruments).

This process resulted in the identification of 120 emotional music excerpts (i.e. 30 for each combination of arousal and valence). The emotional ratings, albums, timepoints, and track numbers for the 120 excerpts are presented in Table A.1 of Appendix A. An additional 40 orchestral excerpts were selected from the Eerola and Vuoskoski (2010) stimulus set, to allow direct comparisons between orchestral and electronic music in the same participants. The excerpts were selected to represent the full emotional range of valence and intensity present in the stimulus set produced by Eerola and Vuoskoski (2010). A full table of the emotional ratings, albums, track numbers, and timepoint for the 40 excerpts is presented in Table A.2 of Appendix A.

Table 3.1. Soundtracks selected for stimulus set

Film	Composer	Year	Genre
American Beauty	Thomas Newman	1999	Drama, Romance
Batman Begins	Hans Zimmer, James Newton Howard	2005	Action, Adventure
The Dark Knight	Hans Zimmer, James Newton Howard	2008	Action, Crime, Drama
Far Cry 4	Cliff Martinez	2014	Video Game: Action, Thriller
The Lego Movie	Mark Mothersbaugh	2014	Animation, Action, Adventure
Gone Girl	Trent Reznor and Atticus Ross	2014	Drama, Mystery, Thriller
A Series of Unfortunate Events	Thomas Newman	2004	Adventure, Comedy
A Road to Perdition	Thomas Newman	2002	Crime, Drama, Thriller
Spring Breakers	Cliff Martinez	2012	Action, Crime, Drama
The Social Network	Trent Reznor and Atticus Ross	2010	Biography, Drama
300 Movie	Tyler Bates	2007	Action, Fantasy, War
Cinderella Man	Thomas Newman	2005	Biography, Drama, Sports
The girl with the Dragon Tattoo	Trent Reznor and Atticus Ross	2011	Crime, Drama, Mystery
Finding Nemo	Thomas Newman	2003	Animation, Adventure, Comedy
The Knick	Cliff Martinez	2014	TV: Drama

3.2.2 Participants

A total of 242 participants were recruited through an online crowdsourcing service, CrowdFlower[3], to provide evaluation of the musical excerpts on a number of emotional dimensions. The participants came from 48 different countries including India (10%),

[3] Note: this platform was rebranded as 'Figure Eight' in 2018.

Indonesia (7%), Serbia (6%), Bosnia and Herzegovina (6%), Portugal (5%), and the USA (5%). The ages of the participants ranged from 18 to 68 years (M = 29.04, SD = 9.05; 26% female). This meant a reasonably diverse population evaluated the emotional quality of the music stimulus set. Thirty-four percent of participants indicated having no musical training at all, 25% had less than a year of training, 20% had 1-3 years of training, and 21% had over 3 years of musical training. One percent of participants replied that on a typical day they didn't listen to any music, 20% replied that they listened to music for less than an hour, 33% 1-2 hours, 21% 2-3 hours, 11% 3-4 hours, and 14% more than 4 hours. Participants were paid $0.50USD.

3.2.3 Procedure

In order to obtain emotion ratings on the music excerpts, a ratings task was administered via the Qualtrics platform. A total of 160 audio files were evaluated: the 120 chosen by the musical experts, and an additional 40 orchestral excerpts selected from the study by Eerola et al., (2009). The orchestral excerpts represented negative and positive valence, crossed with low and high energy[4]. Excerpts were divided into 4 groups of 40 each (approximately balanced based on categorization of excerpt provided by the experts), and each participant listened to 40 excerpts (≈20 minutes). Each participant was assigned to rate either valence or arousal so that they could focus on one dimension. Furthermore, within each group of participants, half of the participants rated *felt emotion* – that is, how the music made them feel personally. The other half rated *perceived emotion* – that is, the emotion they believe the music is intending to convey. While the felt and perceived emotion of music will likely tend to be in agreement, it is possible that music can convey one emotion (such as sadness) but make the participant feel another (such as a positive appreciation of beauty).

[4] The arousal scale in Eerola, Lartillot, and Toiviainen (2009) splits arousal into 2-dimensions: *energy* and *tension*. When controlling for valence, the energy dimension is closely aligned with arousal.

After each excerpt, participants were presented with three 9-point Likert scales on which they rated the affect (felt or perceived, valence or arousal – according to their assigned ratings condition), how much they liked or disliked the excerpt, and how familiar the excerpt was to them. Each excerpt was rated by 14 to 17 listeners. Likert scales were set low to high for arousal and familiarity, and negative to positive for valence and preference. Audio file order was randomized for each participant.

After providing the ratings, participants completed a short questionnaire adapted from Belcher & Haridakis (2013), which asks questions about the participant's music interest, musical preference, and musical training. This information was collected for demographic purposes and to inform other projects, but was not use in analyses.

3.3 Resulting Electronic Music Stimulus Set

3.3.1 Emotion Ratings

Affect ratings for the 120 electronic music excerpts across Felt Arousal, Perceived Arousal, Felt Valence, and Perceived Valence conditions appear in Table 2. The individual ratings for each excerpt are presented in Appendix A. All analysis is conducted using the R programming language and "psych" package. A Pearson's Product Moment correlation analysis with Holm p-value correction for multiple comparisons shown Table 3.2 revealed a very strong positive correlation between the felt and perceived conditions of both arousal ($N = 120$, $r = 0.848$, $p < .001$) and valence ($N = 120$, $r = 0.846$, $p < .001$). However, a two-sample paired t-test, in which the music excerpts are used as subjects, reveals significant differences between felt and perceived emotion in both arousal ($t = 3.8$, $df = 119$, $p < .001$, $d = 0.264$) and valence ($t = 7.33$, $df = 119$, $p < .001$, $d = 0.4$). Music was rated as more positive and more arousing by participants who rated the felt emotion compared to those who rated the perceived emotion. However, Table 3.3 shows that despite this shift in ratings as a function of task, felt and perceived ratings are highly correlated. Also of note is that liking and familiarity seem to have high positive

correlations with valence. One possible explanation for this is in the nature of film music. Film music has adapted some tried and tested techniques for effectively conveying emotions to global audiences. Because of this, some excerpts may be very reminiscent (cliché) and quite catchy - resulting and higher ratings of both familiarity and liking. The difference between felt and perceived valence may suggest the "negative/opposite" or "unmatched relationship" effect (Gabrielsson, 2002; Schubert, 2013). This is the case when the music may convey one emotion, but the listener may feel another. For example, an excerpt may convey negative valence, but the user may enjoy it and experience positive valence.

Table 3.2. Summary statistics for ratings in each condition

Condition	Median		Mean		SD		95% CI Lower Bound		95% CI Upper Bound	
	Perceived	Felt	Perceived	Felt	Perceived	Felt	Perceived	Felt	Perceived	Felt
Arousal	4.97	5.33	5.11	5.38	1.44	1.23	4.85	5.16	5.37	5.60
Valence	5.1	5.62	4.99	5.4	1.18	1.15	4.77	5.19	5.20	5.60

All ratings are on a 9-point scale, from low to high for arousal and negative to positive for valence.

Table 3.3. Correlation between item ratings in each condition

Condition	Felt Arousal	Perceived Arousal	Felt Valence	Perceived Valence	Liking
Felt Arousal					
Perceived Arousal	.848***				
Felt Valence	.231.	.206			
Perceived Valence	.096	.118	.860***		
Liking	.038	-.071	.829***	.786***	
Familiarity	.188	.116	.717***	.706***	.831***
Significance: * p<.05, ** p<.01, *** p<.001, using Holm p-value correction					

3.3.2 Comparison to Orchestral Ratings

Participants also rated 40 excerpts taken from Eerola et al. (2009), which in the original study were assessed on 3-dimensions: Energy, Tension, and Valence, however in this study are assessed along the Arousal and Valence dimensions. In order to establish that the ratings for the orchestral excerpts showed consistency across studies, I performed correlation analyses between the valence, energy, and tension ratings reported in the original study and the valence and arousal ratings collected in this study for the same excerpts. Overall, these high correlations across studies, calculated using a Pearson's Product Moment correlation analysis with Holm p-value correction for multiple comparisons and shown in Table 3.4, support the reliability of the ratings collected here. Analysis showed a strong positive correlation between current ratings in the felt and perceived valence conditions and the original valence ratings, demonstrating consistency in the valence dimension. Furthermore, the energy ratings from Eerola et al. (2009) were positively correlated with both felt and perceived arousal and felt and perceived valence, suggesting that the energy dimension is a construction of both arousal and valence. Tension showed a strong negative correlation with felt and perceived valence, indicating that tension is a product of negative valence.

Table 3.4. Correlation between comparison ratings and original ratings (across the top) and ratings collected from this study.

	Energy	Valence	Tension
Felt Arousal	.803***	.193	.221
Perceived Arousal	.675***	.197	.144
Felt Valence	.532**	.938***	-.650***
Perceived Valence	.560**	.879***	-.584***
Significance: * = p<.05, ** = p<.01, *** = p<.001			

In order to compare the emotional ratings made by our non-selected participants for the 40 excerpts from Eerola et al. (2009) to the electronic excerpts selected by our

expert panel, I split the axis of each emotional dimension into four quartiles: extreme low (10%), moderate low (20%–40%), moderate high (60%–80%) and extreme high (90%), (as in Eerola and Vuoskoski, (2010), and performed a $4x2x2$ ANOVA (4 percentiles, 2 stimulus set, 2 emotion states), comparing sets across emotional dimension using the ratings from my participants only. This process yielded no significant main effects (besides the expect effects of quadrants), indicating that the electronic excerpts showed the same affective range as orchestral excerpts. Figure 3.2 demonstrates the spread of ratings across emotion quadrants, and spread within quartiles respectively. As can be seen, the electronic music excerpts from this study cover an emotional range as large as acoustic excerpts selected from Eerola et al. (2009).

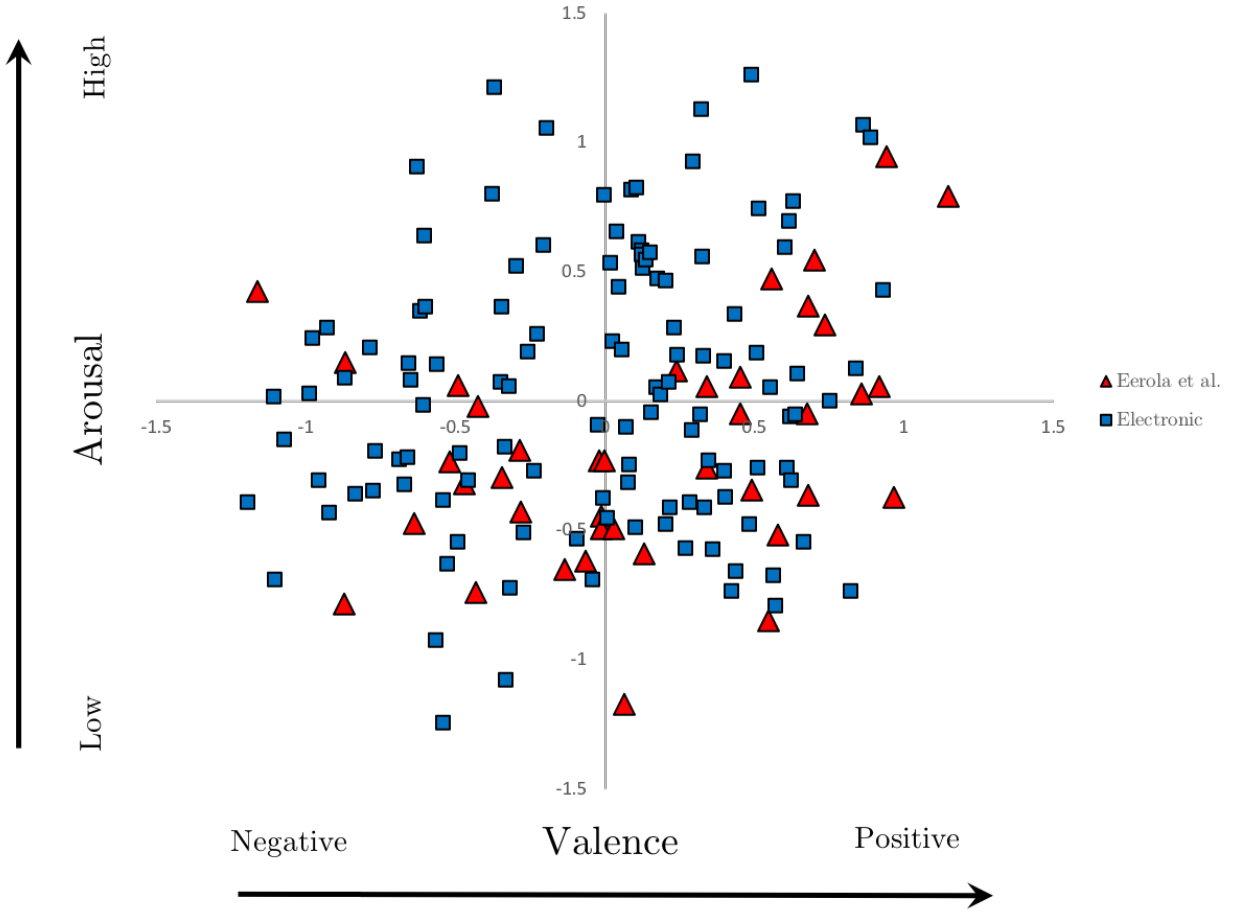

Figure 3.1 This plot illustrates the range of ratings for music excerpts collected in this study across the felt (experienced) arousal and valence space. It compares ratings of electronic-based excerpts selected by experts with ratings on excerpts originally selected in Eerola et al. (2009). As illustrated, our excerpts cover a similarly large space as orchestral musical selections described in previous work.

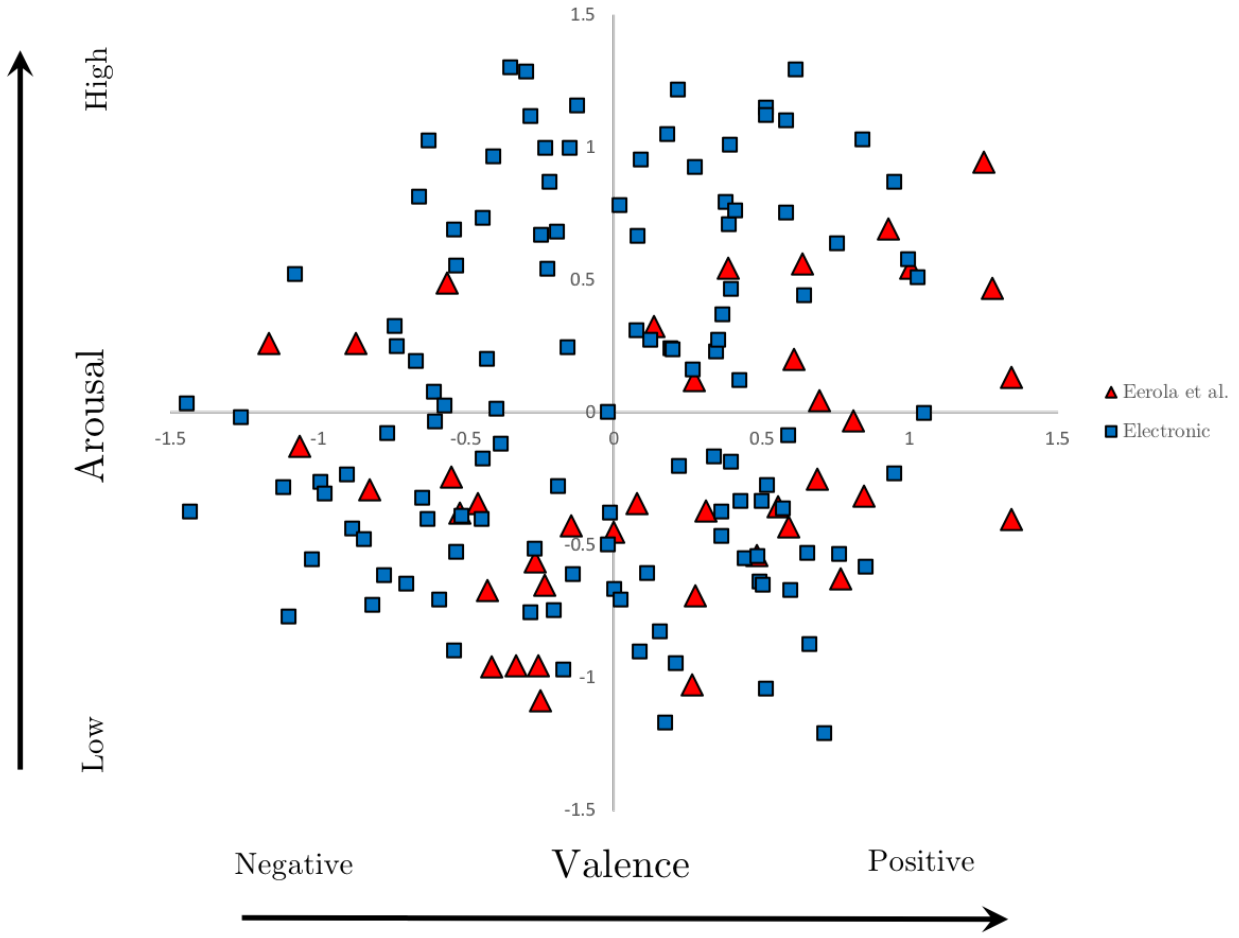

Figure 3.2. This illustrates the range of ratings for music excerpts collected in this study across perceived arousal and valence space. It compares ratings of electronic-based excerpts selected by experts with ratings on excerpts originally selected in Eerola et al. (2009). As illustrated, our excerpts cover a similarly large space as orchestral musical selections described in previous work.

3.4 Interim Summary

The objective of the first phase of this study is to produce a stimulus set based on modern electronic film music. I assembled a panel of experts to preselect music excerpts, then I had the music excerpts rated by naive participants on dimensional emotion scales of valence and arousal. Analysis of these ratings show that the electronic music excerpts both cover a similarly large emotional range of the orchestral stimulus set produced by Eerola et al. (2009), thus showing that electronic music can induce the full range of

affective experience as demonstrated previously with orchestral music. Reliability of the ratings are accessed by comparing ratings of 40 excerpts originally collected in Eerola et al. (2009) with ratings on the same excerpts in this study and the results presented in Table 3.4 show that the ratings between the two studies on the same excerpts are consistent.

3.5 Method for Performing Musical Feature Analysis

The methodology of the next phase of this study consists of performing feature extraction, feature selection, prediction, and evaluation components in order to explore the potential for developing a personalised music affect induction system that can predict emotional responses of individuals based on qualities of the music presented to them. The Essentia library (Bogdanov et al., 2013) is used to extract 443 spectral, rhythmic, and tonal features from each excerpt of music and to compute summary statistics for each feature: the minimum, maximum, median, mean, variance, mean of the derivative, variance of the derivative, mean of the second derivative, and variance of the second derivative for each feature across the excerpts; these features are listed in Table 3.5. The features can broadly be classified as being related to rhythm, tone, and low-level features. Many of the low-level features are related to timbre, but some have no obvious musical correlate.

Table 3.5 Musical Features extracted with Essentia Toolbox

Group	Abbreviation	Features
Low-level features	L	Average Loudness Energy of the Barkbands
		Energy of the Erbbands
		Energy of the Melbands
		Dissonance
		Dynamic Complexity
		HFC (High Frequency Content)
		Pitch Salience
		Silence Rate
		Spectral Centroid
		Spectral Complexity
		Spectral Energy
		Spectral Energy Band High
		Spectral Energy Band Low
		Spectral Energy Band Middle High
		Spectral Energy Band Middle Low
		Zero Crossing Rate
		GFCC (Gammatone Feature Cepstral Coefficients)
		MFCC (Mel-Frequency Cepstral Coefficients)
Rhythm features	R	Beats Loudness
		Beats Loudness Band Ratio
		BPM (The mean of the most salient tempo)
		BPM Histogram
		Onset Rate
Tonal features	T	Chords Changes Rate
		Chords Number Rate
		Chords Strength
		Key Strength
		Chords Histogram
		HPCP (Harmonic Pitch Class Profile)

The next step is to reduce the dimensionality of the feature vector. There are three motivations for this process. First, with too many features (greater than 10% of sample size) the regression algorithms will likely overfit the data. Second, the features often contain a large amount of redundant information. Finally, there is often no direct correspondence between individual low-level music features and perceptual music qualities

– which means drawing a conceptual relationship between some low-level features and emotional responses would be very difficult. Grouping these features into factors that map on to psychological constructs clarifies the contribution of some of these low-level features to emotional response.

For this task, I implemented factor analysis with oblimin rotation. An oblique rotation is chosen because it allows factors to be correlated with one another - this makes sense in the context of music features because musical effects are often composed of a combination of features and the effects of combined features on emotional perception are additive (Eerola et al., 2013). The number of factors generated are chosen through parallel analysis, and of those created, those that correlate with emotional responses are kept for further analysis. I then use Random Forest feature selection to a) identify which factors contribute to emotional response, and to b) examine their partial dependency plots and explain in which ways they are contributing to emotional responses.

For developing a predictive model for music emotion, three models are evaluated as potential solutions: Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest Regression. MLR is a standard regression algorithm that assumes linear effects of features and is often used as a baseline model to which other models can be compared, whereas SVR and Random Forest regression are able to approximate non-linear functions and are more suitable when relationships may not necessarily be linear. The SVR accomplishes non-linearity through the use of a kernel function, and it optimises the generalization bounds for regression through a loss function that is used to weight the actual error of the point with respect to the distance from the correct prediction. The SVR available in Matlab R2016a is used for this. Random Forest is a bagging technique comprised of a collection of decision trees; each node of the tree takes an input variable and selects a sub-branch based on the node criteria. This input is passed down the tree to a leaf that makes a decision on the output. In Random Forest,

all trees provide a vote and the result with the highest number of votes wins. Random Forest is amongst the most popular and successful algorithms in use currently.

For evaluation of the models' performances, ten iterations of 10-fold cross-validation are performed on each model. The performances of regression models are then compared using Wilcoxon signed rank test.

3.6 Results of Music Feature Analysis

3.6.1 Selecting Factors

Parallel analysis, performed using the psych package in R, determined that from 443 independent musical features, 15 is the optimal number of factors. The R package uses Ordinary Least Squares (OLS) to find the minimum residual (minres or MR) solution. A scree plot from the parallel analysis is shown in Figure 3.3. A table of factor loadings for each music feature is presented in Appendix B.

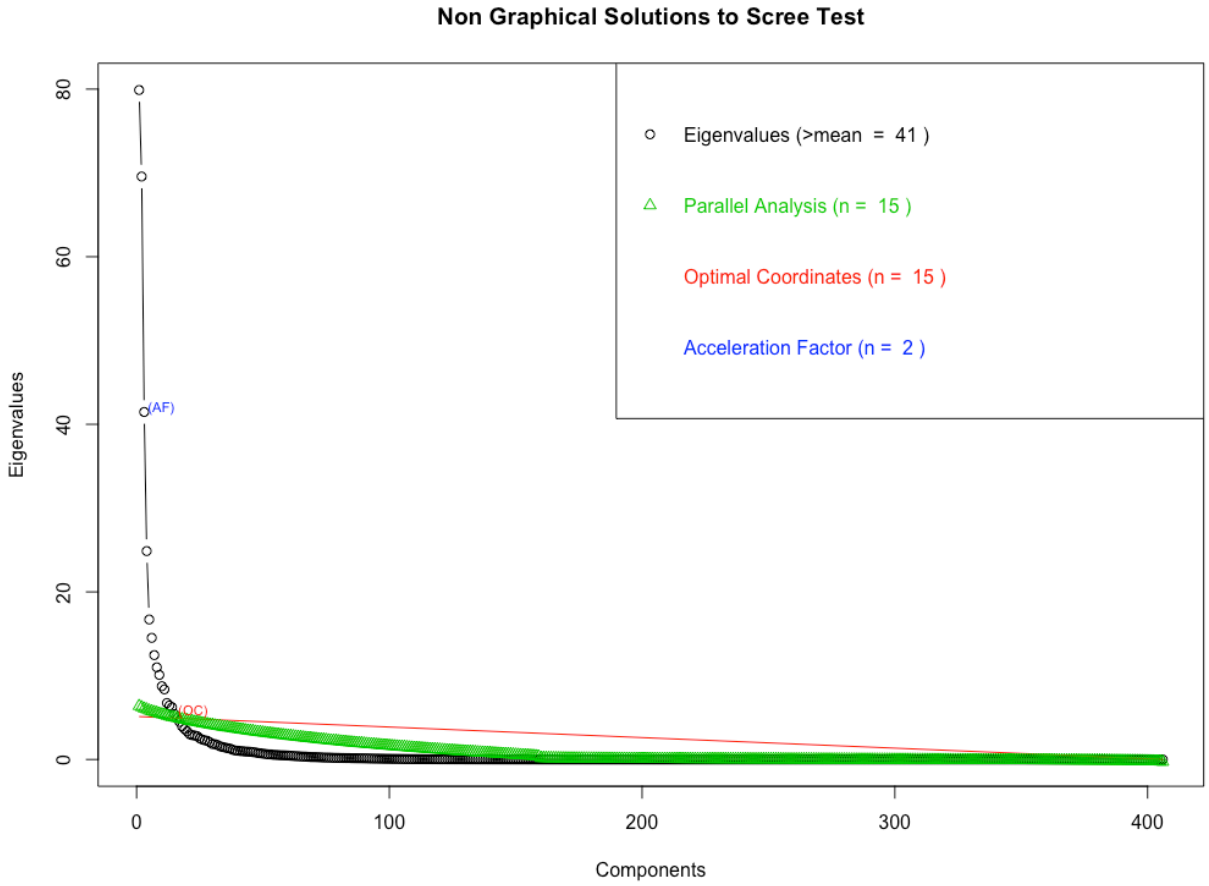

Figure 3.3. Scree plot of parallel analysis which suggest the extraction of 15 factors. The acceleration factor (AF) corresponds to the second derivative of the plot, and thus identifies where the 'elbow' of the plot occurs. The optimal coordinates (OC) corresponds to an extrapolation of the preceding eigenvalue by a regression line between the eigenvalue coordinates and the last eigenvalue coordinates.[5]

3.6.2 Understanding which factors contribute to emotional response

I perform Random Forest feature selection to identify factors that contribute to the emotional ratings, where the mean affective ratings for felt arousal and felt valence are set as outcome variables and the factor scores for each music excerpt are set as predictor variables. Felt (as opposed to perceived) ratings are chosen because the goal is to predict emotional response to better induce affect in subjects. Variable importance is

[5] nFactors R Package: https://cran.r-project.org/web/packages/nFactors/index.html

determined by incremental node impurity of variables, which in terms of regression forests refers to the increase in residual sum of squares if a variable is randomly permuted.

From examining the node impurity plot in Figure 3.4, I can observe that four factors are contributing to the arousal response to music excerpts. Factor MR14 is chosen as the cut-off because its impurity is almost a factor of two greater than MR9, and factors following MR9 make no significant difference to node impurity. The four factors chosen for arousal consist of three factors that appear to represent timbral qualities of the music, and one (MR14) that has a rhythmic component to it. This is not surprising, as both timbre and rhythm are consistently linked to the energy or intensity of the music in the literature (Gabrielsson & Lindström, 2001; Lu, Liu, & Zhang, 2006). The partial dependency plot in Figure 3.5 shows these factors to have a monotonic relationship with arousal ratings – an increasing relationship with factors MR1, MR11, and MR14, and a decreasing relationship with MR2. MR2 is composed of factors related to spectral flatness, entropy, and kurtosis, which would indicate that it is likely capturing a timbral quality related to the level of smoothness (vs. spikiness) in the timbre. Its inverse relationship with arousal would indicate that as the timbre becomes more spiked, emotional arousal increases. Also included in this factor is the dissonance feature, which has a negative impact on the factor, and thus an increase in dissonance is related to an increase in arousal. MR1 is constructed from higher order moments of spectral features, which indicates that this factor represents articulatory qualities in timbre, or Attack Decay Sustain and Release (ADSR). As articulations become more pointed, punchy, or choppy, arousal is likely to increase. This is reflected by the monotonic increasing relationship shown in the figure. MR11 is comprised of moments taken from spectral contrasts coefficients, and would seem to indicate a component of noise within the timbre. As shown in the figure, as noisiness in the timbre increases, so too does the experience of arousal. Lastly, MR14 is composed of silence rate, spectral RMS, spectral energy, and spectral flux; and represents a rhythmic

component. As demonstrated in the partial dependency plot, as this rhythmic component increases, so does the experience of arousal in participants.

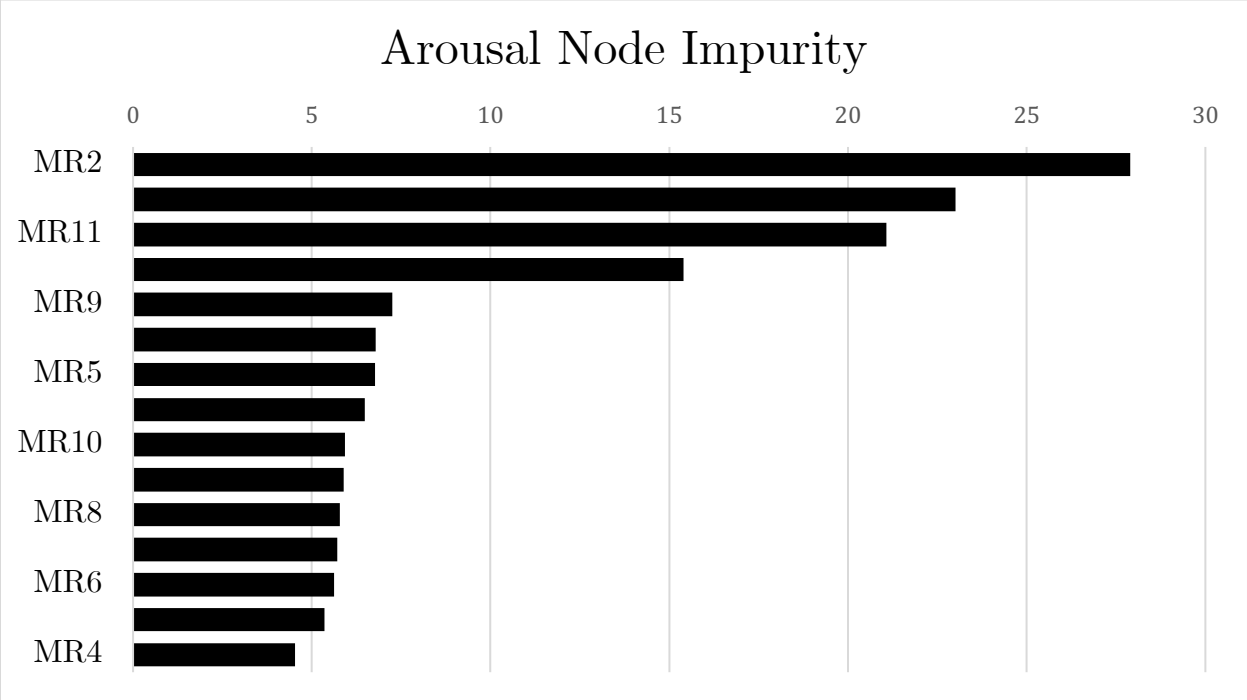

Figure 3.4. Graph of Node Impurity in factor selection for arousal as determined through Random Forest. The drop off after MR14 suggest that 4 factors are sufficient.

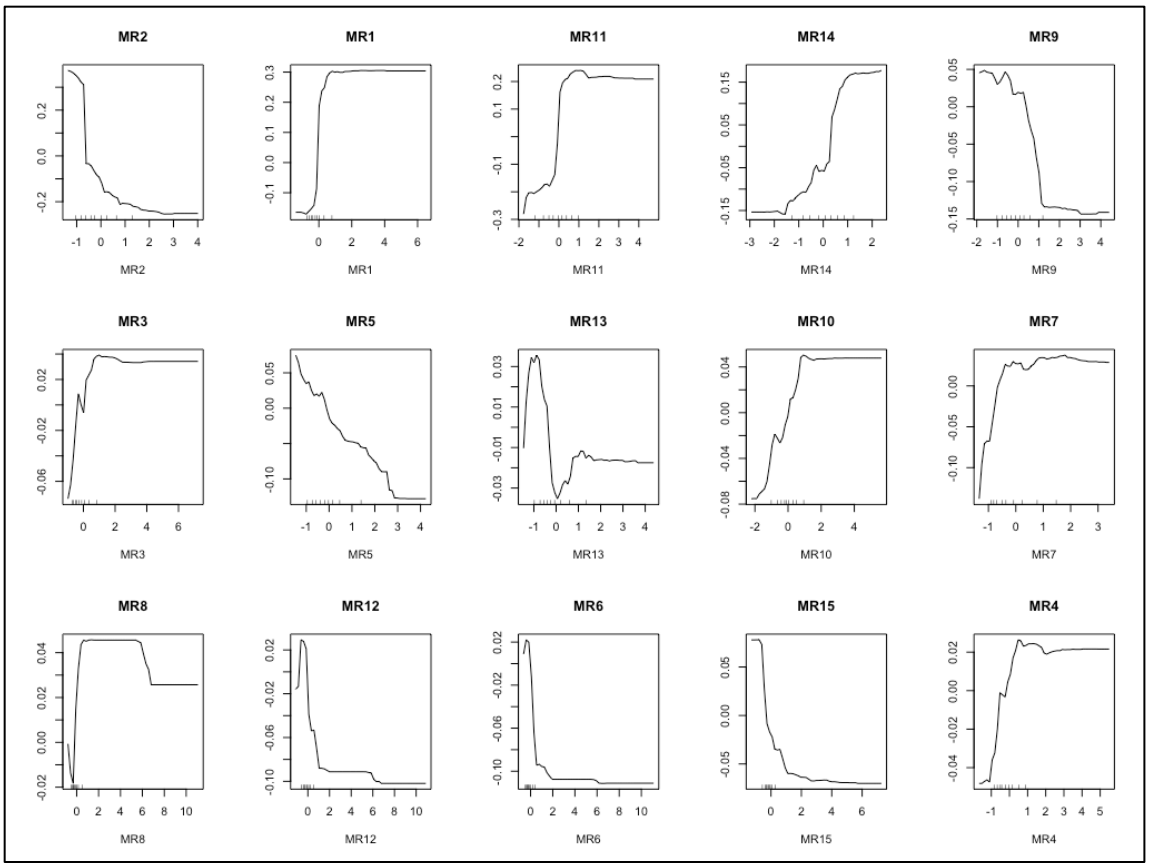

Figure 3.5. Partial dependency plot of factors and arousal. The first four plots from the top left indicate the factors chosen to contribute to arousal. The y-axis represents factor score whereas the x-axis represents affective ratings.

Examining the node impurity plot for valence, in Figure 3.6 reveals that two factors primarily contribute to the affective valence responses, MR4 and MR2, both of which are composed of timbral variables. Valence has often been associated with tonality and timbre; typically, timbres with higher numbers of harmonics and dissonant tonalities have been associated with negative valence (Blood, Zatorre, Bermudez, & Evans, 1999; Koelsch, Fritz, v. Cramon, Müller, & Friederici, 2006). MR4 has a monotonically increasing relationship with valence responses and is comprised entirely of spectral contrast and spectral valley features which indicates that as timbre becomes more dynamic, valence response increases – that is, emotional response becomes more positive. As with arousal, MR2 has a monotonically decreasing relationship - as timbre becomes more pronounced valence becomes more negative.

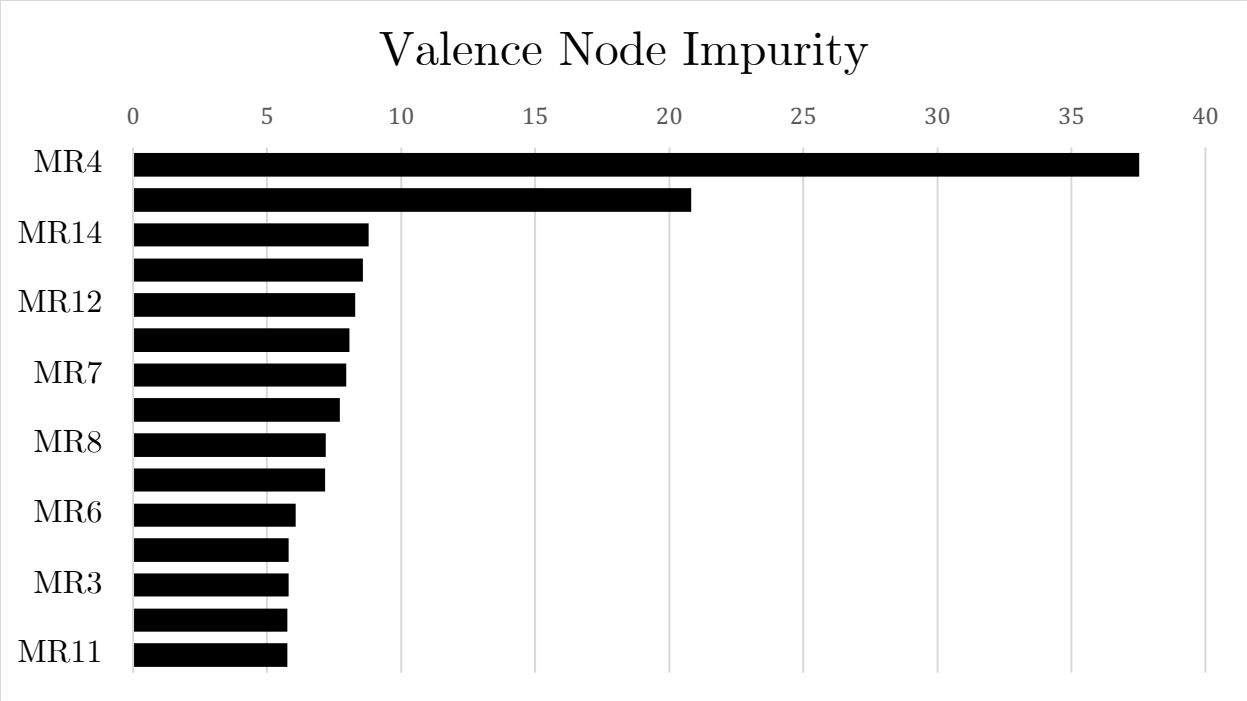

Figure 3.6. Graph of Node Impurity in factor selection for valence as determined through Random Forest. The drop off after MR2 suggest that 2 factors are sufficient.

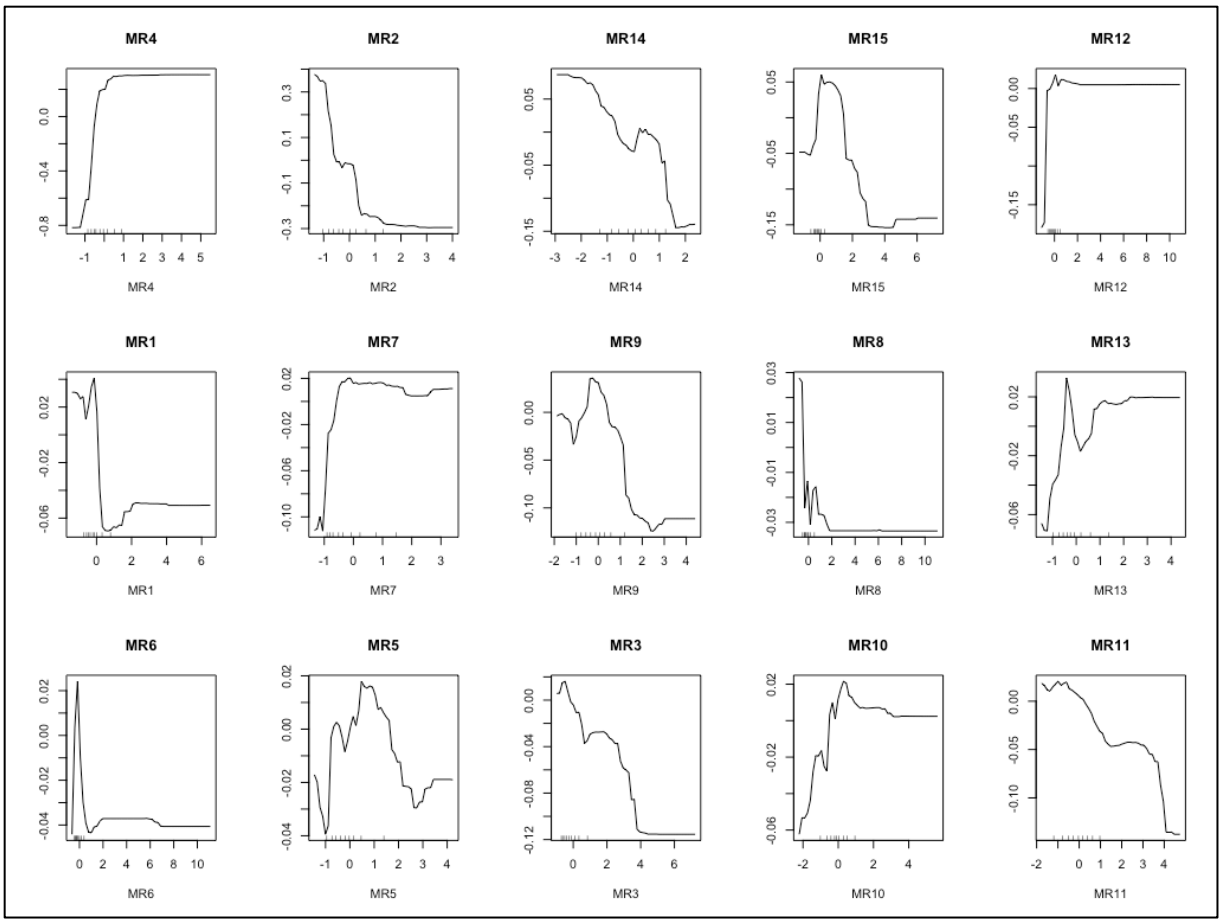

Figure 3.7. Partial Dependency plot of Factors and valence. The first 2 plots indicate the factors chosen to contribute to valence.

3.6.3 Using factors to predict average emotional responses

Once contributing factors are identified, I want to determine if these factors could be used to form predictive models of emotional responses. In order to accomplish this, I evaluate three machine learning models that are commonly used in regression problems: Linear Regression, Random Forest Regression, and Support Vector Regression – these models provide an understanding of the predictive power of the musical factors. I perform the evaluation using ten iterations of 10-fold cross-validation on the machine learning algorithms in Matlab2016. The performance of these models is shown in Figure 3.8 and Figure 3.9, and are determined to be equivalent in predictive value through a Wilcoxon comparison (used to detect significant differences). The figures indicate that it is indeed possible to use these basic models and musical factors to predict averaged emotional

ratings on ratings scales of 1 to 9 within a deviation of about 0.88 for arousal and about 1 for valence, and that factors can be extracted from the musical excerpts that can form predictors for emotional responses. This indicates that it may be possible to use musical features to create more fine-grained personal affective induction models in future studies.

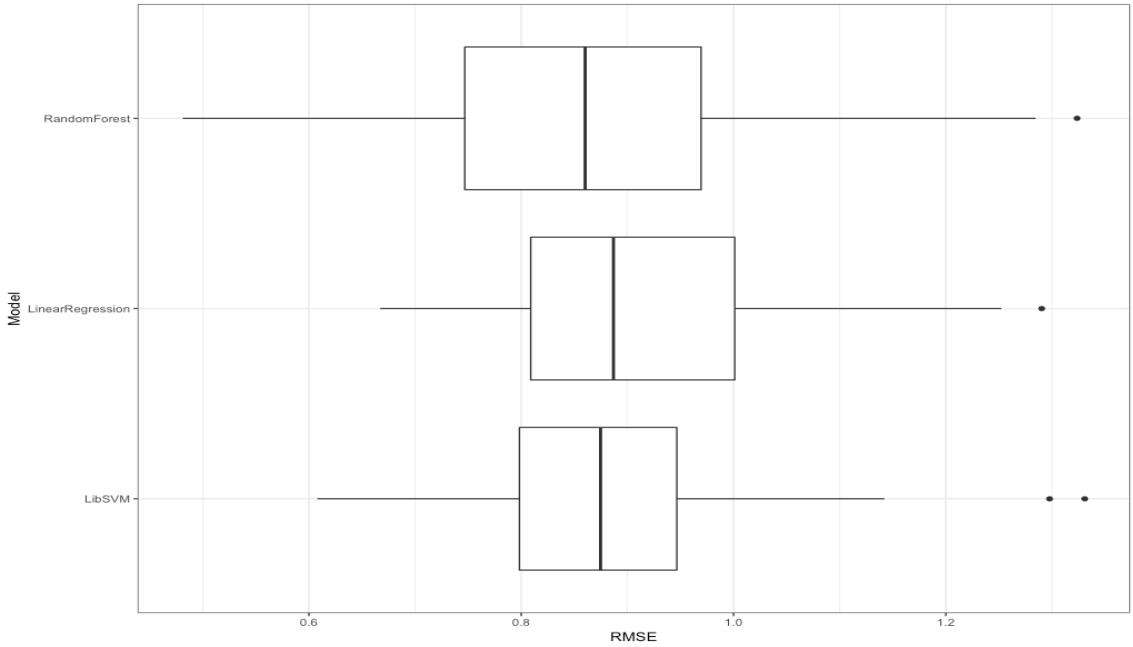

Figure 3.8 Ten iterations of 10-fold cross-validation is performed on 3 standard machine learning models to determine the predictive ability of the musical factors. This figure shows the performance of these algorithms in predicting arousal ratings for music excerpts. There was no significant difference in performance between algorithms.

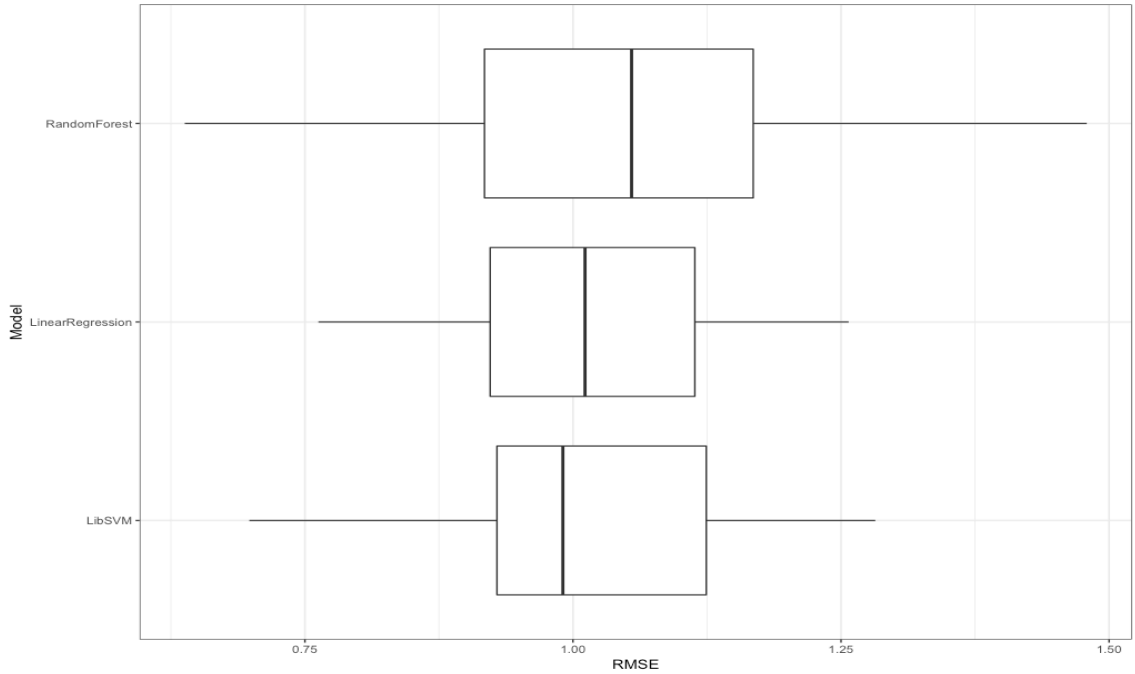

Figure 3.9 Ten iterations of 10-fold cross-validation is performed on 3 standard machine learning models to determine the predictive ability of the musical factors. This figure shows the performance of algorithms in predicting valence ratings for music excerpts. There was no significant difference in performance between algorithms.

3.7 Chapter Conclusion

This study is designed to identify a set of electronic music that could be used to effectively induce affective responses in individuals, to validate that these excerpts could be as effective in inducing emotional responses as commonly used orchestral excerpts, to identify components of the music that contribute to the emotional responses, and to validate that these components could be used to predict the averaged emotional experience induced by the music excerpts. This study shows that electronic music can effectively induce self-reported emotion in individuals. The felt emotional responses to the music excerpts presented covered the majority of the two-dimensional affective space and a large emotional range, similar to that of the orchestral excerpts taken from Eerola et al. (2009). This chapter highlights key factors in music that induce affective responses and demonstrates the ability of these factors to predict general affective ratings of the

electronic music stimulus set. This is an important validation step towards the development of an individualised affect induction system because it shows that not only is electronic music effective in inducing affect in individuals, but also that specific music factors can be extracted to develop predictive affect models.

Chapter 4 Predicting Individual Emotional Responses

4.1 Introduction

While it has been shown, on average, that the manipulation of certain music components affect the self-reported emotional responses to music excerpts (see Section 3.6.3), still not much research has been conducted to understand how musical components affect emotion at the individual level. Individual differences such as people's personalities, social influences, musical listening history, and preferences can all factor into an individual's emotional experience when listening to an excerpt of music. The necessity of understanding and accounting for these factors in affect induction has been made clear in the literature (Juslin & Västfjäll, 2008), and failing to account for them can lead to inconsistencies in study results, failure to replicate in psychological studies, and the inability to systematically control variables. The ability to better explain the relationship between music and emotion at an individual level, will not only improve the psychological research, but also has implications for applied fields. For example, this understanding could expand the possibilities of music therapy, by giving music therapists the ability to more precisely identify emotional music for a particular patient. In entertainment applications, such as film and videogames, this research could also give the creators the

ability to create more personalised and adaptive musical experience, allowing the game or content to respond to the human's own emotional response.

4.2 Chapter Goals/Objectives

In this chapter, I lay the foundation for developing a personalised emotion prediction system. The aim of this system is to predict an individual's emotional response to novel excerpts of music (i.e. excerpts they haven't listened to), based on their emotional rating responses to music excerpts that they have already listened to. This research consists of three steps. In Section 4.3 I highlight the differing emotional responses to music, by comparing responses to similar emotional music features at the cohort level (i.e. grouping individuals who exhibit similar emotional responses). This exploratory step provides a coarse view of the individual differences that exist within emotional responses to musical features, and highlights the need for more personalised emotion induction in psychological research. In Section 4.4, I implement and evaluate several standard approaches to collaborative and content-based recommendation systems for developing the personalised emotion prediction system. These recommender systems are designed to produce personalised experiences for the user, and serve as baseline measures of the potential success in developing a personalised music emotion induction system. Section 4.5 extends existing recommender research with the development and evaluation of a novel content-based Convolutional-Recurrent Neural Network (CB-CRNN).

4.3 Discovering and comparing emotional responses between similar individuals

Although there are statistically significant relationships between specific musical features and emotional responses (see Chapter 3), there are also clear individual differences in the mapping between music and emotion. Many factors contribute to an individual's emotional responses while listening to music, and these factors often result in different people experiencing different emotional responses when listening to the same musical

features. In this study, I highlight these differences at a coarse level through the extraction of 'emotional cohorts', or groups of individuals that display similar emotional responses to musical features.

4.3.1 Procedure

For this study, I used the ratings data for the electronic music stimulus set developed in Chapter 3. However, given that the focus in Chapter 3 was on the identification of musical features with emotional qualities, relatively few individuals rated each specific excerpt. This presented a challenge when trying to use the data from this same stimulus set for identifying individual differences in emotional response and extracting cohorts. The study in Chapter 3 was designed with four groups of about 15 raters per musical excerpt with no crossover participants between groups, meaning no participants in different groups listened to the same music excerpt. Thus, to identify cohorts of people with similar emotional response to similar music features, I first combined (through clustering) all music excerpts with similar emotional music features into a single entity (i.e. cluster). I then averaged each individual's arousal or valence rating for each excerpt they rated within a music cluster, and considered that to be a measure of an individual's emotional response to the music features of that cluster. In this way, I was able to construct cohorts of individuals with similar emotional responses to music regardless of whether they listened to and rated the same excerpts or not.

4.3.2 Method

4.3.2.1 Music clustering

The first step was to identify similar low-level features that define *clusters* (groups) of musical excerpts with common emotional properties. K-means is commonly used as a method for grouping musical excerpts according to low-level features, and has been applied in the development of several music recommendation solutions (McFee, Barrington, & Lanckriet, 2012; Pauws & Eggen, 2002; Schedl, Knees, McFee, Bogdanov, & Kaminskas,

2015). As such, I used k-means clustering on the music excerpts from Section 3.3, with musical feature vectors constructed of the top 30 most relevant musical features from both the arousal and valence conditions (identified in Section 4.4.2.3).

4.3.2.2 Cohort development

Similarly, the k-means method was also used to develop cohorts of individuals (Dakhel & Mahdavi, 2011; Ungar & Foster, 1998). After averaging each user's ratings for musical excerpts within each cluster (i.e. groups of music with similar affective features), I applied k-means again to identify cohorts of individuals who had similar emotional responses. This revealed connections between sets of affective musical features and cohorts of individuals with similar emotional responses.

4.3.2.3 Gap analysis for determining the optimal k in k-means

As a divisive clustering technique (i.e. a technique that splits data points into exclusive clusters based on their distances from neighbouring points), it is necessary to determine the optimal k number of clusters to divide the data into. To determine the optimal k, I used the gap statistic, which is a standard heuristic for determining the optimal number of clusters for each analysis (Tibshirani, Walther, & Hastie, 2001). The gap statistics are computed by running the k-means clustering algorithm i times and calculating the difference between the log mean dispersion of a bootstrapped sample of a reference distribution,

$$E_n^*\{log(W_k)\}, \qquad (4.1)$$

and the log mean dispersion of the original dataset,

$$log(W_k). \qquad (4.2)$$

Dispersion is defined as the sum of all point distances from the cluster mean

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r, \qquad (4.3)$$

where n_r is the number of data points in cluster r, and D_r is the sum of the pairwise distances for all points in cluster r. The gap is thus defined by the equation

$$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k), \quad (4.4)$$

where n represents the sample size, k is the number of clusters being evaluated, and W_k is the within-cluster dispersion.

For the gap analysis system, 500 iterations of bootstrapping were implemented using the *factoextra* package in R, and the *firstmax* method. The firstmax method gives the location of the first local maximum, which is then used as the optimal number of clusters.

4.3.2.4 Semantic annotations for cohorts and music clusters

As a further step in analysing the musical clusters, expert excerpt annotations collected in the previous study (Section 3.2.1) were used to identify the musical qualities (i.e. articulation, volume, timbre, harmony, tempo, rhythm, mode, and pitch) that characterised each emotional music cluster. While these annotations came from a relatively small sample of experts and did not fully represent all musical pieces in the clusters, they provided a broad semantic description of each cluster's musical features. These annotations assisted in understanding the make-up of music in different clusters, and helped provide a tangible understanding of how different cohorts of individuals have differing emotional experiences to the same musical features. For example, these semantic labels can help a reader distinguish the difference between clusters with fast and slow tempo, or firm and flowing rhythms. For each music excerpt cluster, the annotations are summed to show how frequently experts associated the music of that cluster with the given semantic quality.

4.3.3 Results

In this section I evaluate the results of music clustering procedure, and cohort discovery for arousal and valence conditions.

4.3.3.1 Music clustering

The first step in cohort development is identifying clusters of music excerpts with similar emotion-related music features (as determined in section 4.3.2.2). Gap analysis (shown in Figure 4.1) determined the optimal number of clusters to be four – meaning that the 160 music excerpts could be collapsed into four clusters based on the extracted emotion relevant features. The optimal number of clusters is determined by the first local maximum, however in this case the maximum spans between approximately three and five, so I chose the four-cluster solution a reasonable medium given the relatively small difference of dispersion between these three clustering solutions, they would result in clustering solutions very close in n-dimensional space with very little difference in effect. Figure 4.2 shows a plot of the first two principal components extracted for visualization of each music excerpt for three, four, and five cluster solutions. These principal components were extracted by performing Principal Component Analysis (PCA) on the 42 music excerpt feature vectors that were used for clustering. The PCA representation is a useful method for visualizing how the excerpts are similar or different in 2D space. As seen in Figure 4.2, the plot demonstrates very little separation between the three-, four-, and five-cluster solutions in the two largest component axes.

To semantically characterise each of the four music excerpt clusters, I examined the experts' annotations of the musical excerpts in each cluster. Figure 4.3 shows these annotations in a graphical representation. Each cluster is described according to the frequency of qualities in each semantic space (i.e. articulation, volume, mode, pitch, rhythm, tempo, and harmony). Cluster 1 is described by its pointed articulation, flowing rhythm, medium to slow tempo, medium-low pitch, and medium-soft volume. Also of note, many of the excerpts in Cluster 1 include clarinet, plucked strings, piano, and pitched percussion. Cluster 2 is comprised of excerpts that feature medium-high pitch, contrasting bright and dark timbres, fast tempo, firm rhythm, medium to high volume, and choppy or pumping articulation. The excerpts of Cluster 3 are mostly in minor mode,

low-pitched, dark in timbre, with dissonant harmony, and slow to medium tempo. Cluster 4 excerpts are primarily major in modality, medium-pitched, rhythmically flowing, and low to medium tempo.

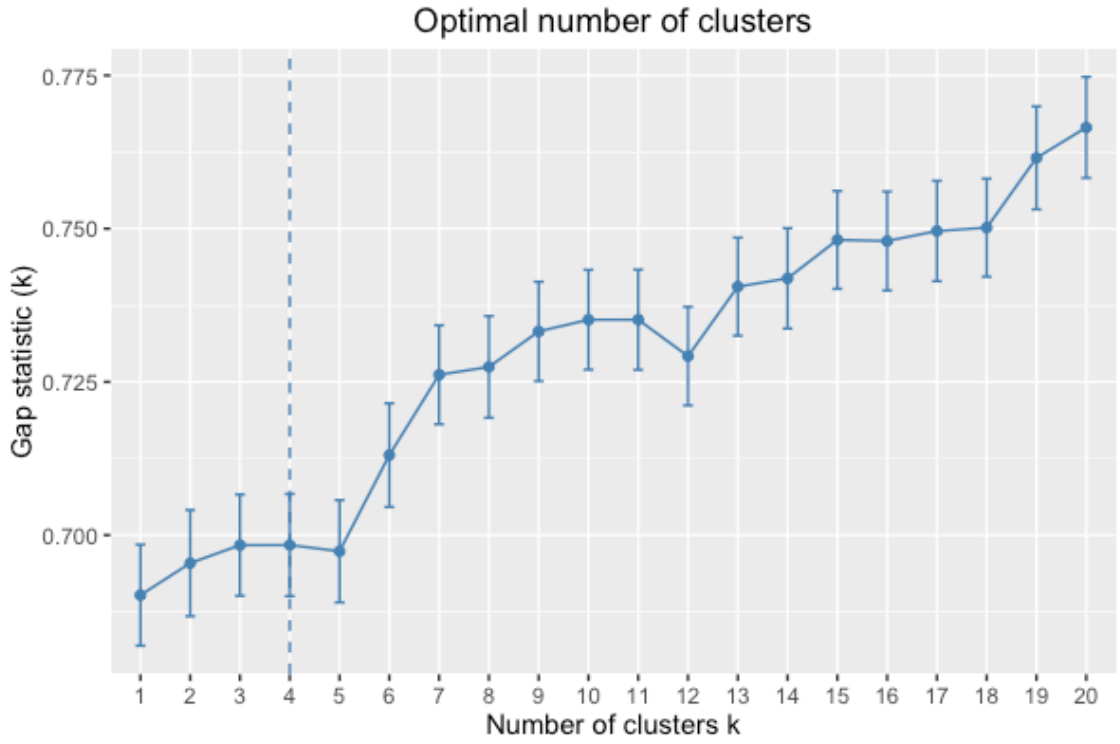

Figure 4.1. A line graph showing the results of gap analysis on the first step of clustering music excerpts, and illustrating that the optimal number of music clusters is four. The optimal number of clusters is determined by the first local maximum, which is a standard indicator used in Gap Analysis.

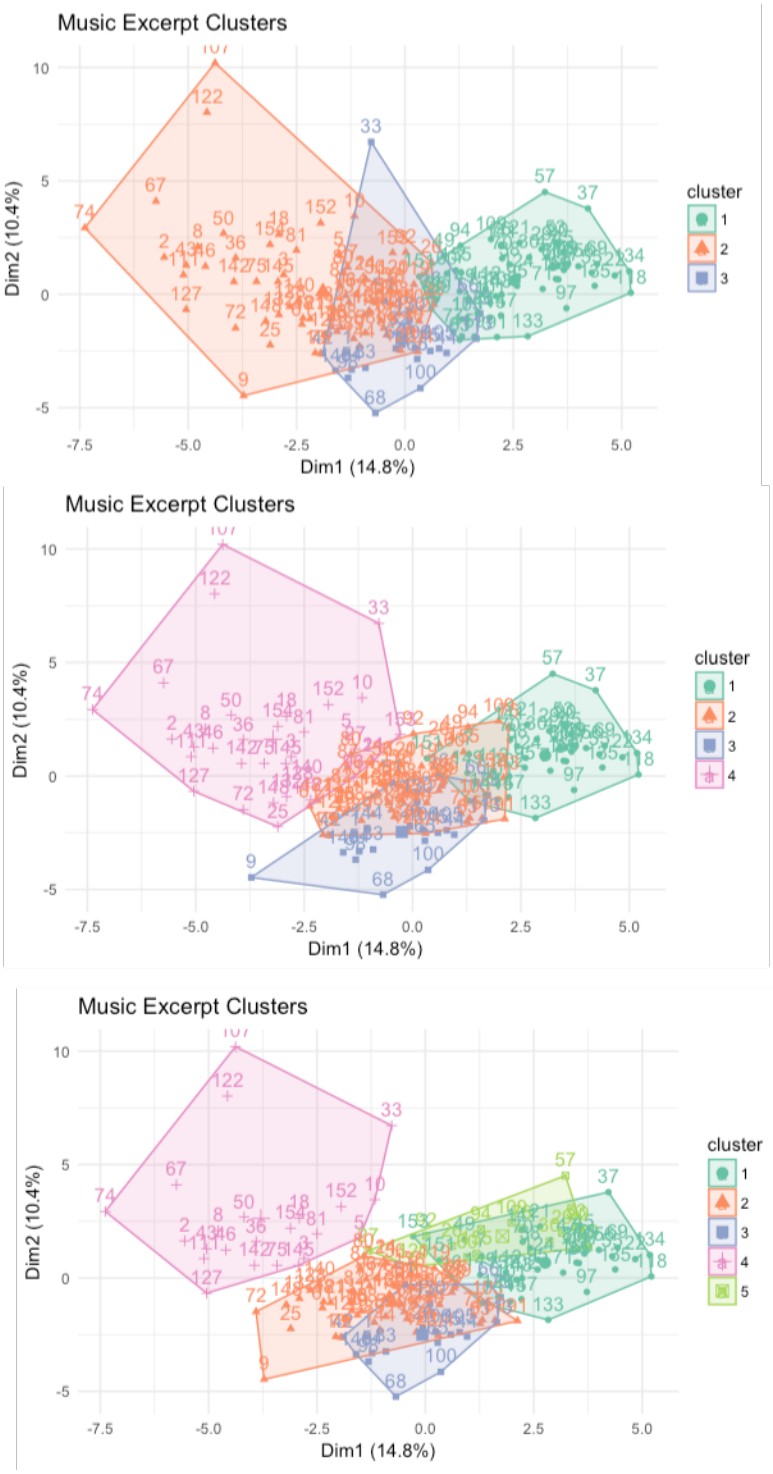

Figure 4.2. Shows the first 2 principal components plot on two-axes for three, four, and five cluster solutions respectively. The plot demonstrates very little separation between the solutions in the two largest component axes.

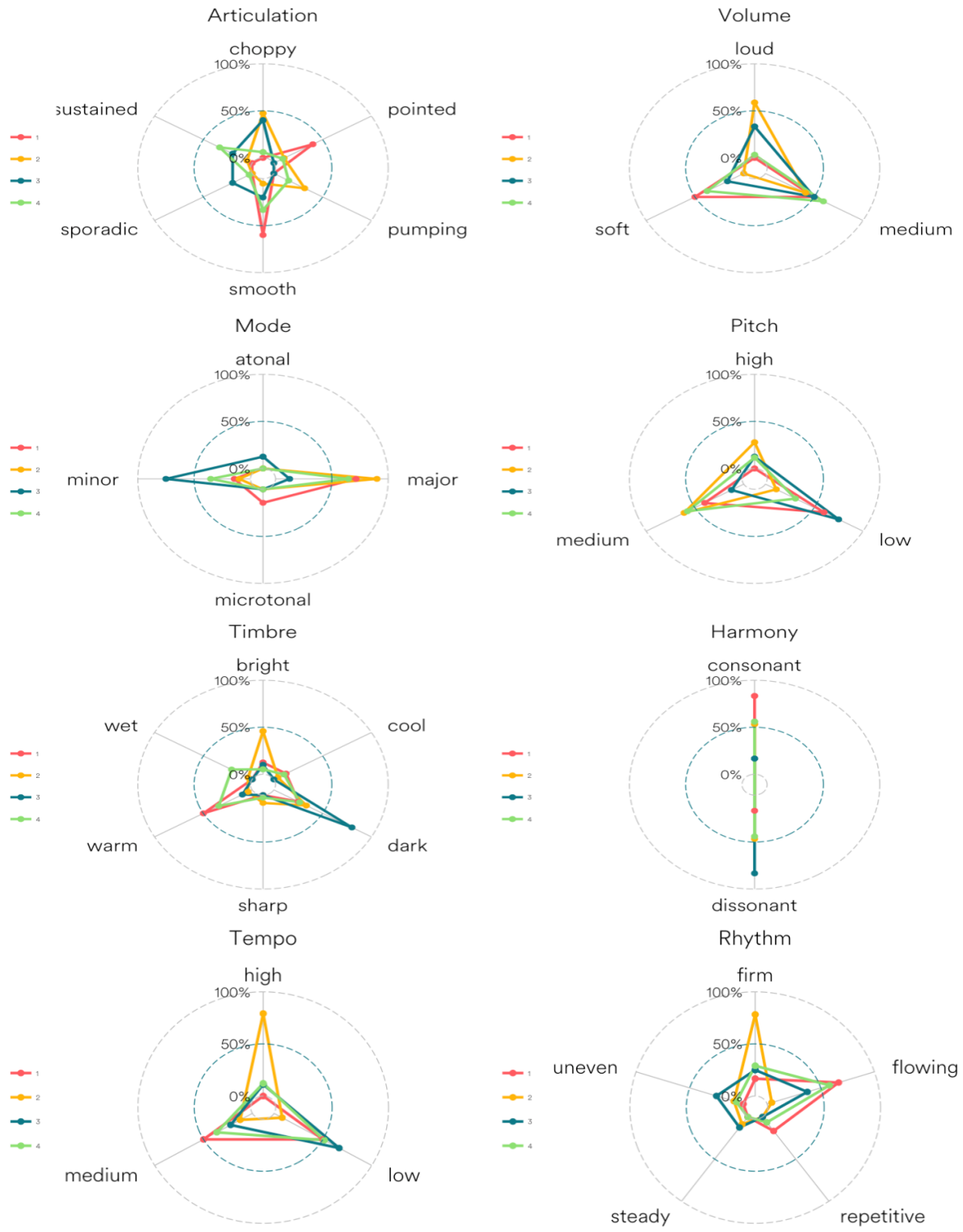

Figure 4.3. Radial plot depictions of the experts' annotations for musical excerpts, averaged across clusters. The coloured lines refer to the cluster numbers, and the percentage to the percentage of excerpts within a cluster that were associated with the semantic quality.

4.3.3.2 Arousal cohorts

After forming clusters of music with similar emotional features (as in Section 4.3.3.1), I used those clusters to identify cohorts of individuals with similar arousal and valence responses to the excerpts within a cluster. This approach segments users based on similar responses to similar content.

First, I assessed arousal cohorts. Participants' arousal ratings for excerpts lying within each music cluster were averaged, resulting in four ratings per participant (i.e. a mean rating for each cluster). These ratings were used to group (segment) participants into cohorts. As such, we can predict how any given participant will rate a novel musical excerpt based on how others in their cohort rated that music. Gap analysis, shown in Figure 4.4, determined that the optimal number of participant cohorts for arousal was six. The cohorts are visualized in Figure 4.5, which shows a clear separation on the first two principal components. The optimal number of cohorts was determined by the first maximum in standard error.

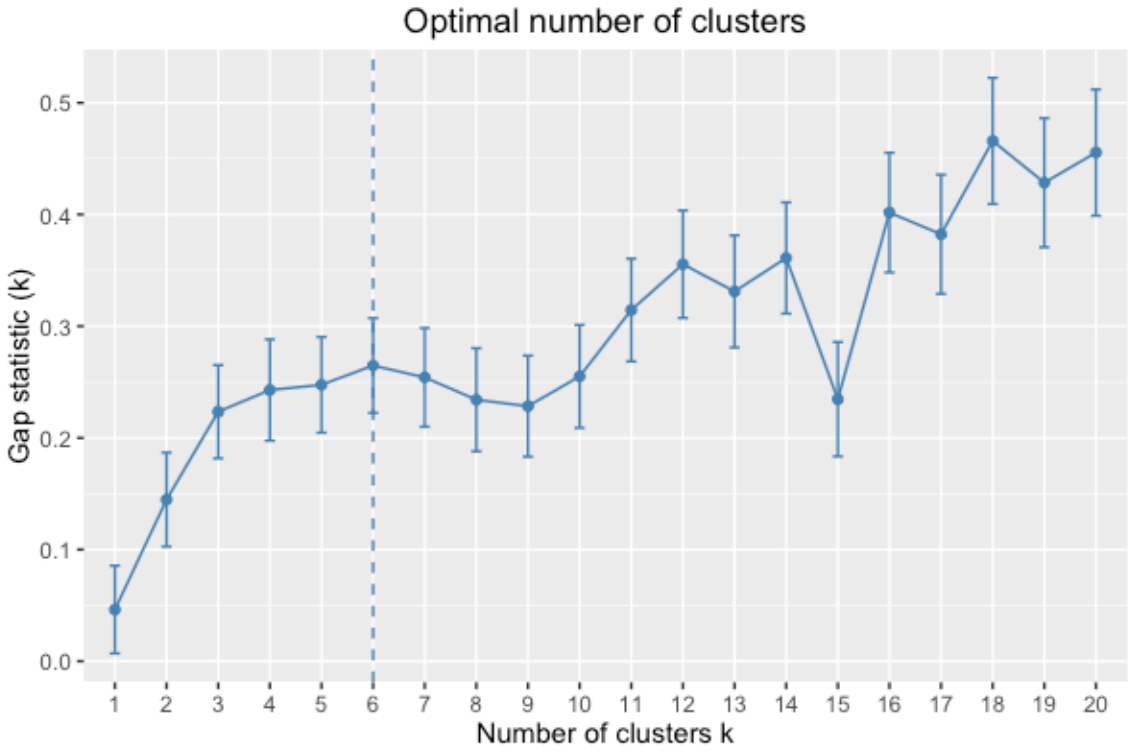

Figure 4.4. A line graph, showing the results of gap analysis for arousal cohorts, illustrating that the optimal number of cohorts is 6. The optimal number of cohorts is determined by the first local maximum, which is a standard indicator used in gap analysis.

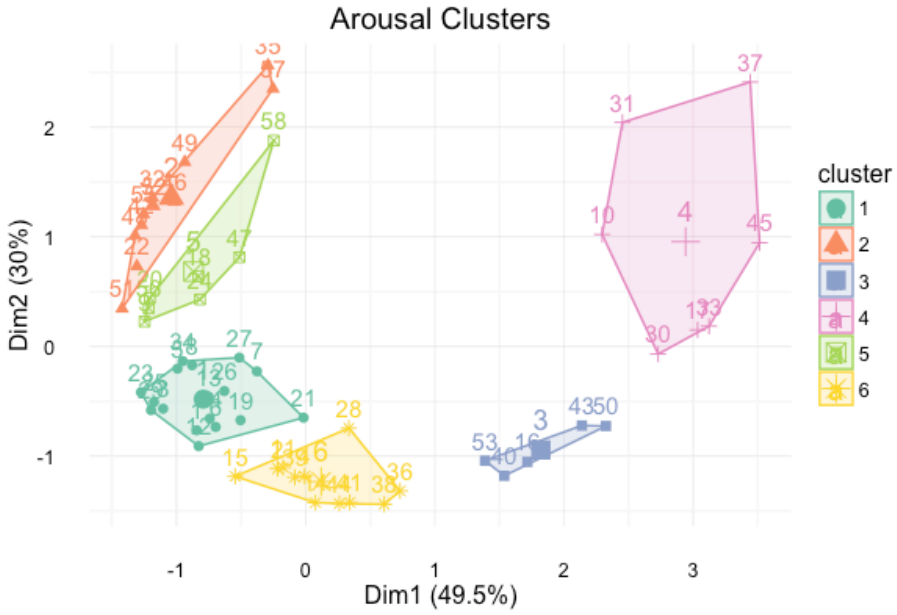

Figure 4.5. Shows the first two principal components plotted on two-axes, with a clear separation between the arousal cohorts (each point is a person).

The next step is compare the emotional responses of cohorts to observe differences. As shown in Table 4.1, cohorts of individuals can have contrasting responses to similar music features. For example, participants in Arousal Cohort 5 have contrasting emotional responses to those in Cohort 4 across all music excerpt clusters. Furthermore, the differences between responses are intricate, where two cohorts may respond similarly on certain music clusters, but can disagree on others. For example, participants in Arousal Cohorts 1 and 6 have similar arousal responses to excerpts in Clusters 1, 2, and 4; but differ in the arousal induced by Cluster 3. The variable nature of individuals' arousal responses to musical features, as demonstrated through cluster analysis on these small sample sizes, illustrates the need for a more personalised emotional stimulus recommendation. As shown in Table 4.1, participants fitting into different cohorts (rows) could have contrasting responses to stimuli within the same cluster (columns), making consistency in implementation difficult without a personalised recommendation approach.

Table 4.1 Average normalised arousal ratings as rated by each arousal cohort (rows) for excerpts from the different music clusters (columns).

	Cluster			
Cohort	1	2	3	4
1	-0.973	1.228	0.201	-0.456
2	-0.870	0.205	1.244	-0.579
3	0.539	0.881	-1.296	-0.124
4	1.062	-0.385	-0.922	0.245
5	-1.290	0.683	0.781	-0.174
6	-0.400	1.419	-0.607	-0.412

4.3.3.3 Valence cohorts

The process described above for identifying groups of individuals (cohorts) that respond similarly to musical features was then similarly repeated for valence responses. Participants' valence ratings for each of the excerpts lying within each cluster were averaged, resulting in one rating for each of the four clusters, for each participant. These ratings were used to group participants into cohorts and the gap analysis, shown in Figure

4.6, determined that the optimal number of participant cohorts for valence was four. Again, the optimal number of cohorts was determined by the first local maximum. A plot of the first two principal components (see Figure 4.7). shows a clear separation between valence cohorts. As in Section 4.3.3.1, several very close maxima appear together and so the median was selected.

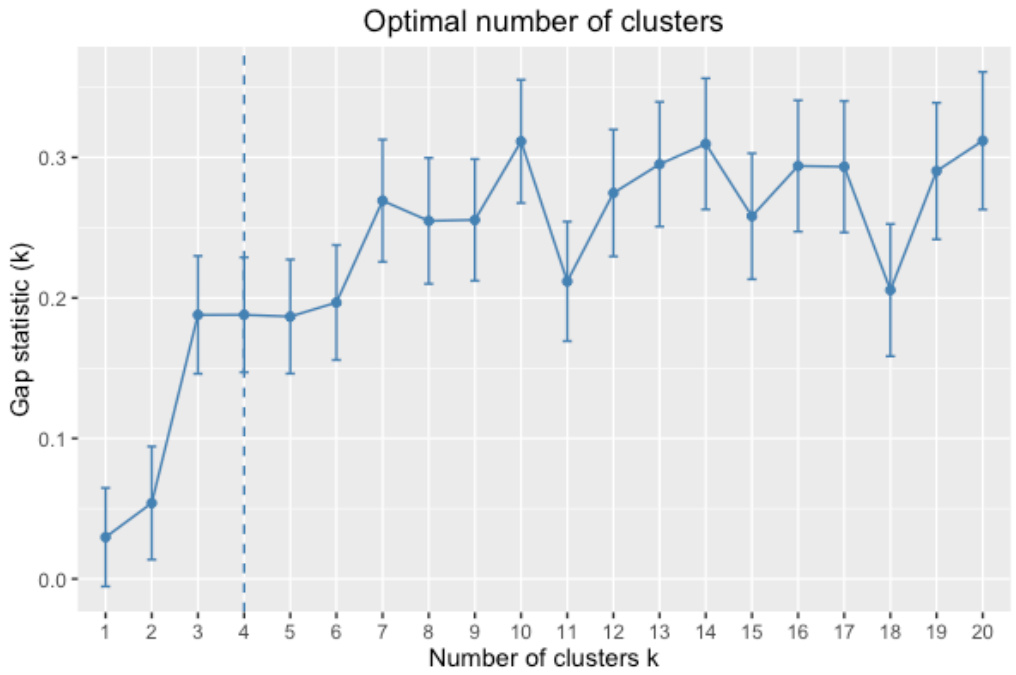

Figure 4.6 A line graph showing the results of gap analysis for valence cohorts, illustrating that the optimal number of cohorts is four. The optimal number of clusters is determined by the first local maximum, which is a standard indicator used in gap analysis.

Figure 4.7 Shows the first 2 principal components of valence cohorts plotted on two-axes (each point is a person).

As seen in Table 4.2, it appears that cohorts of individuals (rows) can have contrasting valence judgements to similar music clusters (columns). For example, while participants in Cohort 4 found excerpts of Cluster 1 to be of negative valence, participants of Cohort 1 found them to have a positive valence. As with the arousal condition (Section 4.3.3.2), it appears that individual differences can play a large role in the emotional experience of the music excerpt, so failure to account for individual differences could interfere with the ability to produce reliable experimental results in affective research. Once again, the undeniable clustering of valence cohorts was exhibited even by this small sample of individuals. This validated the need for a personalised prediction system that can intelligently recommend music in such a way that the individual's unique emotional responses are reflected.

Table 4.2. Average normalised ratings; showing how each valence cohort (rows) rated excerpts from music clusters (columns).

	Cluster			
Cohort	1	2	3	4
1	0.731	0.255	-1.294	0.309
2	0.860	-0.887	0.524	-0.496
3	-0.114	1.274	-0.768	-0.393
4	-1.068	0.743	0.141	0.184

4.4 Developing recommender systems

In Section 4.3, cohorts of individuals were extracted from the ratings collected in Chapter 3. The development of the cohorts in Sections 4.3.3.2 and 4.3.3.3 highlights, at a coarse level, that there are individual differences in the emotional responses to music, and thus supports the need for a more personalised music affect induction system for psychological research. In the next section I evaluate several existing recommendation techniques that could be used to develop a more personalised music emotion induction system for researchers to use for more controlled emotion induction.

4.4.1 Procedure

In recommender systems, content-based and collaborative filtering techniques can be used to better predict individuals' preferences and responses to novel items. I looked at several variations of both content-based filtering and collaborative filtering approaches, to evaluate their performance in making personalised recommendations of music excerpts from the electronic music stimulus set. The content-based filtering approach to recommendation engines is built on the premise that similar content will be rated similarly, so if music excerpts have similar music features, they will be rated as emotionally similar by individuals. Collaborative filtering on the other hand, does not use information about the musical features, instead assuming that individuals with similar rating patterns are likely to rate novel items similarly. As such, collaborative filtering approaches identify

different cohorts of individuals who have been found to rate items similarly, in order to predict future rating behaviour for novel items. For example, in the context of music excerpts, if two participants are in the same given cohort, and participant A has rated an excerpt that participant B has not, we can assume that participant B would give a similar rating to that provided by participant A.

4.4.2 Methods

4.4.2.1 Collaborative and content-based filtering

I evaluated five different collaborative and content-based filtering systems:

1. item average,
2. user-based collaborative filtering,
3. item-based collaborative filtering,
4. singular value decomposition approximation, and
5. distance-weighted knn algorithm (content-based).

The first method, known as the *item average* approach, is a non-personalised approach that simply takes the average rating for each excerpt and uses that average to fill in missing user ratings. This method functions as a good comparison for more personalised recommendation techniques.

The second approach, *user-based collaborative filtering*, predicts ratings based on a person's 'nearest neighbours' – that is, the n users that have ratings most similar to theirs. For this study, n was set at five and the distance function used was cosine similarity.

The *item-based collaborative filtering* approach, assumes that users will rate similar items similarly. Similarity of items in this technique is calculated based on similar item ratings, and ratings are predicted as a result of a participant's ratings on similar items.

The singular value decomposition (SVD) approximation is a popular matrix factorisation technique used in recommender systems. This collaborative filtering

technique decomposes a $m \times n$ ratings matrix (e.g. m participants, n music excerpts) into three matrices using SVD:

1. U: $m \times r$ matrix (m participants, r latent factors)
2. S: $r \times r$ diagonal matrix (strength of each latent factor)
3. V: $r \times n$ matrix (r factors, n music excerpts)

The predictions are then generated by multiplying the product matrix U and diagonal matrix S by the transpose of matrix V.

The final approach evaluated is a standard content-based filtering algorithm known as distance-weighted knn. The knn approach to content-based recommendation typically serves as a baseline for comparisons of recommendation solutions (Davidson et al., 2010; Linden, Smith, & York, 2003). The objective of the distance-weighted knn is to give the ratings of music excerpts that are nearer in n-dimensional space greater influence. Like item-based collaborative filtering, content based filtering is built on the assumption that users will rate similar items similarly. However, the similarity in content-based is computed based on an item's features (attributes) and not participants' ratings. The development of the distance-weighted knn algorithm is accomplished in four steps:

1. Extracting feature vectors that describe each item.
2. Taking the inverse of absolute difference for k nearest neighbours (k=5).
3. Dividing each inverse distance by the sum of all inverse distances (the resulting inverse distances sum to 1).
4. Multiplying each of five k-nearest neighbours' ratings by their inverse distance, and summing the result to produce the predicted rating.

Crucially, for each of the collaborative and content filtering techniques described above, I held out 5% of ratings from 30% of participants as novel test data. This allowed for the system to be tested on excerpts that were not used in training the models.

4.4.2.2 Feature extraction for content-based filtering

One of the challenges of creating a content-based recommender system for personalised music emotion induction is selecting appropriate musical features that correspond with individual differences in the experience of emotion. In the previous study (Section 3.6.3), I used factor analysis to demonstrate that, when averaging across individuals, musical features could be extracted to predict the emotional state induced by music. In the present section, for the development of a more personalised emotion prediction system, I focused on the more granular attributes that could be combined in individualised ways, utilising the features extracted in Section 3.5. These are standard musical features, many of which are used in many Musical Information Retrieval (MIR) studies on emotional identification in music.

4.4.2.3 Feature selection for content-based filtering

In general, using too many or too few features can lead to 'overfitting' or 'underfitting'. In such cases predictive models would be poor at generalising because they would either (a) be too specific to the original training data, or (b) not learn enough relevant information to form good predictions. I therefore use the ReliefF feature selection algorithm to find the most important musical features for predicting arousal and valence. As opposed to other feature selection methods such as the correlation coefficient, information gain, and signal to noise ratio, ReliefF feature selection takes feature interrelationship into account in selecting the best features (Yang & Chen, 2011). To determine the number of features to retain from the ReliefF weighting, I implement an At Most One Change (AMOC) changepoint analysis using the *ChangePoint* package in R. For this changepoint analysis I set the parameters to test for changes in the mean and variance of the sequence of ReliefF weight values, to identify any significant change in the sequences' normal distribution (Hinkley, 1970). I set the methods penalty to 'asymptotic' with a penalty value of 0.05 to ensure that changepoints are detected with 95% confidence. The benefit of this approach is that as opposed to arbitrarily selecting a number or

percentage of features to retain, the selection is made based on statistically significant changes in the ReliefF weighting distribution. The results of the changepoint analysis are shown in Figure 4.8 and Figure 4.9 for arousal and valence respectively. The figures show with 95% confidence that a changepoint in the mean and variance of the normal distribution of ReliefF weights occurs after 35 features for arousal and 32 features for valence. From this determination, I selected the 30 highest ranked musical features for arousal and valence (resulting in 42 unique features), shown in Figure 4.10 and Figure 4.11. These ReliefF features were determined by the algorithm to be more highly predictive than other individual features of emotion ratings.

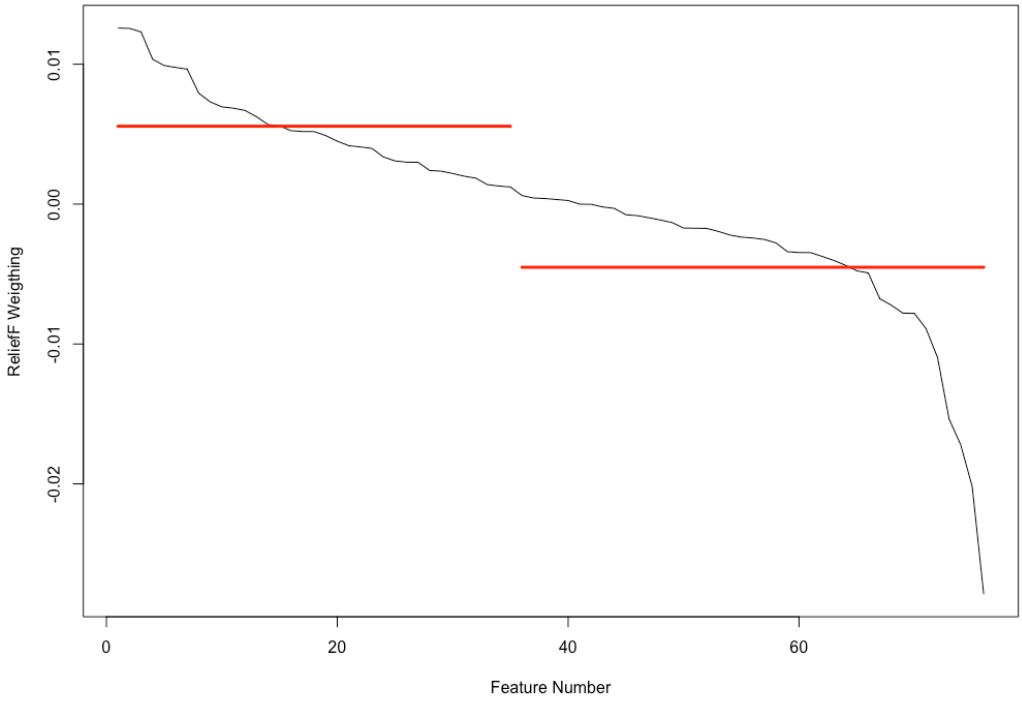

Figure 4.8. Graph showing the AMOC changepoint analysis on arousal music feature ReliefF weights. A change in the mean and variance was found to occur after 35 features, determined with 95% confidence. The red bars shows the cut-off of the first section of 35 features, and the beginning of the second section (where the mean and variance changes).

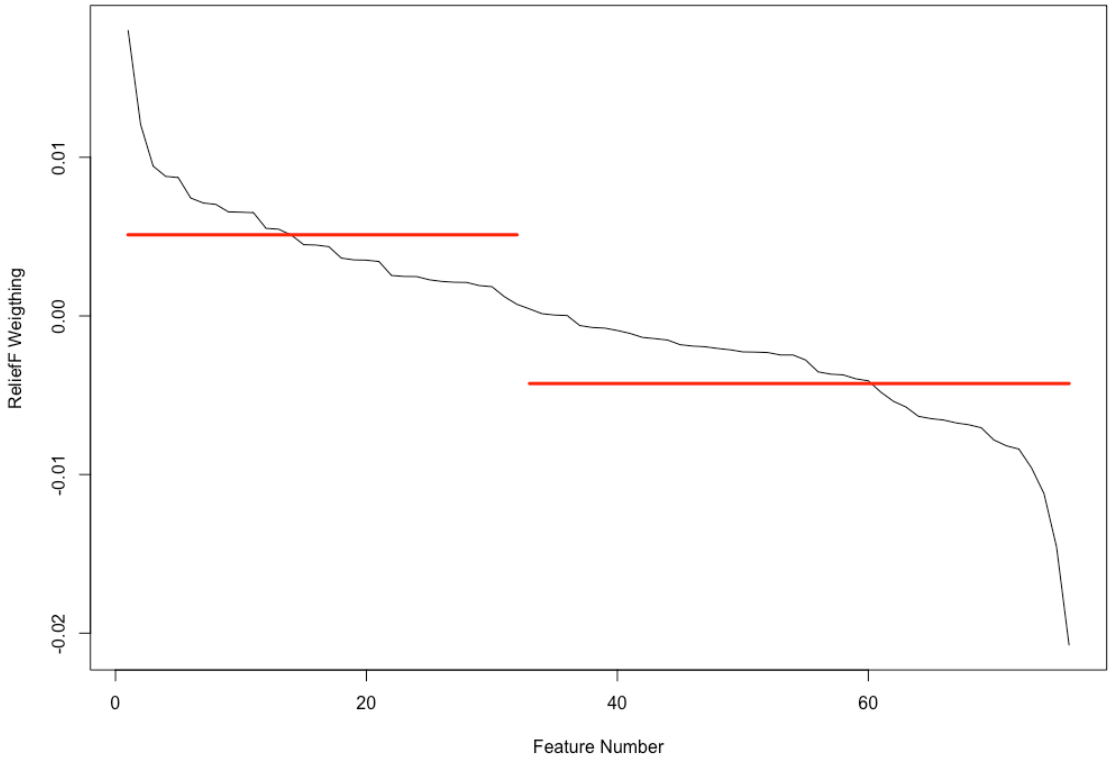

Figure 4.9 Graph showing the AMOC changepoint analysis on valence music feature ReliefF weights. A change in the mean and variance was found to occur after 32 features, determined with 95% confidence. The red bars shows the cut-off of the first section of 32 features, and the beginning of the second section (where the mean and variance changes).

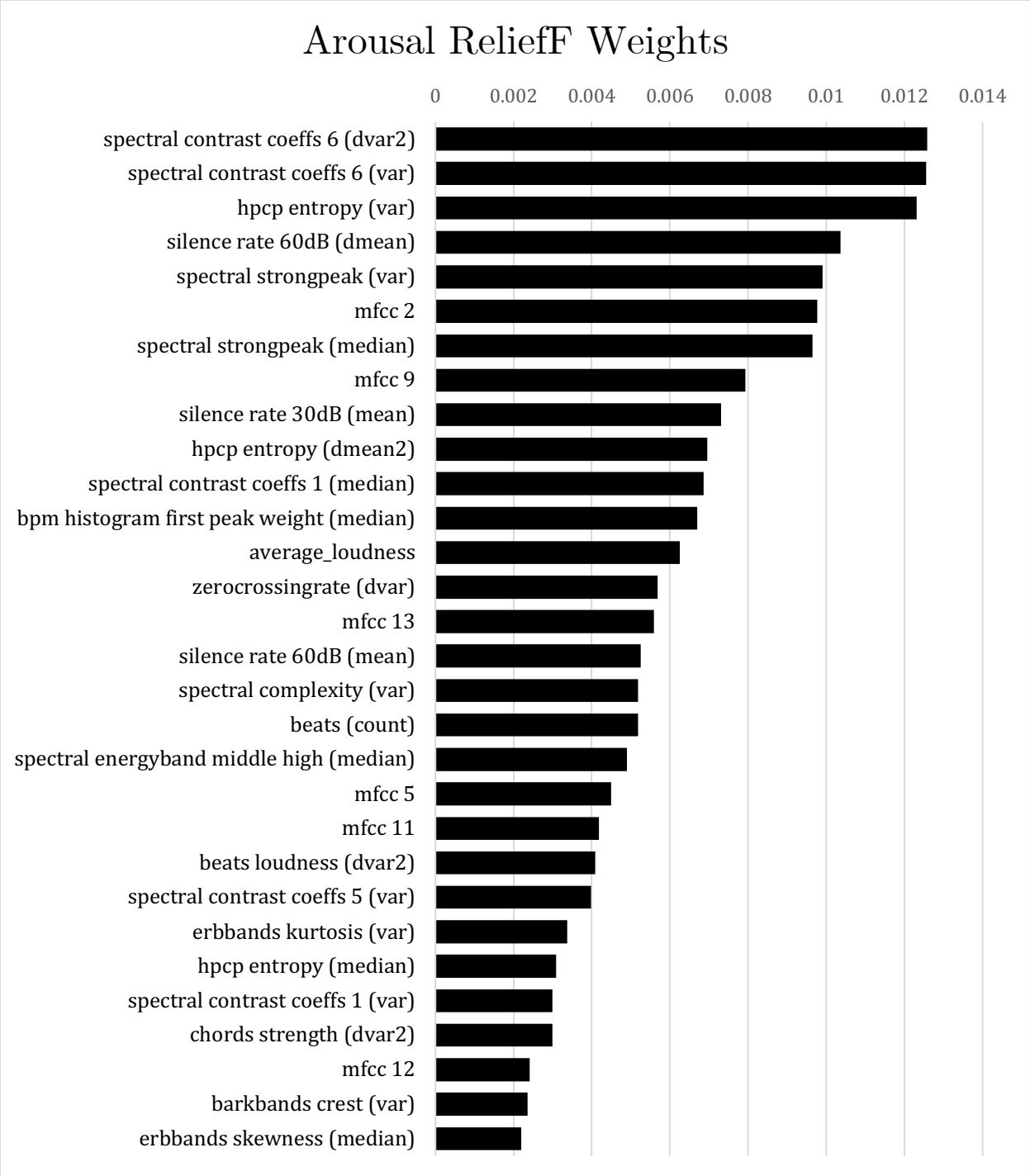

Figure 4.10. Bar graph showing the ReliefF weights for the top 30 features extracted for arousal using the ReliefF feature selection algorithm.

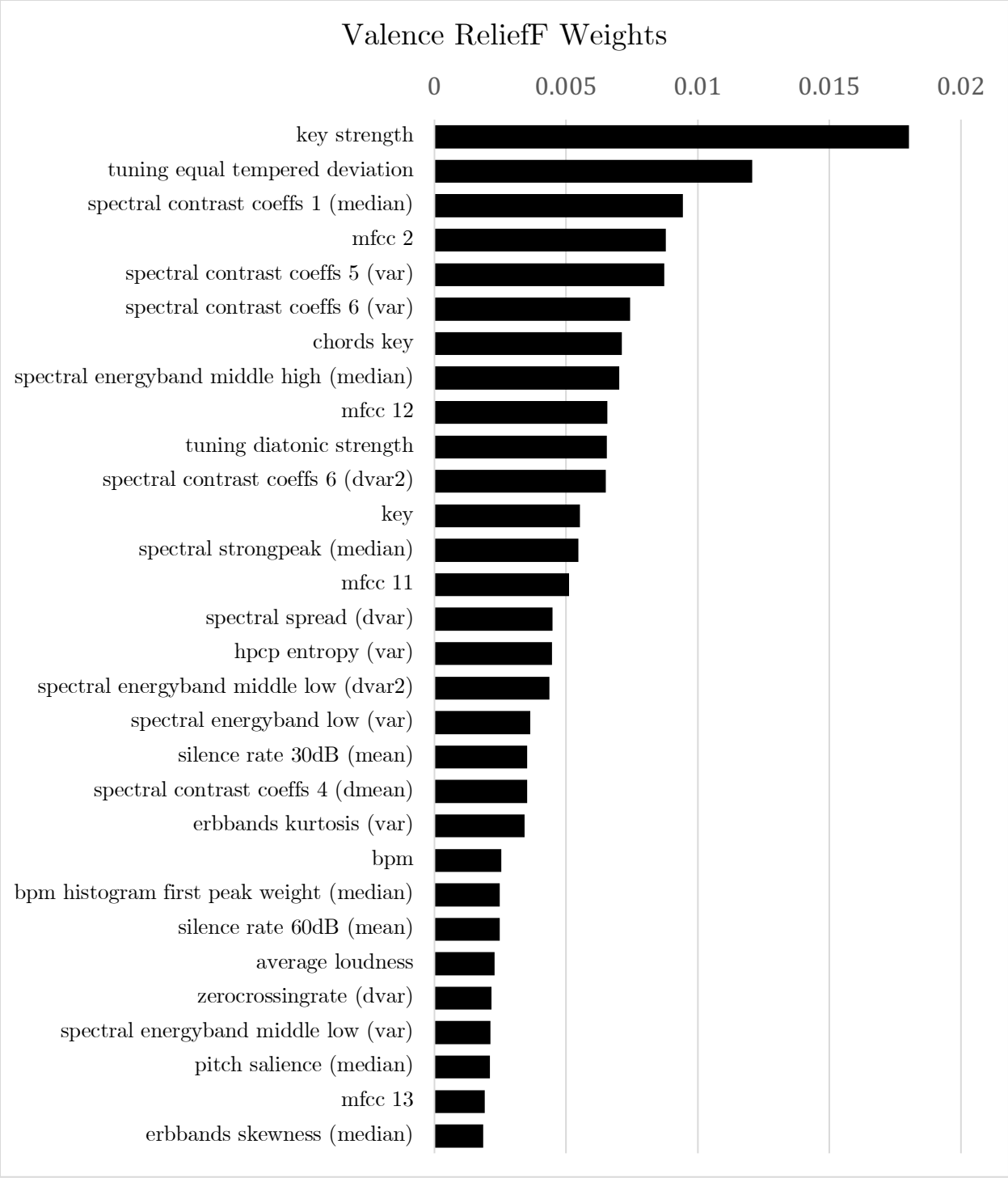

Figure 4.11. Bar graph showing the ReliefF for the top 30 features extracted for valence using the ReliefF feature selection algorithm.

4.4.3 Results

This section provides an overview of the experimental results from my evaluation of the arousal based and valence based personalised emotion prediction systems. Five methods of recommendation (1 non-personalised, 3 collaborative filtering, and 1 content-based filter) were evaluated on the four groups of participants 10 times for both arousal and valence. I used Root Mean Square Error (RMSE) to measure how well the algorithm predicts the actual value rating of individuals for excerpts and participants that were withheld. This is important, as the predictions should correspond with actual user ratings of excerpts. Also, because the user ratings in this system refer to the magnitude of an emotional experience (e.g. very negative or only moderately negative), the systems should be able properly predict the magnitude for an individual.

The results of the recommendation evaluations, presented in Figure 4.12 and Figure 4.13, illustrate that each personalised recommender (with the exception of distance-weighted knn and SVD in the arousal predictions), significantly outperforms the non-personalised recommender system in both arousal and valence predictions. The significance of each pairwise comparison between recommender methods is evaluated using a Wilcoxon Test with Holms p-value correction and shown in Table 4.3 and Table 4.4.

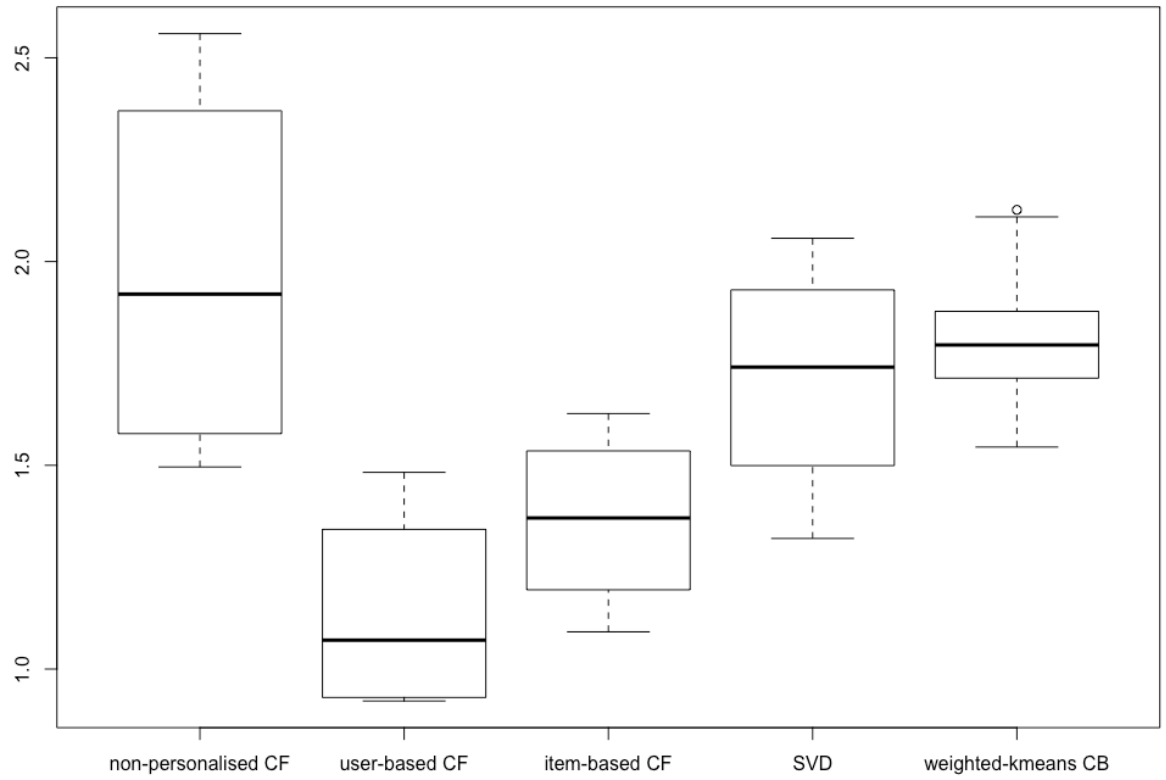

Figure 4.12 A box plot showing the performance of collaborative filtering (CF) and content-based filtering (CB) algorithms on personalised arousal predictions. The y-axis represents the RMSE scores of the recommenders. Across 10 iterations on four groups of participants, the user-based collaborative filtering approach performed best in predicting missing user ratings. The error bars represent the variability across iterations.

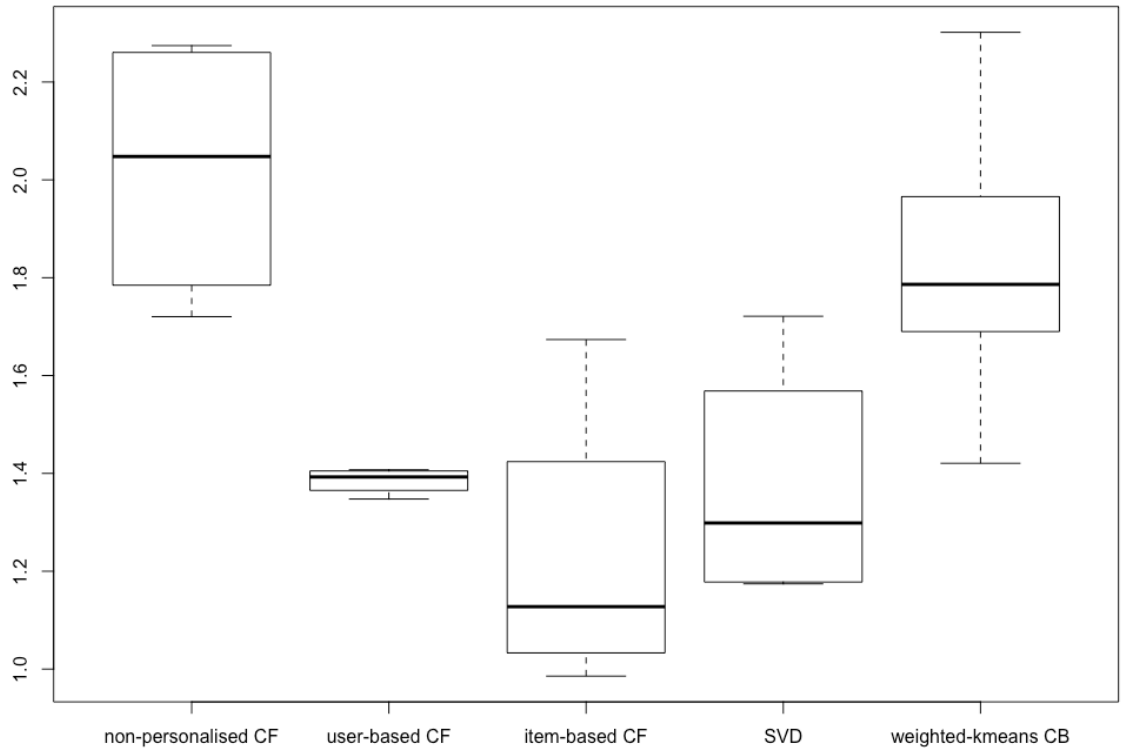

Figure 4.13 A box plot showing the performance of collaborative filtering (CF) and content-based filtering (CB) algorithms on personalised valence predictions. The y-axis represents the RMSE scores of the recommenders. Across 10 iterations on four groups of participants, the item-based collaborative filtering approach performed best in predicting emotional response values. The error bars represent variability across the iterations.

Table 4.3 Results of Wilcoxon Test of significance on the difference in Arousal Recommender performances using Holms p-value correction for multiple comparisons.

Model	non-personalised CF	user-based CF	item-based CF	SVD
user-based CF	$< .001$			
item-based CF	$< .001$	$< .001$		
SVD	.16	$< .001$	$< .001$	
weighted-knn CB	.63	$< .001$	$< .001$	.25

Table 4.4 Results of Wilcoxon Test of significance on the difference in Valence Recommender performances using Holms p-value correction for multiple comparisons.

Model	non-personalised CF	user-based CF	item-based CF	SVD
user-based CF	$< .001$			
item-based CF	$< .001$	$< .001$		
SVD	$< .001$	1.00	$< .001$	
weighted-knn CB	$< .001$	$< .001$	$< .001$	$< .001$

The user-based collaborative filtering approach ($M = 1.14$, $SD = 0.23$) performed best in predicting arousal, with the item-based collaborative filtering ($M = 1.36$, $SD = 0.2$) as a close second. Each of SVD ($M = 1.71$, $SD = 0.27$), weighted k-means ($M = 1.81$, $SD = 0.14$) and the non-personalised ($M = 1.97$, $SD = 0.43$) recommenders performed worse on the arousal condition and showed no significant difference in means. However, for the valence prediction, all recommenders significantly outperformed the non-personalised recommender ($M = 2.02$, $SD = 0.25$) with item-based collaborative filtering ($M = 1.23$, $SD = 0.27$) performing best, followed by SVD ($M = 1.37$, $SD = 0.23$) and user-based collaborative filtering ($M = 1.39$, $SD = 0.02$)[6] with statistically equivalent performances, and distance-weighted knn ($M = 1.82$, $SD = 0.20$).

4.4.3.1 Discussion

Overall, results showed that the personalised recommenders performed better at predicting how music made people feel, in comparison with the non-personalised recommenders. This indicates that it is indeed possible to use the predictions of the collaborative and content-based filtering techniques to produce a proxy of people's felt emotions to provide improved music emotion induction in psychological research. However, there are still some limitations with these models and the current evaluation of them. First, collaborative filtering methods typically require a lot of ratings for each item

[6] Of note, the variability of participants' ratings of the valence of music excerpts (SD = 1.9) was less than that of arousal (SD = 2).

or else suffer from the 'cold-start problem'. The cold-start problem arises when there are simply not enough ratings to identify substantial nearest neighbours from which to preform precise predictions. Also, standard collaborative filtering methods typical rely on a substantial user rating history to make good neighbour comparisons and predictions, and they are locked into the dataset for which they are created. This is problematic if a user's mood influences their emotional response, or if researchers want to introduce new music excerpts into data sets.

The content-based filtering approach resolves the issue of the cold start problem by depending on the items features, rather than users' ratings. Content-based filtering approaches also allow new stimuli to be introduced to the stimulus set, because they predict based on user responses to item features. However, content-based features can only work well if the musical features successfully identify and differentiate individuals' emotional responses. In this study, despite the small amount of data, the content-based method was outperformed by the collaborative filtering methods, indicating that the extracted features may not have been the most useful in differentiating individual emotional responses. To resolve this limitation, I introduce a content-based convolutional-recurrent neural network (CB-CRNN) method in Section 4.5, to personalise feature engineering (feature creation) for content-based filtering.

4.5 Content-based convolutional-recurrent neural network (CB-CRNN)

In this section, I introduce and evaluate a novel content-based technique based on both convolutional (CNNs) and recurrent neural networks (RNNs), which can build memory of events and their outcomes over time. Furthermore, given that the features used for content-based filtering performed poorly compared to collaborative filtering approaches and non-personalised recommendation in Section 4.3.3, I evaluate two state-of-the-art approaches to music feature extraction - using CNNs and RNNs to extract features of music – with the aim of extracting musical features that differentiate

individuals' emotional responses and summarising those features in a way that accounts for the features' short-term and long-term effects on emotional experience.

4.5.1 Methods

4.5.1.1 Feature extraction

Certain musical features can trigger shorter-term or longer-term effects on individuals' emotional experiences, depending on the features that proceed or succeed them in music time. As such, it is important to account for the temporal context when predicting individual's emotional responses to musical features. Thus, in order to develop more personalised music affect induction based on musical features, emotional musical features must be able to describe time dependent effects. I introduce and evaluate two state-of-the-art approaches to music feature extraction that are based on CNNs and RNNs and designed to account for the dynamic aspect of emotional experience: *kernel=2D, convolution=2D* (k2C2) and *convolutional recurrent network* (feature-CRNN) (Choi, Fazekas, Sandler, & Cho, 2017). Each approach extracts features from a *mel-spectrogram* representation of the music signal. The mel-spectrogram is a human perception inspired time-frequency representation of the audio signal derived by weighted averaging of the absolute values squared of a short-term fourier transform (STFT). For this study, I computed the mel-spectrogram for 10 second clips of audio from each excerpt and used parameters that are commonly used in music information retrieval (MIR): N_FFT = 2048, N_MELS = 128, HOP_LEN = 256, Sample_Rate = 22050 (Dörfler, Bammer, & Grill, 2017). This resulted in a matrix of size 128 × 862 (mel-frequencies × time), meaning that each frame size was approximately 0.093 seconds, with a hop length of 0.023 seconds. The resulting matrices then represented the audio excerpts in the neural network. Figure 4.14 illustrates the process of a mel-spectrogram representation being fed into a CNN architecture which learns to extract emotional features from the music. The shaded area of the square refers to a single $n \times n$ region for which a $n \times n$ convolution is performed.

Each layer has n convolutional filters that are applied to the output of the previous layer. The key benefit of this approach is that the CNN learns features from the mel-spectrogram that are highly predictive of emotional experience on an individual basis (see Section 4.5.2).

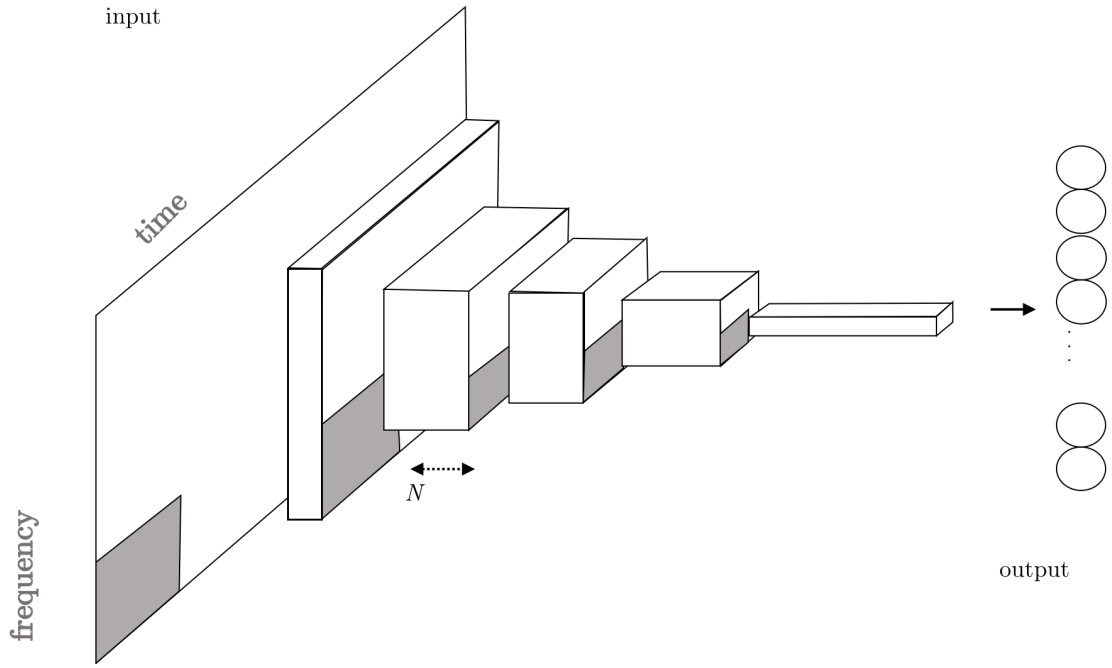

Figure 4.14. A mel-spectrogram representation is fed into a convolutional neural network architecture on the left, which learns to extract emotional features from the music. The shaded area of the square refers to a single $n \times n$ region for which a $n \times n$ convolution is performed. Each layer has n convolutional filters that are applied to the output of the previous layer.

4.5.1.2 CNN feature extractor architectures

I used two CNN- and RNN-based approaches to music feature extraction, namely k2c2 and feature-CRNN, and evaluated them at 0.1×10^6 and 0.26×10^6 parameters (Choi et al., 2017). The full description of each feature extraction model's layer types, layer widths, kernel sizes, maxpoolings, and activation functions are presented in Table 4.5 and Table 4.6.

Table 4.5 Description of the k2c2+ CNN feature extractor. This model affords time and frequency invariances in different scales by gradual 2D sub-samplings. Layer width refers to the number of units in each layer (each network is tested at two sizes), kernel refers to the size of the convolutional kernel, maxpooling refers to the dimensions of the maxpooling applied after each convolutional layer, and activation refers to the activation function used.

k2c2 +				
Type	**Layer width (No. params {0.1, 0.25} × 106)**	**Kernel**	**Maxpooling**	**Activation**
Convolution	{20, 33}	(3, 3)	(2, 4)	ELU
Convolution	{41, 66}	(3, 3)	(2, 4)	ELU
Convolution	{41, 66}	(3, 3)	(2, 4)	ELU
Convolution	{62, 100}	(3, 3)	(3, 5)	ELU
Convolution	{83, 133}	(3, 3)	(4, 4)	ELU
Convolution*	{30, 48}	(1, 1)		ELU
Convolution*	{15, 24}	(1, 1)		ELU

** Note: 1×1 convolutions were added to reduce the dimensions of the feature vector and match the size of CRNN feature vectors for comparison.*

Table 4.6 Description of CRNN feature extractor. This model uses two gated recurrent unit layers to summarise local features extracted using 4 convolutional layers. The assumption of this model is that there are underlying temporal patterns that are better captured by using RNNs than by averaging. Layer width refers to the number of units in each layer (each network is tested at two sizes), kernel refers to the size of the convolutional kernel, maxpooling refers to the dimensions of the maxpooling applied after each convolutional layer, and activation refers to the activation function used.

CRNN				
Type	**Layer width (No. params {0.1, 0.25} × 10^6)**	**Kernel**	**Maxpooling**	**Activation**
Convolution	{30, 48}	(3, 3)	(3, 3)	ELU
Convolution	{60, 96}	(3, 3)	(2, 2)	ELU
Convolution	{60, 96}	(3, 3)	(3, 3)	ELU
Convolution	{60, 96}	(3, 3)	(4, 4)	ELU
GRU	{83, 133}			ELU
GRU	{30, 48}			ELU

The first to be evaluated was the k2c2 architecture, and it was constructed of five convolutional layers with 3×3 kernels and max-pooling layers ((2×4)-(2×4)-(2×4)-(3×5)-(4×4)). It was designed initially for the task of music autotagging (Choi, Fazekas, & Sandler, 2016) and its ability to capture local time-frequency relationships makes it highly suitable for extracting temporal features, such as ostinatos and trills, that effect individuals emotional responses (see Section 5.6.2). In addition to the layer specified in the original research paper (Choi et al., 2016), I added two 1×1 convolutional layers at the end in order to reduce the dimensions of the final music feature vector to match the size of feature-CRNN feature vectors for comparison.

The second architecture I evaluated was the feature-CRNN, which was constructed of four convolutional layers with 3×3 kernels and max-pooling layers (2×2)-(3×3)-(4×4)-(4×4), followed by two RNN layers with gated recurrent units (GRU) to summarise temporal patterns of the CNNs. The summarising of temporal patterns with RNNs rather than statistical moments such as mean and standard deviation allow it, like k2c2, to better explain the short-term and long-term effects of musical features on emotions.

4.5.1.3 Gated recurrent units

RNNs are a type of neural network devised to model variable lengths of sequential data, thus making them suitable for time-series. The ability of RNNs to model long term dependencies in dynamic temporal data make them ideally suited for use in the development of a personalised emotion prediction system that needs to (a) develop a memory of an individual's emotional responses to musical features, and (b) account for the effects of temporal musical features on emotional experiences. RNNs have an internal hidden state that allows them to integrate input from the current time step and previous time steps.

A standard of RNN is described by the update function:

$$h_t = g(Wx_t + Uh_{t-1}) \qquad (4.5)$$

where W and U are weight matrices, x_t is the input at the current time t, h_{t-1} is the previous state, and g is an activation function. Intuitively, the weighting matrices can be thought of as providing a measure of the influence that a feature of input (current or historical) has on the current prediction. However, traditional RNNs suffer from the vanishing gradient problem, an issue which prevents the backpropagation process from affecting weights for more than a few steps (Bengio, Simard, & Frasconi, 1994). Two of the most successful solutions for this problem are the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and the gated recurrent unit (GRU) (Cho, Van Merriënboer, Bahdanau, & Bengio, 2014) RNN models, both of which implement complex gating mechanisms that allow them to learn arbitrarily long dependencies in time-series data. The memory cells of a GRU is illustrated in Figure 4.15.

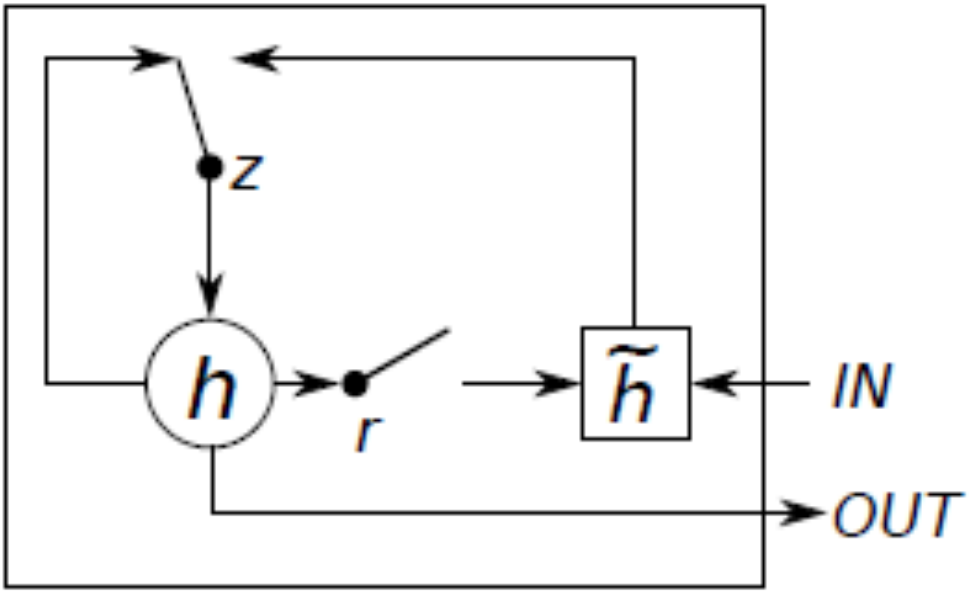

Figure 4.15. GRU Architecture

The GRU is potentially useful in two aspects of developing a personalised emotion prediction system. First, from the feature extraction perspective, the GRU can capture

dynamic aspects such as certain musical features having more immediate or sustained effects on emotional responses. Second, from an emotion prediction perspective, the GRU can learn which present and historical musical features and subsequent emotional responses are highly indicative of an individual's future emotional responses to music features. The innovation in the GRU model is that it uses update and reset gates to decide what information from the present and past should pass through and affect the current outcome.

The GRU's update gate z_t determines how much of the past information is to be passed along to the future:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \qquad (4.6)$$

where x_t is the input to the network at time t, h_{t-1} holds the information about the GRU's activations at previous times, and W_z and U_z are their respective weight (influence) matrices. The products are summed and a sigmoid activation function is used to range them between 0 and 1.

The reset gate r_t is used to determine how much of the past to forget, and is calculated like the update gate:

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \qquad (4.7)$$

with the difference being the value of weight matrices of W_r and U_r. The reset gate is then used to form a memory content:

$$\tilde{h}_t = tanh(W x_t + r_t \circ (U h_{t-1})) \qquad (4.8)$$

where the element-wise product between, r_t and $U h_{t-1}$, determines what to remove from previous time steps, and is summed with the product of a weight matrix W and input x_t. The sum is passed through the non-linear activation function $tanh$.

The final GRU architecture is defined as:

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_0 \qquad (4.9)$$

where the activation h_t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$. The update gate z_t is used to determine what is collected from current memory $\tilde{h}_0$ and previous steps h_{t-1}.

Given these advanced memory cells, RNN models provide the ability to continuously predict an individual's emotional responses throughout time, in relation to the responses it has previously observed from the user to other musical features, and to emotional patterns it has been trained to predict from other users (training data). At the feature level, the models can also summarise the features in a way that accounts for their long and short term impact on emotional experience, and use features at the recommendation level in a content-based filtering approach. The hybrid collaborative and content-based filtering approach created by RNNs captures both explicit content based and latent collaborative emotional music features and their relative impact emotion.

4.5.1.4 Siamese-network architecture

I used two different siamese network architectures that are based on RNNs. A Siamese Network, such as the one shown in Figure 4.16, is a special kind of neural network architecture in which two identical branches of a neural network (shared weights) are simultaneously fed different inputs and forced to learn a similar representation (Bromley, Guyon, LeCun, Säckinger, & Shah, 1994). For example, I want the network to learn how a listener's emotional response changes between two different music excerpts, however I only want the network to learn one set of musical features that can explain individual emotional responses to music. The Siamese architecture allows for this single representation to be learned and is the common choice for tasks that involve finding similarity or a relationship between two comparable things.

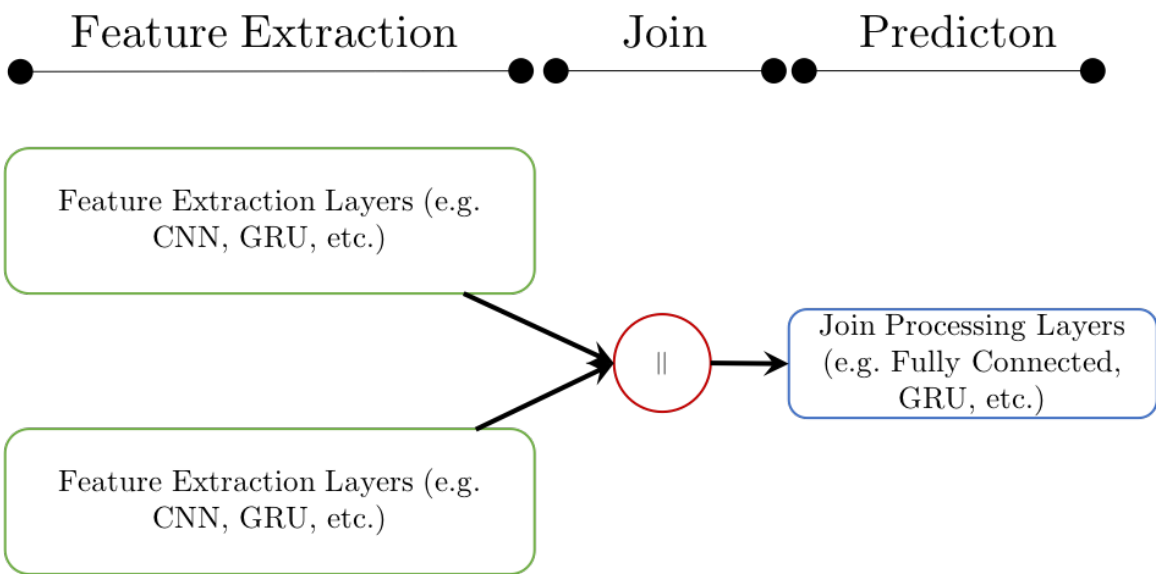

Figure 4.16 Example of a siamese network architecture in which: (1) two feature extraction branches with identical weights extract features, (2) those extracted features are joined through some operation (i.e. addition, concatenation, etc.), and (3) that joint representation is then passed through a second path of the network which form the output prediction.

Both Siamese networks are trained on feature vectors constructed from current music features (unrated), previous music features (rated), and the previous ratings (affect, liking, familiarity), where music feature vectors, are passed through a Siamese network before the outputs of which are concatenated with the ratings of the previous music excerpt in the session and passed through the rest of the network. The network then predicts the person's rating for the current music excerpt. The architecture for this model is shown in Figure 4.17. The RNN approaches are evaluated by splitting participants into training (70%) and testing (30%) blocks and performing RMSE on predictions. The *Keras* Python package (Chollet, 2015), which is a deep learning package that combines both *Theanos* and *Tensorflow* backends, is used for training the RNNs.

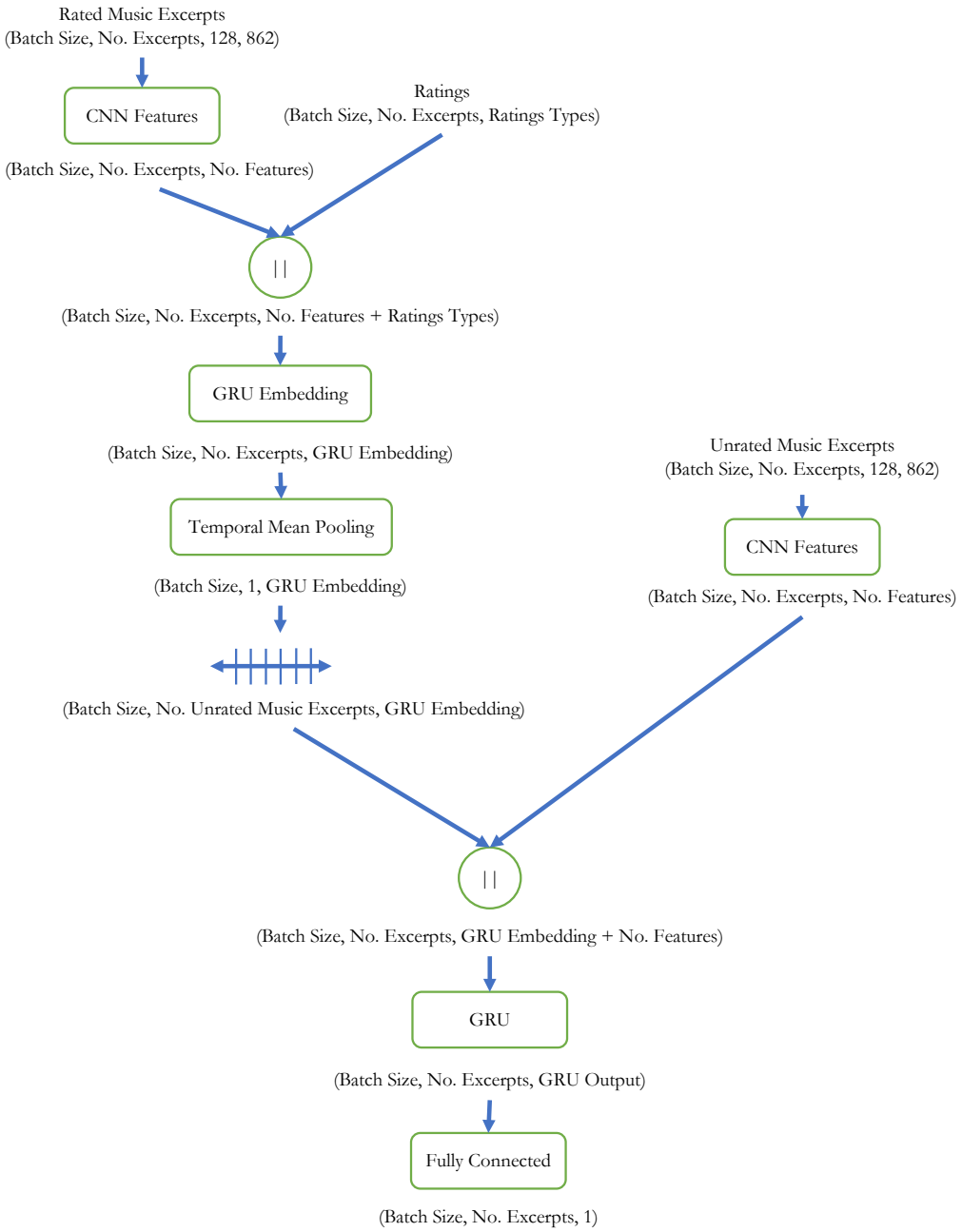

Figure 4.17 CB-CRNN architecture for affective rating prediction. The model takes three inputs: the melspectrogram of rated music excerpts, the ratings, and the melspectrogram of unrated music excerpts. The model outputs the predicted emotion ratings of the unrated music excerpts.

4.5.2 Results

For each condition (arousal and valence), I evaluated 54 parameter configurations for the CB-CRNN using the *Hyperopt* python package (J. Bergstra, Yamins, & Cox, 2013), which implements tree-structured parzen estimators (J. S. Bergstra, Bardenet, Bengio, & Kégl, 2011), to determine the best combination of learning rates, dropout rates, CNN feature extractor, and RNN units. The objective of the hyperparameter optimization is to discover parameters that neither overfit nor underfit the training and testing data. Each model is optimised using RMS propagation, with a 50% decay rate of the learning after every 10 epochs. All models significantly outperform the non-personalised and content-based model evaluated in Section 4.4.3.

The performance of the Siamese models (k2c2 and CRNN) are compared with the results of the five standard models presented in Section 4.4.3. The best performing arousal model (RMSE 1.66), outperformed each of SVD ($M = 1.71$, $SD = 0.27$), weighted-kmeans ($M = 1.81$, $SD = 0.14$) and the non-personalised ($M = 1.97$, $SD = 0.43$). Similarly, the best performing valence model (RMSE 1.61) outperformed the distance-weighted knn ($M = 1.82$, $SD = 0.20$) and the non-personalised recommender ($M = 2.02$, $SD = 0.25$). In both cases, the k2c2 feature extraction architecture outperformed the CRNN.

Table 4.7 Top 5 parameter configurations for each feature extractor on the arousal condition. The k2c2+ feature extractor performed best overall.

Feature Extractor	Rank	Learning Rate	Feature Dropout	GRU Embedding Units	GRU Model Units	Recurrent Dropout	Feature Parameters	RMSE
k2c2+	1	0.003	0.3	32	8	0.04	0.25 × 106	1.66
	2	0.006	0.03	16	8	0.06	0.25 × 106	1.69
	3	0.013	0.02	16	4	0.03	0.1 × 106	1.80
	4	0.003	0.3	32	8	0.04	0.1 × 106	1.81
	5	0.035	0.3	32	4	0.1	0.25 × 106	1.82
CRNN	1	0.003	0.3	32	8	0.04	0.25 × 106	1.76
	2	0.003	0.3	32	8	0.04	0.1 × 106	1.78
	3	0.011	0.3	4	32	0.4	0.25 × 106	1.86
	4	0.001	0.1	32	32	0.4	0.25 × 106	1.87
	5	0.011	0.3	4	8	0.1	0.1 × 106	1.89

Table 4.8 Top 5 parameter configurations for each feature extractor on the valence condition. The k2c2+ feature extractor performed best overall.

Feature Extractor	Rank	Learning Rate	Feature Dropout	GRU Embedding Units	GRU Model Units	Recurrent Dropout	Feature Parameters	RMSE
k2c2+	1	0.013	0.1	16	16	0.1	0.1 × 106	1.61
	2	0.047	0.2	16	16	0.1	0.1 × 106	1.64
	3	0.009	0.03	16	32	0.07	0.1 × 106	1.69
	4	0.003	0.3	32	8	0.04	0.25 × 106	1.69
	5	0.006	0.1	16	8	0.1	0.1 × 106	1.70
CRNN	1	0.003	0.3	32	8	0.04	0.1 × 106	1.70
	2	0.003	0.3	32	8	0.04	0.25 × 106	1.72
	3	0.005	0.1	8	32	0.1	0.1 × 106	1.75
	4	0.001	0.01	8	32	0.05	0.1 × 106	1.79
	5	0.014	0.02	4	32	0.04	0.25 × 106	1.91

4.6 Chapter Conclusion

In this chapter I performed three studies to show that there are individual differences in emotional responses to musical features, and that personalised emotion prediction systems are better suited for predicting emotional responses than non-personalised approaches. In the first study, cohort analysis revealed that a group of relatively homogeneous individuals can vary greatly from other groups in their emotional responses to musical features. I then evaluated the predictive ability of five recommendation approaches in predicting individuals' emotional responses based on the responses of similar people (collaborative-filtering), or similar musical features (content-based filtering), showing that these personalised methods outperform non-personalised prediction. As a final step, I showed the advantages of CNN and RNN methods in their ability to (a) create custom feature extraction and (b) predict the emotional responses of an individual based on their previous responses.

While these results are promising for the personalised emotion prediction system, the stimulus set used for the evaluation was still quite small (120 excerpts). Furthermore, the number of participants used to develop these cohorts was also relatively small. Subsequently, this limitation demanded further investigation to determine whether these methods would still be effective using a larger database of music stimulus and greater population of participants. As such, in Chapter 5, I further develop the personalised emotion prediction system by conducting a similar experiment to Chapter 4 with far greater quantities of data (i.e. with a database of over 1,307 manipulated audio loops and approximately 2,000 participants to provide emotional ratings).

Chapter 5 A Reliable System for Personalised Emotion Prediction

5.1 Introduction

The purpose of this study is to further develop the personalised prediction techniques that were introduced in Chapter 4 (i.e. collaborative and content-based recommender systems), to account for individual differences and more precisely predict emotional response. Most notably, one of the techniques introduced was a novel content-based Convolution-Recurrent Neural Network (CB-CRNN) model that uses combinations of RNNs and CNNs to (a) generate music feature representations, and (b) predict a participant's individual emotional response to music excerpts.

In the previous study (Section 4.4), the collaborative filtering techniques (i.e. user-based and item-based), consistently outperformed both the content-based techniques (i.e. distance-weighted knn and CB-CRNN) and the non-personalised emotion prediction model in predicting individuals' emotional responses to music stimuli. This is not surprising given that collaborative filtering techniques tend to perform well when there are large user and item spaces as they rely on similarities between users or items to calculate ratings. As such, within a largescale emotion induction research setting (using a finite set of prepared stimuli), collaborative filtering techniques appear to create a valuable

prediction model, demonstrating a clear performance advantage over non-personalised prediction techniques.

On the other hand, although outperformed by the collaborative filtering techniques in predicting individuals' emotional responses, content-based techniques offer the benefit of being more adaptable than collaborative filtering techniques in the sense that they are not limited to the stimulus set on which they are trained (i.e. new stimuli can be incorporated without requiring new participant ratings to retrain the model). For example, the novel CB-CRNN technique I introduced in Section 4.5 is a Siamese network which operates by taking samples of (a) the participant' s emotional response (i.e. ratings) to a set of music excerpts as one input, and (b) a set of unrated musical excerpts as a second input, and then combining the output of Siamese branches to predict the participant's response to the unrated excerpts. Theoretically, the excerpts added to either side of the Siamese network could be completely novel, as the network is trained to predict on musical features, as opposed to querying a collected database of users and item ratings. This would be particularly useful for researchers or game developers wanting to introduce new musical excerpts as emotional stimuli and predict personalised responses.

Although the evaluation of collaborative and content-based recommendation techniques in Chapter 4 validate that both are more effective than non-personalised techniques in predicting the induced emotional responses of participants, there were some limitations to that study:

1. The models were developed on a limited stimulus set (160 excerpts).
2. The study had a relatively small number of participants (120 participants).

The present study attempts to address these limitations by (a) using a larger and more musically diverse stimulus set (i.e. with a broader range of features), to improve the content-based filters' ability to generalize to music it has not been trained with, and (b) recruiting more participants, to increase the probability of "tightly aligned" neighbours for

the collaborative filtering techniques to use in predicting a participant's emotional responses.

A second goal of this study is to determine whether the personalised emotion prediction techniques can be used by researchers and practitioners to develop stimuli that manipulate emotional responses. I hypothesized that if the personalised emotion system has been trained to predict an individual's emotional response to a set of loops, then a researcher or practitioner should be able to use these predictions to select loops and create customized musical compositions that will intentionally induce any given emotion for each individual participant (i.e. the person's predicted responses would inform the selection of loops to be used in creating the musical composition, allowing the system to select loops that will induce the particular emotion).

5.2 Chapter Goals/Objectives

This chapter consists of two studies, with several modifications designed to extend the previous personalised emotion prediction systems:

1. I increased the number of participants for this study (1,943 in total).
2. I increased the number of excerpts that were evaluated (1,307 in total).
3. I reduced the musical difference between excerpts by creating excerpts that were more similar to each other (i.e. variations of each other). This modification allows the content-based system to learn at a more resolute level what manipulations of musical features result in changes in individuals' emotional responses. It also allows collaborative filtering systems to capture the rating differences between highly similar excerpts.

In Section 5.3, I discuss the development of a new loop-based stimulus set, to collect people's ratings and form an amenable stimulus. Using the personalised emotion prediction system, researchers will be able design more effective and reliable emotion induction using this loop-based stimulus set. In Section 5.3, I describe the collection of ratings for the new

stimulus set. In Section 5.5, I evaluate the recommender techniques introduced in Section 4.4 and Section 4.5 using the data collected in this study. In Section 5.5, I analyse the features of the different content-based filtering techniques, and show how changes in musical features can be used to manipulate emotions. Finally, I summarize the results of the system in Section 5.6.

5.3 Development of Stimulus Set

For the purposes of this study, I used loop libraries as 'building blocks' to create the stimulus sets. Loops are useful because they can be combined sequentially and simultaneously to quickly and easily create a vast quantity of novel stimuli (i.e. without having to create stimuli from scratch). Loops are short (typically 10-30 seconds), self-contained musical excerpts that can be repeated, and combined horizontally (i.e. sequentially in time) or vertically (i.e. in harmony or polyphony), to form larger musical compositions. Individual loops are typically in the form of instrumental tracks (e.g. drums, bass, pads, strings, etc.) and are presented in libraries (i.e. groups of compatible loops that can be combined).

Typically, loop-based libraries are sold on popular loop websites, such as Producer Loops[7] and Loop Master[8]. These websites have demos of the loops that are prepared by professional musicians, but the loops themselves are hidden behind a paywall. In this section, I discuss the selection and development of a loop-based stimulus set from loop libraries.

5.3.1 Procedure

As cinematic music is usually designed with the explicit intent of inducing emotions across a maximum range of audiences (see Section 3.3), I selected film music again to

[7] www.producerloops.com
[8] www.loopmaster.com

create the stimuli for this study. I identified 19 loop libraries using the single tag "cinematic" on three popular loop library websites. However, to be suitable for developing stimulus sets, loop libraries were required to be (a) of good quality, (b) able to effectively induce emotional states (i.e. Low Arousal, High Arousal, Negative Valence, Positive Valence), and (c) easily combined in sequence-based AAC compositions.

As it is difficult to discern the loops' construction and quality based on the online demos, I used an expert panel to select the initial loop libraries (as in the studies of Chapter 3). The experts were limited to (a) the information they could read about the loop library, and (b) the audio demo provided on the websites. The expert committee consisted of one music professor, four PhD students in music, and one professional musician (see Section 3.2.1 for expert selection).

The expert panel was presented with the 19 loop libraries and asked to answer the following questions about each, by providing ratings on a 9-point Likert scale (1 = lowest, 9 = highest):

1. *How well does the loop library represent each of the emotional states (Low Arousal, High Arousal, Negative Valence, Positive Valence)?*
2. *How do you rate the quality of the loops used in this library?*
3. *How difficult would it be for a machine to create music that conforms to the rules of music using this library?*
4. *How effective would this library be at inducing affect?*
5. *Would you recommend that we use this library for our study?*

5.3.2 Results

The chart in Figure 5.1 shows the normalized (standard score) ratings that the experts provided on the quality, difficulty, effectiveness, and recommendation for each library in decreasing order of recommendation from left to right (i.e. the library with the highest recommendation rating is presented first). The scale is normalized to a zero mean, and I acquired all the libraries for which the experts gave a positive (> 0) or neutral

recommendation. This process resulted in the selection of ten loop libraries in total (i.e. File_01 to File_18 in order of the chart).

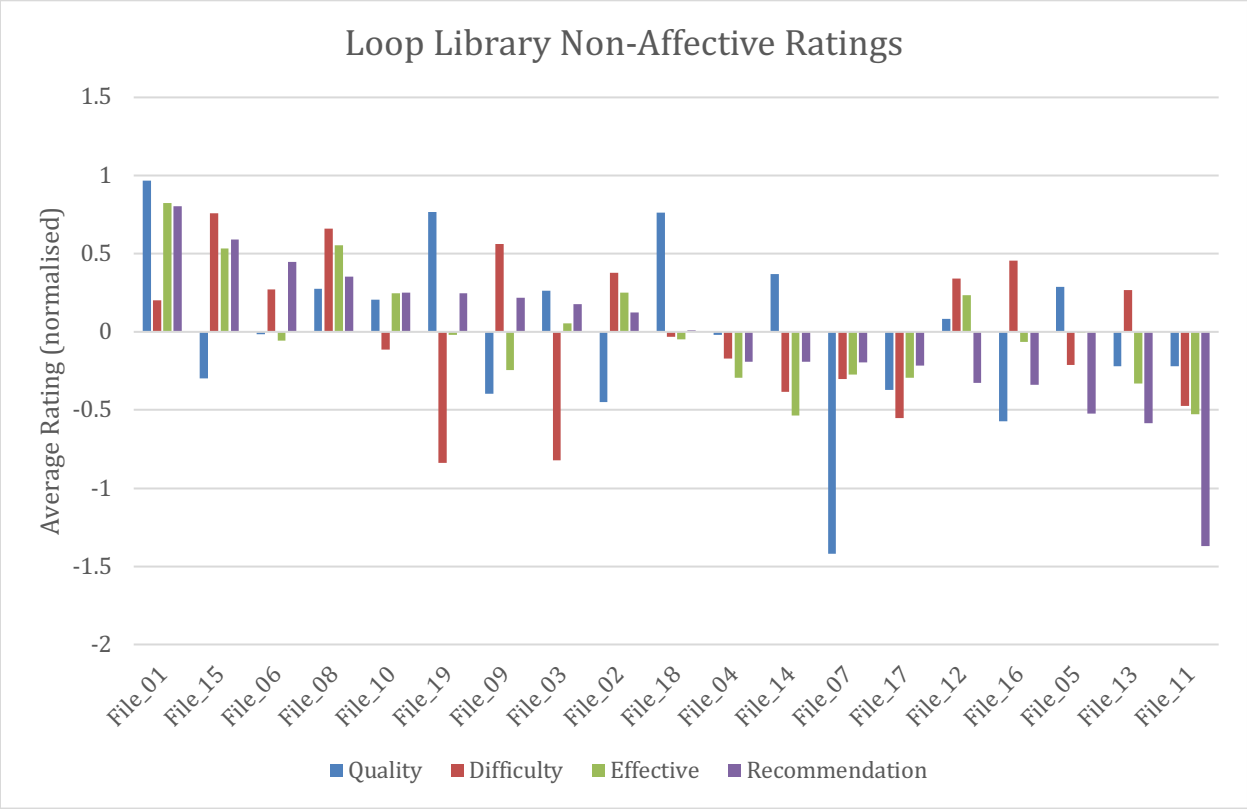

Figure 5.1 A column graph showing the expert panel's normalized ratings for the quality, difficulty, effectiveness, and recommendation of each library. Results are presented in decreasing order of recommendation (i.e. highest recommendation first, from left to right). The first ten libraries exhibiting positive or neutral recommendations (i.e. File_01 to File_18), were selected for further consideration.

As seen in Figure 5.1, most of the acquired libraries also received positive ratings in the other categories (i.e. quality, difficulty and effectiveness); the exceptions being "File_03 and "File_19", which experts suggested may be difficult to use in a sequenced-based AAC system. However, upon acquiring the libraries, only eight files were kept for further preparation, with two (i.e. "File_06" and "File_08") excluded because they were

determined to be too difficult to use in sequence-based AAC systems[9]. This selection process ultimately resulted in the selection of eight loop libraries to be used as 'building blocks' in the construction of the stimulus sets.

5.3.3 Stimulus Preparation

There were found to be two different types of loop-based libraries within the final selection of files. Each loop in the first type of library contains an instrumental layer (e.g. drums, synth, etc.) and is grouped into pre-composed music excerpts that share attributes (e.g. key and tempo) but are free to be combined in any way at any given point in time (i.e. sequentially/horizontally or simultaneously/vertically). The second type of library consists of loops that differ only in instrument type, but are not grouped into pre-composed music excerpts and are also free to be combined in any way. These tracks also share attributes such as key and tempo.

For the first type of loop, I prepared four excerpts in two ways (1) a full example with all voices active (sounding simultaneously), and (2) three partial excerpts randomly generated with a 50% chance of dropout on each instrument voice. The dropout simulates the effect of musical components being added or removed during a musical composition, and allows the personalised emotion prediction system to learn any granular effects that the manipulation of musical features can have on individuals' emotional responses.

For the second type of loop library, I generated approximately ten examples per instrument with a 50% chance of voices dropping out. Ultimately, I produced a stimulus set of 1,307 excerpts, based on 240 loop families.

[9] After exploring the contents, these files were found to be "Construction Kits" rather than loop-based libraries. Construction kits are through-composed pieces of music with libraries constructed of small cues of music that are designed to fit in at specified times in a musical composition (i.e. they are thus unable to be repeated and combined in the same way that loop-based libraries are).

5.4 Collecting ratings for loop-based excerpts

5.4.1 Participants

A total of 1,943 participants were recruited using the CrowdFlower platform[1]. Participants came from 92 countries, including Venezuela (8%), Serbia (7.7 %), the United States (6.4 %), India and Russia (6%). Ages ranged from 18 to 75 (M = 33.5, SD = 11), with 33% of participants identifying as female. All participants were paid $0.50USD to complete the ratings task.

5.4.2 Procedure

Participants were presented with 40 musical excepts (see Section 3.2.3 for procedure), and asked to answer a set of questions about their emotional response to each. Responses were collected as a set of four 9-point Likert scales for each excerpt, on which participants rated (a) the level of arousal they felt (1 = low, 9 = high), (b) the level of valence they felt (1 = very negative, 9 = very positive), (c) how much they liked or disliked the excerpt, and (d) how familiar the excerpt was to them. The 40 excerpts were randomly allocated to participants from a total stimulus set of 1,307 musical excerpts (prepared from the expert-selected loop libraries, see Section 5.3). The task was administered to participants via the Qualtrics platform.

5.4.3 Results

As expected, the count of all rating values, the users' average ratings, and the excerpts' average rating are normally distributed and centred around five (i.e. the central value of the 9-point Likert scale). Each user rated 40 items and each of the 1,307 items received between 25 and 75 ratings. The normalised ratings of each loop are plotted in Figure 5.2, against the normalised ratings of the electronic music stimulus set. As there is much more similarity between excerpts within the loop-based stimulus set, and this effects

[1] Note, this platform was rebranded as 'Figure Eight' in 2018.

the standard deviation in z-score normalisation, each set was normalised individually. The figure shows that the loop-based stimulus set covers a similarly large space as the electronic stimulus set and thus represent an acceptable emotional range.

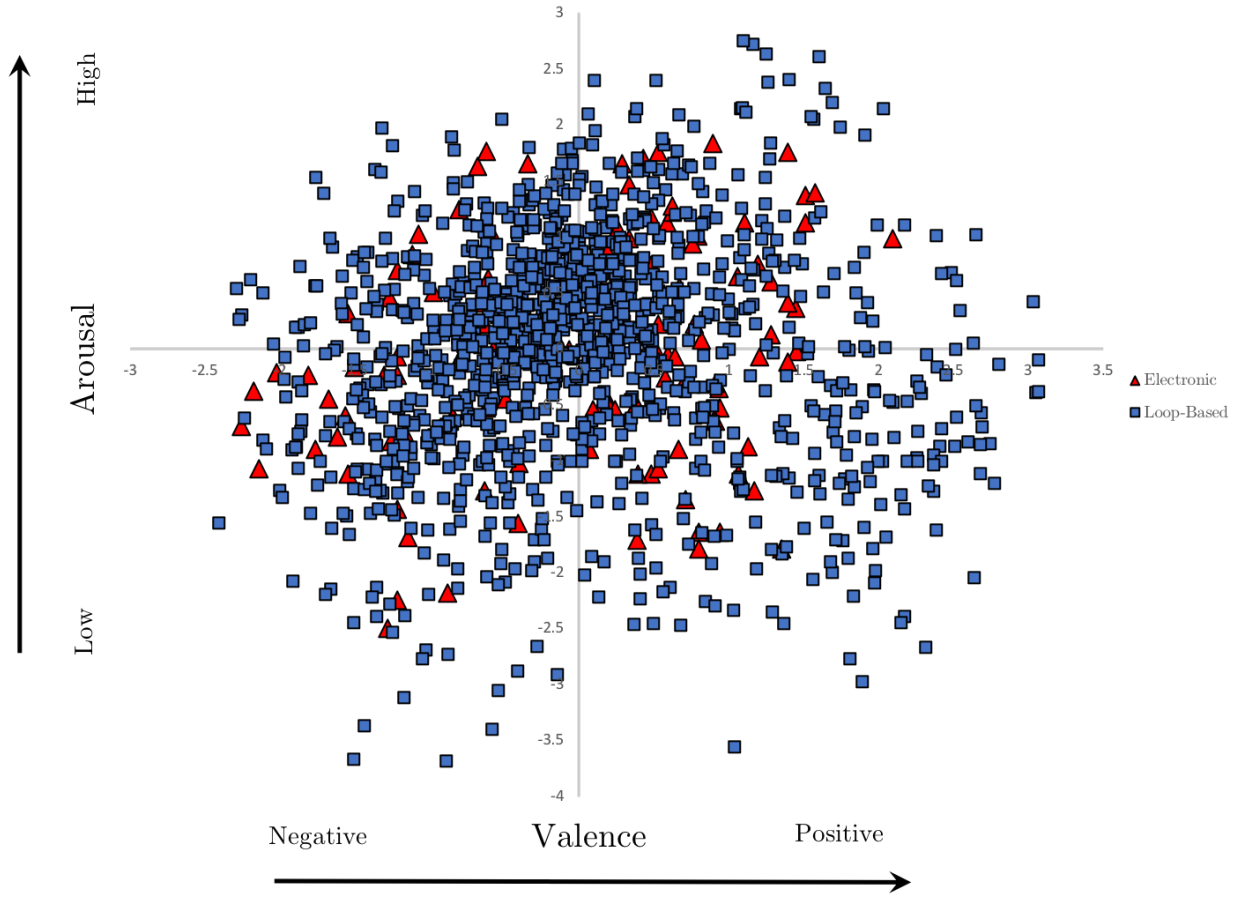

Figure 5.2 This plot illustrates the range of ratings for loop-based music excerpts collected in this study, across the arousal and valence space. It compares the normalised ratings of the electronic stimulus set, with the normalised ratings of the loop-based stimulus set. As illustrated, the loop-based excerpts cover a similarly large space as the electronic-based stimulus set.

I performed a visual inspection of cohorts of raters in Figure 5.3 and Figure 5.4 for arousal and valence respectively. To do this, I calculated the cosine similarity of every pair of users and used agglomerative hierarchical clustering to form dendrograms. Cosine similarity ignores any items that have not been rated by both users, and only includes mutually rated items in the calculation of similarity. Although I had previously used k-means clustering in Section 4.3.2.2 (to also validate the existence of cohorts of individuals

and show that the responses to musical excerpts could be subject to individual differences), the k-means visualisations would not be as effective in this study, because they are less capable of demonstrating coarse separations in this instance. The current study produced more sparse ratings (less overlapping ratings between individuals), and demonstrated higher musical similarity between excerpts (less separation in musical feature space). As such, I elected to use dendrograms instead in Figure 5.3 and Figure 5.4 to visualise these results, as dendrograms are able to give both a coarse and fine grain intuition of how cohorts form. This intuition gives insight on how the predictive systems are expected to perform, and the reason for the performances.

A visual inspection of how participants experienced arousal, shown in Figure 5.3, reveals that participants fall into roughly four cohorts. These cohorts in the dendrogram are colour coded, and I have identified these by splitting by the top four nodes in the tree. These groups are of roughly similar size, suggesting that multiple factors contribute to individual differences in the arousal. This contrasts with the valence dendrogram in Figure 5.4, of which a visual inspection does not yield such substantial cohorts. I split the cohort into 6 clusters to analyse valence, revealing that most participants belong to one massive cohort, and relatively fewer belong to the others. The contrasting dendrograms indicate that participants tend to be more aligned in the level of valence that was induced by each excerpt, rather than the level of arousal. This suggests that valence responses may be less difficult to predict than arousal responses. This result should be reflected in the non-personalised emotion prediction models (i.e. non-personalised emotion prediction models should perform better in predicting valence responses than predicting arousal responses), and could potentially be reflected in the personalised systems as well.

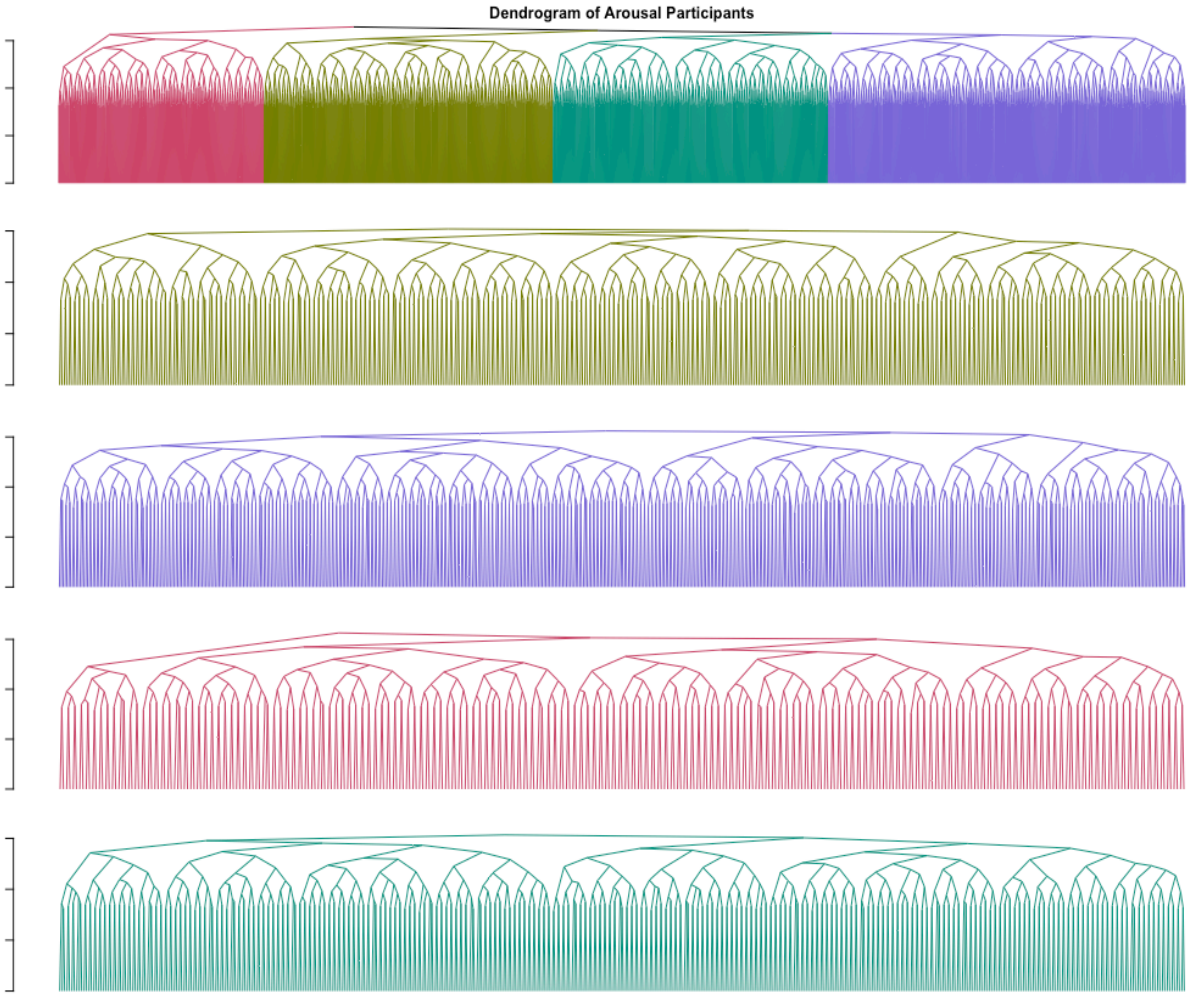

Figure 5.3 A dendrogram showing participants' arousal responses highlights that there are at least four substantial cohorts of ratings. The dendrogram on the top represents the separation into four cohorts. The remaining dendrograms represent each of the four cohorts individually.

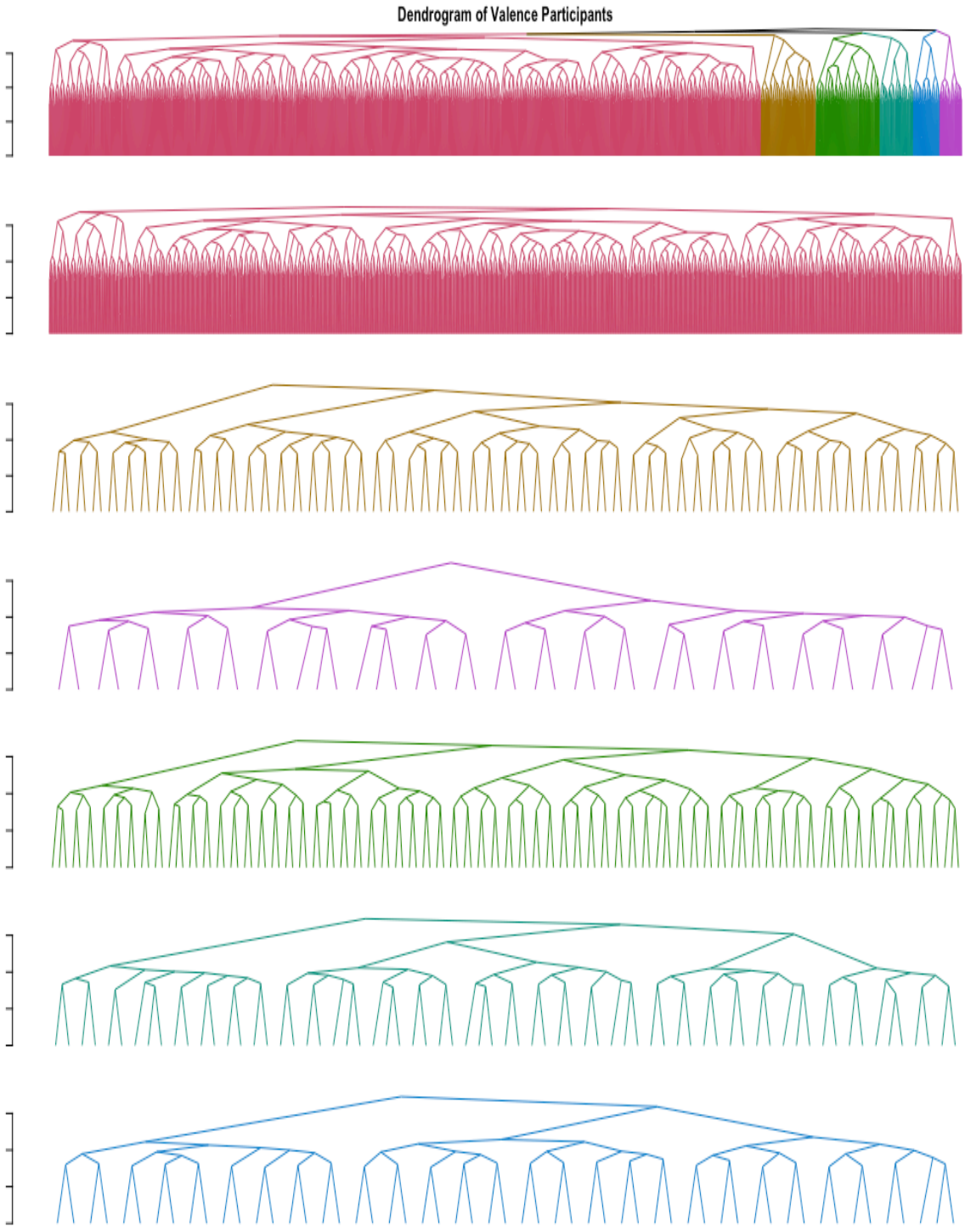

Figure 5.4. Dendrogram showing participants' valence responses highlights one massive cohort and five much smaller ones. The dendrogram on the top represents the separation into six cohorts. The remaining dendrograms represent each of the six cohorts individually. The agreement on valence suggest that the non-personalised prediction might perform better than it does for the arousal condition.

5.4.4 Feature selection for content-based filtering

Content-based filtering works by predicting emotional responses to novel stimuli based on an individual's response to stimuli with similar musical features. However, as stated in Section 3.6.1 and Section 4.3.1, using too many identified features in the analysis will create noise and redundancy and reduce the overall efficiency of the system. Therefore, before evaluating the distance-weight knn content-based filters' ability to predict individuals' emotional responses, I again perform dimensionality reduction.

The method I used to conduct this dimensionality reduction is a repetition of the procedures in Section 4.3. I first used the ReliefF feature selection algorithm to weight the importance of each feature (see Section 4.3 for technical explanation). Next, to select the cutoff threshold for the ReliefF weightings (i.e. to determine at what weighting level the features should be deemed less important), I conducted a AMOC changepoint analysis, with parameters set to test for changes in the mean and variance of the sequence of weight values.

The AMOC changepoint analysis showed with 95% confidence that (a) the first 30 features were the most important to participants' ratings of arousal response (see Figure 5.4), while (b) the first 20 features were most important to participants' ratings of valence response (see Figures 5.5). Consequently, I selected the top ranked features for both arousal and valence as the features to be used in training (see Figures 5.6 and 5.7 for a list of features for arousal and valence respectively). This resulted in a final selection of 36 unique features to be included in the training model (i.e. 14 of the top ranked features were found to be important to both arousal and valence).

Unsurprisingly, most of the feature groups for arousal and valence appear to be the same. Minor differences arise from which statistical moments are represented from the features (i.e. mean, median, variance, etc.). The bigger differences come with the addition of dynamic complexity, onset rate, spectral entropy, and spectral rolloff and the removal of some features related MFCC bands. This is not surprising, as these new features are

commonly identified in music emotion studies, and it is not as clear what low- and high-level timbral features the MFCC bands are capturing (see Section 3.5 for discussion of music features).

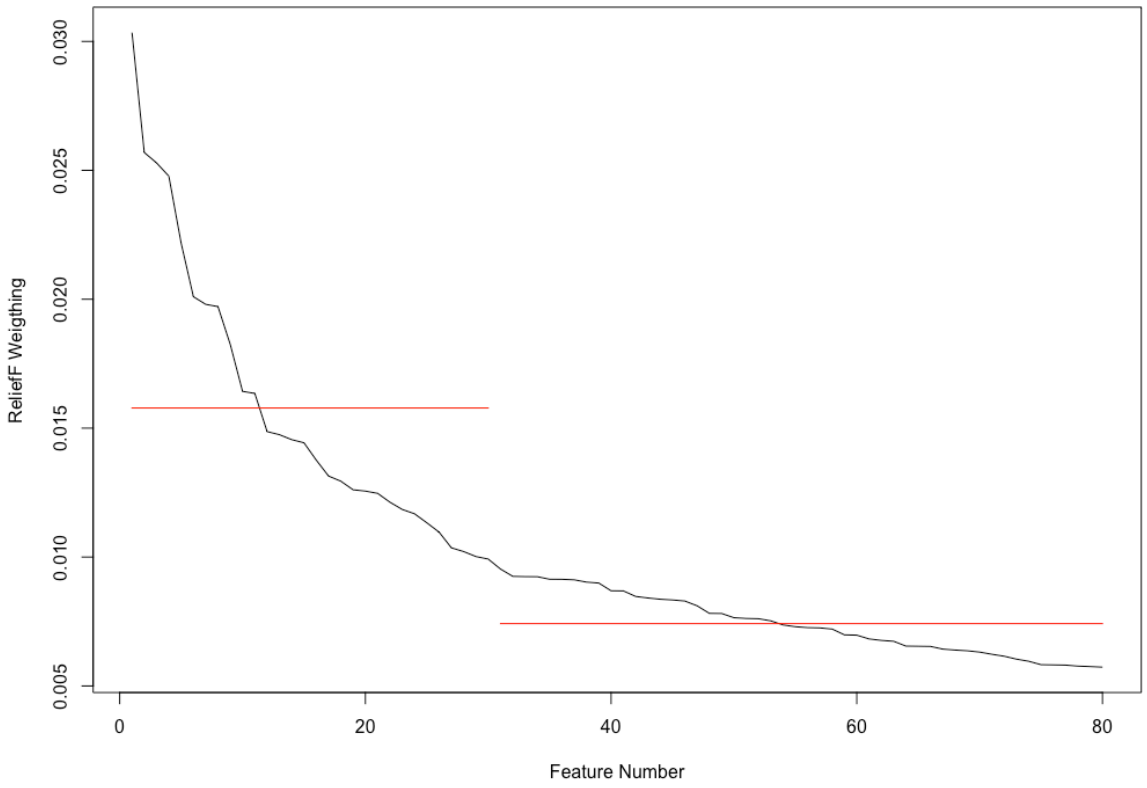

Figure 5.5. AMOC changepoint analysis on Arousal music feature ReliefF weights determined with 95% confidence that a change in the mean and variance occurred after 30 features. The red bars shows the cut-off of the first section of 30 features, and the beginning of the second section (where the mean and variance changes).

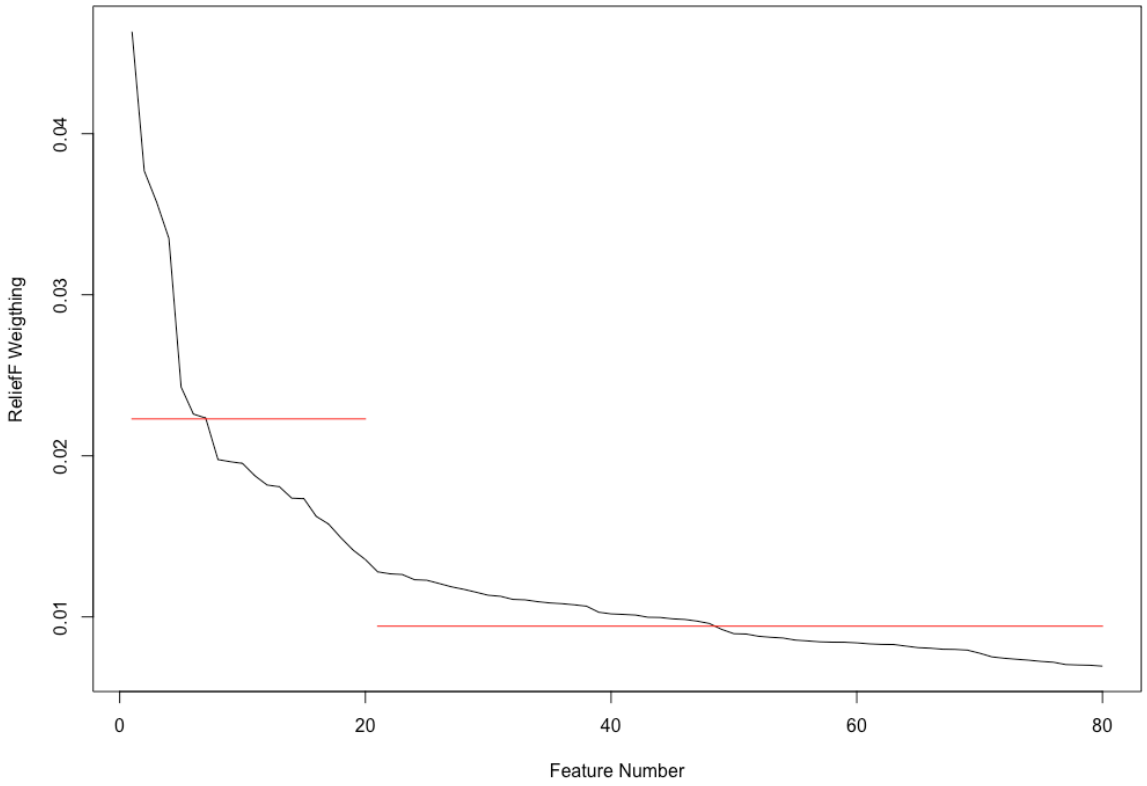

Figure 5.6 AMOC changepoint analysis on Valence music feature ReliefF weights determined with 95% confidence that a change in the mean and variance occurred after 20 features. The red bars shows the cut-off of the first section of 20 features, and the beginning of the second section (where the mean and variance changes).

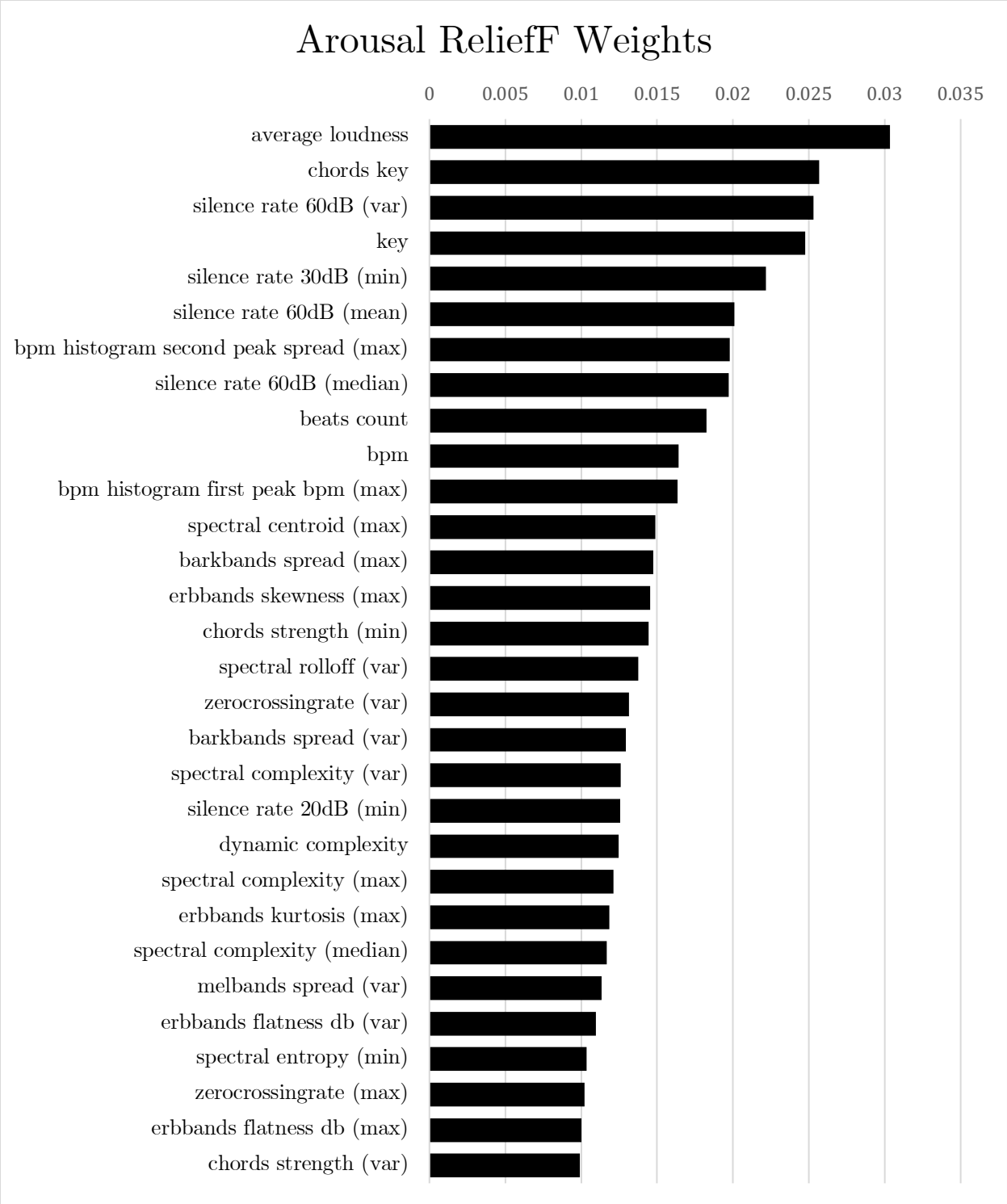

Figure 5.7 Column graph showing the top 30 features extracted by the ReliefF feature selection algorithm for the arousal condition

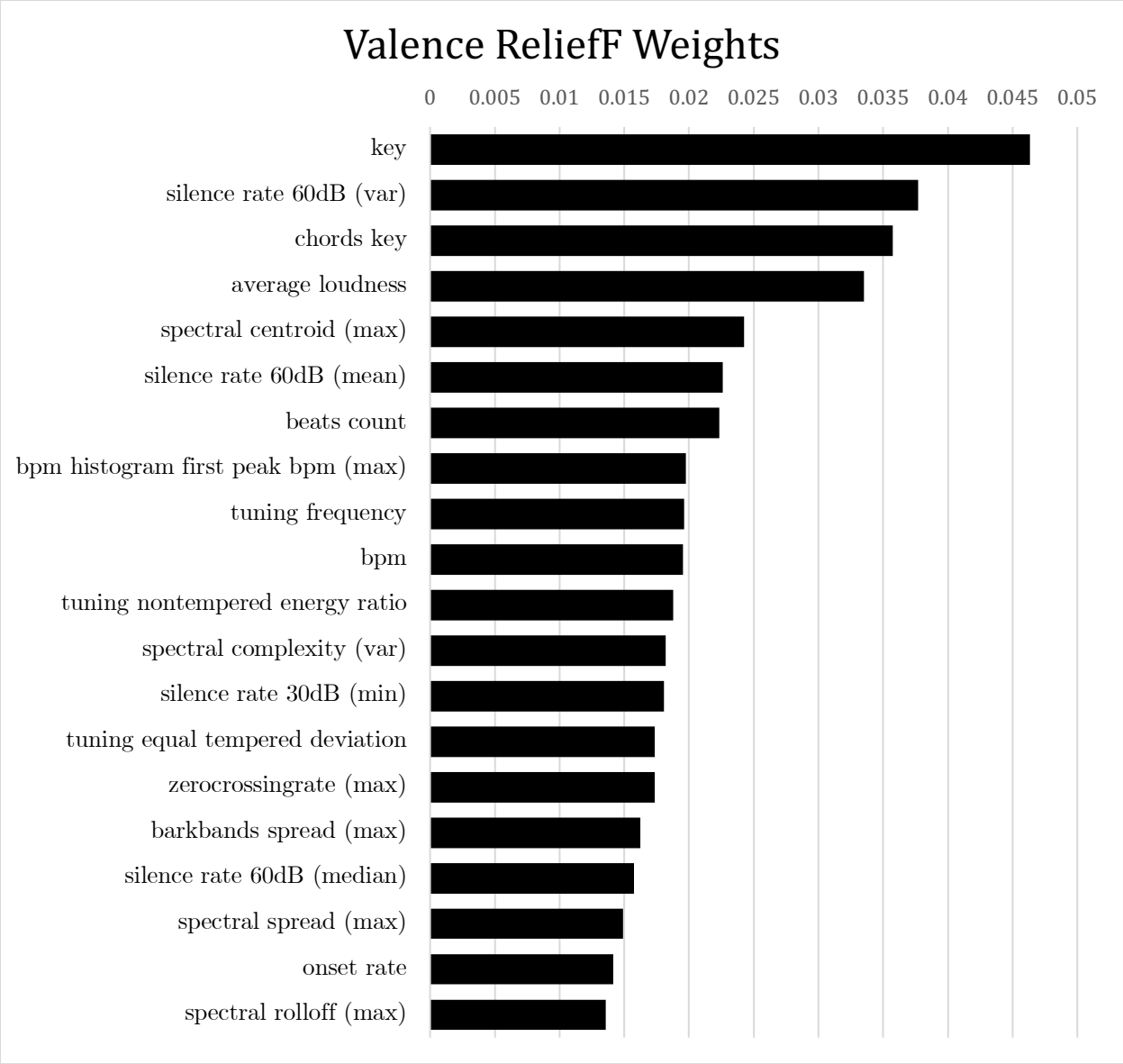

Figure 5.8. Top 20 Features extracted for Valence using the ReliefF feature selection algorithm

5.5 Evaluating Recommender Systems

The studies in Chapter 4 validated that the personalised recommender techniques (i.e. user-based collaborative filtering, item-based collaborative filtering, Singular Value Decomposition (SVD), distance-weighted knn content-based filtering, and CB-CRNN) appear to all predict individuals' emotional responses to musical stimuli better than

traditional non-personalised prediction techniques. Specifically, the user-based filtering system was found to outperform all other methods, followed by item-based filtering systems, then CB-RNN, with distance-weighted knn performing poorest of the personalised prediction techniques. For these re-evaluations, I held out 5% of ratings from 30% of participants (as in Section 4.4.1), to ensure novel test data was kept for each of the collaborative and content filtering techniques (i.e. to allow the system to be tested on novel music excerpts after training).

For the CB-CRNN introduced in Section 4.5, I evaluated both CNN-based feature extractors (k2c2+ and feature-CRNN) at 0.1×10^6 and 0.26×10^6 parameters while varying hyperparameters such as the number of GRU units, learning rate, and dropout rates, and found that the k2c2+ features extractor outperformed feature-CRNN. However, in the present study the stimulus sets are much larger and more diverse, and a greater number of features must be learned. As such, both extractors (i.e. k2c2+ and feature-CRNN) must be re-evaluated to ensure they learn all appropriate features. The hyperparameter search was not repeated for this study, as it would take too long to train given the stimulus set size and the size of the parameter space. Instead, the parameters were set to match those found to consistently rank the best in predicting both arousal and valence in Section 4.5.2 (see Table 5.1).

Table 5.1 Parameters for training CB-CRNN recommendation

Learning rate	Feature dropout	GRU embedding units	GRU model units	Recurrent dropout
.003	.3	32	8	.04

5.5.1 Results

The recommender evaluation results are presented in Table 5.2. The table shows the root mean square error (RMSE) loss for each evaluated technique, whereby lower values indicate less error on the individuals' predicted response. Many of the models were

found to perform better on valence than on arousal responses, corroborating the visual inspection of cohorts from Section 5.4.3. Notably, the non-personalised prediction outperforms both the distance-weight knn and the SVD approach in the valence condition. This confirms my prediction in Section 5.3.3 that the non-personalised approach should work better for predicting valence than arousal (given the level of agreement that was seen in the valence responses).

A comparison of these RSME results (see Table 5.2) with those found in Section 4.4.3, shows that the performance of many of the recommender systems did indeed improve as expected with the increased number of participants. User-based collaborative filtering, the top performing model in both conditions, outperformed all other recommenders, and improved significantly over the previous study (Arousal RMSE=1.14, Valence RMSE=1.39). This improvement confirms, as expected, that increasing the number of participants does indeed improve the model's ability to predict emotional responses. However, the performance of item-based collaborative filtering, did not improve over the previous study (Arousal RMSE=1.36, Valence RMSE=1.23). This is not surprising however, considering that determining items with similar ratings patterns was a much simpler task in the previous study (i.e. participants rated a random selection of the total items in a group for the present study, whereas in the previous study all the items in a group were rated by all the participants). Interestingly, SVD was found to perform poorer in this study than previously (Arousal RMSE=1.73, Valence RMSE=1.37). Future studies should be conducted to determine why this result is observed. Finally, the Convolution-Recurrent Neural Network also outperformed the distance-weighted knn content-based filtering, showing this novel method to be a more viable option for personalised prediction of music emotion induction than the more commonly used content-based knn technique.

Table 5.2. The RMSE performance of recommender methods

Condition	Non-personalised	User-based CF	Item-based CF	SVD	knn CB	k2c2 param_0	k2c2 param_1	CRNN param_0	CRNN params_1
Arousal	1.77	0.80	1.47	2.02	1.70	1.63	1.63	1.63	1.64
						1.63	1.65	1.65	1.64
Valence	1.66	0.81	1.43	1.93	1.73	1.56	1.57	1.57	1.55
						1.57	1.57	1.58	1.58

5.6 Understanding feature representations and embeddings

The topological representation of an item in vectors is known as an "embedding". An interesting and useful property of well-formed embedding spaces are their ability to capture relationships between their embeddings. The most famous examples are the word2vec embeddings (e.g. *king* - *man* + *woman* = *queen* and *paris* - *france* + *poland* = *warsaw*). Embeddings essentially provide a spatial representation of items, with more similar items represented closer to each other in space. Within the present context (i.e. for the personalised emotion prediction system), embeddings can be used to help define a music excerpt within the context of its musical features, and place it relationally on the scale of emotional responses.

As such, in the content based-filtering approaches, I endeavoured to create embeddings of musical excerpts that best captured their effect on emotional responses. A well parameterized embedding of emotional music features allows for better training of the various content-based filtering algorithms and allows them to be more predictive of emotional responses. The k2c2+ and feature-CRNN features for example were directly formed to account for individual differences. Crucially, in the context of this thesis, a well parameterized embedding of emotional music features yields navigable music emotional spaces.

In this section, I use embeddings to validate that the system can create more personalised emotion predictions through the use, combination, and manipulation of musical loops. To accomplish this, I first created a visualization of the emotional spaces created by the different music feature representations. This showed that the neural network based feature representations, which were designed to account for individual differences in emotional responses, did indeed better separate the music-emotion embedding space. These embedding representations thus create more direct paths for potential automated systems to navigate the music-emotion spaces (i.e. more similar

features are represented closer together in representational space, and as such, the system has a more direct path to access similar features).

Next, to provide an example of what kind of emotion inducing music features were being learned by the neural network feature representations, I performed an auditory analysis (i.e. a listening exercise) of the features of *CRNN_param_0*. This allowed me to identify some of the high-level musical features that may be used to predict, differentiate, and manipulate individual's emotional responses. Finally, I showed how random manipulation of certain groups of loops (see Section 5.3.1) resulted in different positioning in the emotional space. This highlights some of the types of manipulations an automated system could learn in order to induce different emotional responses in an individual.

5.6.1 Visualization of Embeddings and Emotional Space

To better understand how these emotional embedding spaces are formed, I plotted the embeddings in 2D space using the well-established dimensionality reduction technique, *t-sne* (Maaten & Hinton, 2008). In Figure 5.9., three visualizations are presented to depict the arousal embedding space, using three different musical feature extraction techniques (tandard features, k2c2+, CRNN). In the top row, each point represents a music excerpt in n-dimensional space, with the average level of arousal that the algorithm has attributed to that excerpt represented on an intensity scale of colour ranging from low (*blue*) to high (*red*) intensity. In the bottom row, a density plot of the actual participant ratings is added as an underlay (using the same intensity scale), to show how these embedding areas relate to the average person's emotional responses. As can be seen in the top row of illustrations, all three representations show a separation of the extreme in arousal and valence levels. However, the density plots underneath show that the features learned by the neural networks are more consistent with the actual data and do a better job of separating the arousal space than the standard feature extractor. This indicates that the music features captured by the neural network are likely more useful for navigating the arousal space.

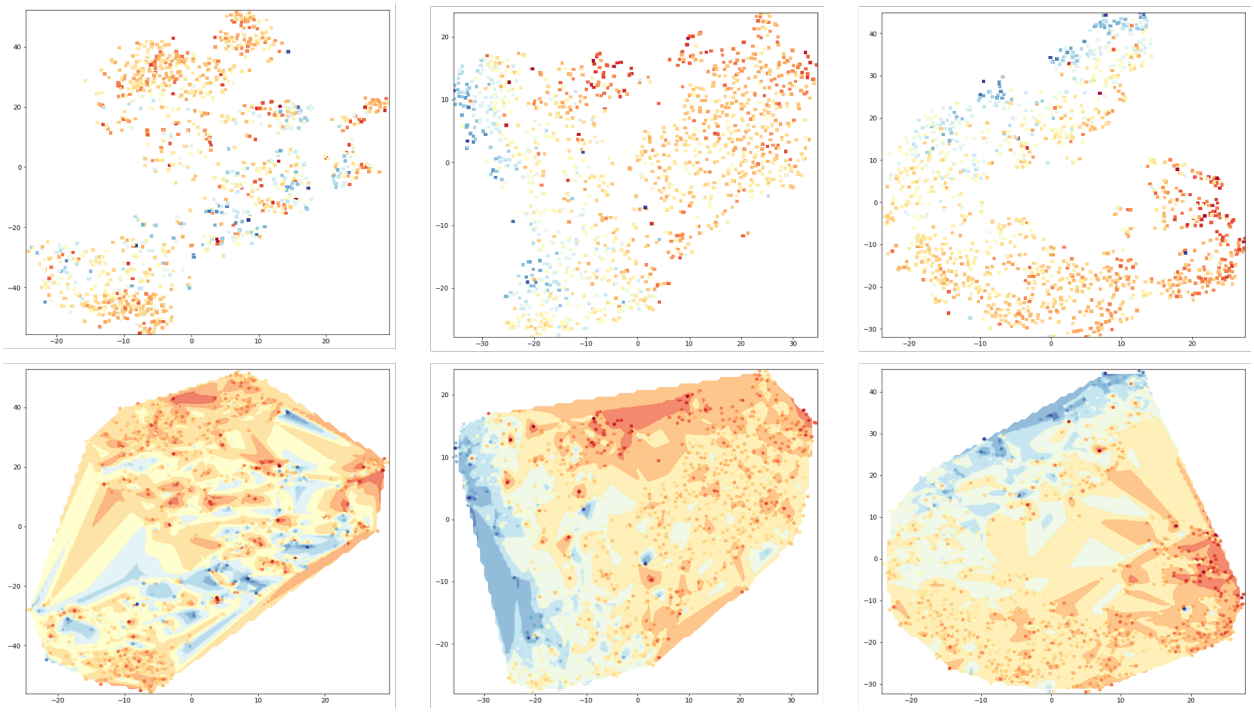

Figure 5.9. T-sne visualizations of arousal music embeddings. These representations of musical excerpts in n-dimensional space were produced by, in columns from left to right, (a) standard feature extraction, (b) the k2c2+ feature extractor, and (c) the CRNN feature extractor. The top row of plots depict a standard t-sne representation of the embeddings, while the bottom row includes a density map of the actual participant rating response averages as an added underlay. In both cases, level of arousal is represented by colour, on an intensity scale ranging from low intensity (blue) to high intensity (red).

Similar results are also observed when the valence emotion spaces are represented in this format. In Figure 5.10, the t-sne embedding plots and density map underlays are used again to represent the valence emotional space for each of the three feature extraction techniques. Once again, the top row of t-sne representations show that all three embeddings separate musical excerpts into the extremes of the valence space well. However, again, the density plot also reveals that the embeddings created by the neural network are more consistent with participant ratings and do a better job of separating the valence space overall. Consequently, these visualisations of both arousal and valence spaces suggest that the neural networks model of embedding may generate feature representations that are highly useful for both predicting and inducing emotion.

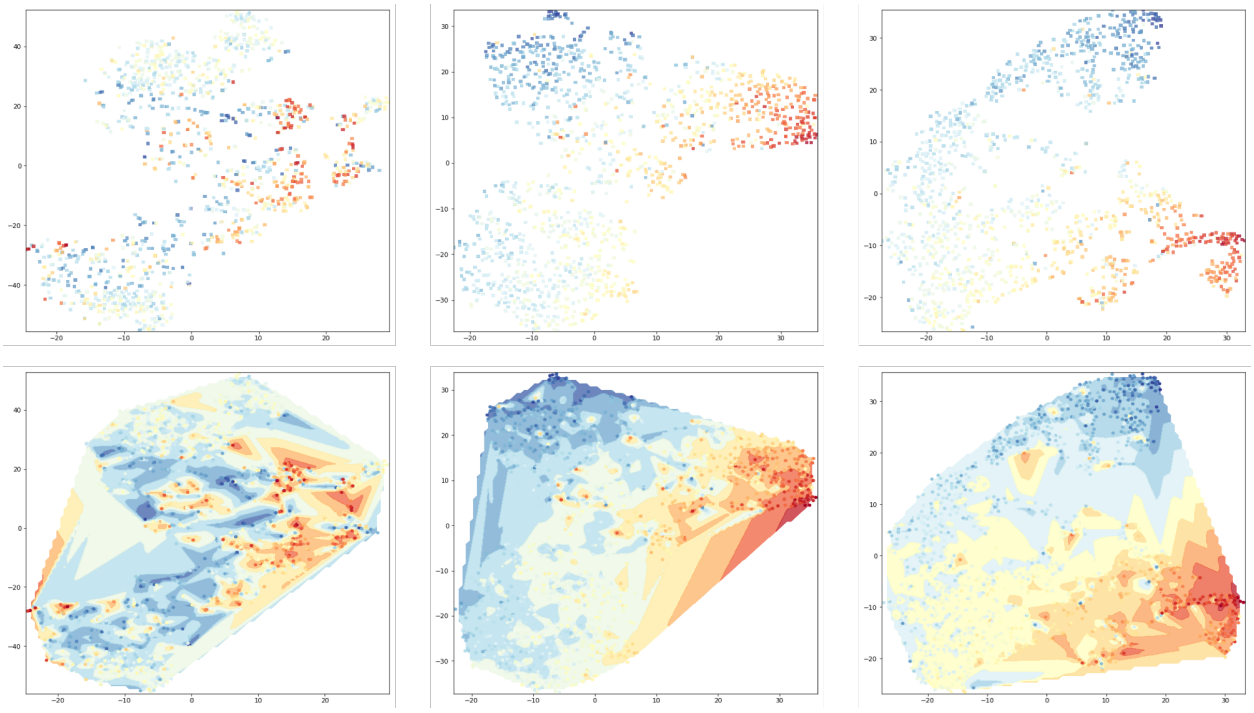

Figure 5.10 T-sne visualizations of valence music embeddings. These representations of musical excerpts in n-dimensional space were produced by, in columns from left to right, (a) standard feature extraction, (b) the k2c2+ feature extractor, and (c) the CRNN feature extractor. The top row of plots depict a standard t-sne representation of the embeddings, while the bottom row includes a density map of the actual participant rating response averages as an added underlay. In both cases, valence is represented by colour, on an intensity scale ranging from negative (blue) to positive valence (red).

5.6.2 Understanding how musical features are represented in the neural networks models

As explained above, neural network models provide a clear benefit over standard feature extraction methods in learning and representing musical features. To quickly recap, in standard feature extraction methods, certain relevant features that have been predetermined to be relevant are selected and assigned to the system by the researcher. In contrast, the neural networks models (i.e. the k2c2+ and feature-CRNN methods used above), can learn from and apply the relationships between inputted data to distinguish, to a much more granular level than is possible from a predetermined set of features, which musical features are most important. As such, these neural network representations inherently account for individual differences in emotional response to music by allowing for the engineering of feature representations that have been derived specifically from the

induced emotional responses (i.e. as opposed to being derived from a predetermined, theoretical set of features).

However, describing what neural networks are learning is an art in itself. The networks themselves can function as somewhat of a 'black box' (i.e. it is a complex system, so the internal workings can be difficult to define in concrete terms), and dissecting what each feature in every network represents musically and how it affects emotional responses would be an exhaustive and time-consuming process. Fortunately however, by examining correlation heatmaps for the k2c2+ and feature-CRNN representations at *param_ 0,* it is clear that the representations are highly correlated for both arousal (see Figure 5.11), and valence (see Figure 5.12). These heatmaps show that the majority features in the feature-CRNN model have highly correlated counterparts in the k2c2+ model, thus showing that similar music features are being used to represent emotion in both models. Therefore, as similar music features appear to be captured across both models, it should suffice to examine just one example representation in detail to understand the types of relationships that are being learned across both. For the purposes of this study, I will use the feature-CRNN representation as this example.

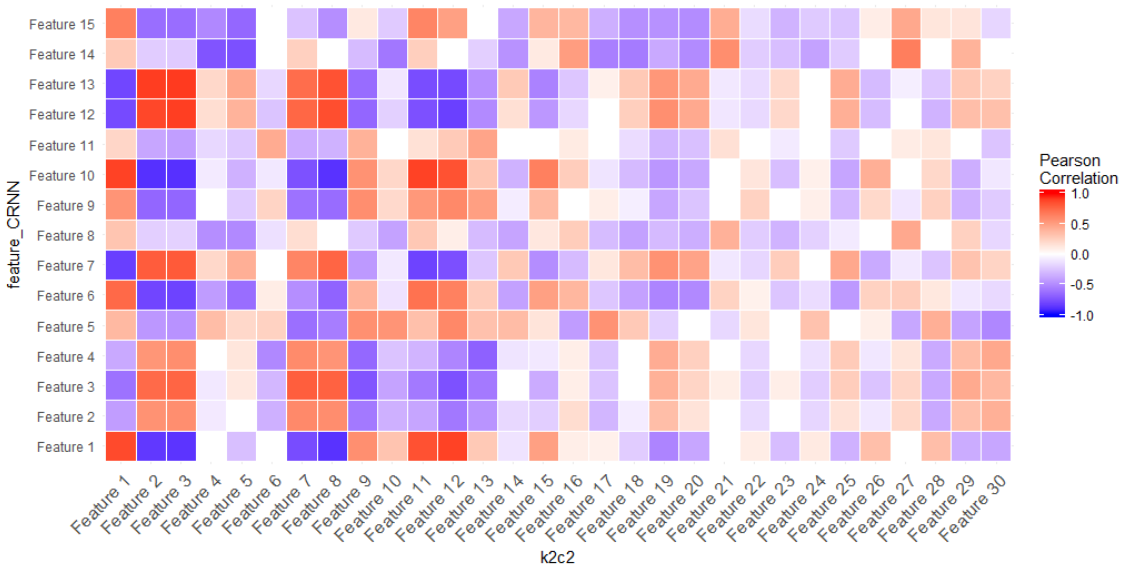

Figure 5.11 A heat map showing the correlation between the k2c2+ and feature-CRNN representations for the arousal condition. The significant Pearson correlation coefficients

(with Holme p value correction for multiple comparisons) are depicted by colour, on a scale from 1 (red) to -1 (blue).

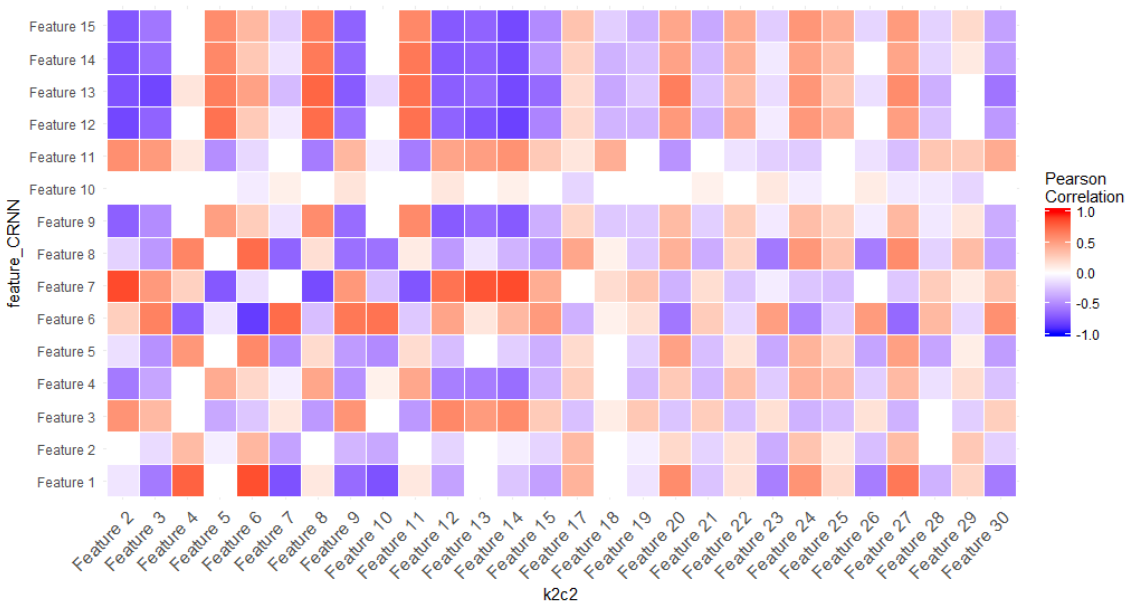

Figure 5.12. A heat map showing the correlation beween the k2c2+ and feature-CRNN representations for the valence condition. The significant Pearson correlation coefficients (with Holme p value correction for multiple comparisons) are depicted by colour, on a scale from 1 (red) to -1 (blue).

Next, I performed an auditory analysis of the musical features of the *CRNN_param_0* representations for arousal and valence to perceptually characterise the type of features that are being learned. This is particularly important to provide practical guidance of how researchers, practitioners, or AAC systems can apply these models to predict or manipulate music for emotion induction. To conduct this auditory analysis, I identified and analysed the top ten most-representative excerpts (i.e. that had resulted in the maximum activation of that particular feature) for each of the features identified by *CRNN_param_0*. This process resulted in the selection of 30 groups of excerpts in total for analysis (i.e. a group of excerpts for each of the 15 features, for each of the arousal and valence emotion conditions). The auditory analysis, revealed that high-level features related to components of time (e.g. rhythm and musical patterns), timbre (e.g. instruments and drones), modes (i.e. major and minor), and mixture between these were

represented within the model. A concise summary of all results from the auditory analysis is provided in Table 5.3. A full list of the top 10 music excerpts for each musical feature type is available in Appendix C.

Table 5.3 Summary of feature-CRNN musical feature types

Feature Type	**Emotion Condition**	**Feature Description**	**Feature Number**
Rhythm	Arousal	Pumping	1
		Driving	10
		Sparsity	9
	Valence	Firm	15
		Sparsity	6
Musical Patterns	Arousal	Harp and Pizzicato like Ostinatos	3
		Static, Repetitive, Long Decay	4
	Valence	Minor Ostinatos	4
		Stepwise and Arpeggio type patterns	10
		Rapid harp like arpeggios and musical trills	14
Drones	Arousal	Drones	7
		Long-Release	8
		Buzzy - Square waves	14
	Valence	Low range - buzzy	5
		Low range and rhythmic	13
Instrument	Arousal	Taiko and Bass Drums	5
		mid to high range sustained	6
		Brass	11
		Light Piano and Pizzicato	13
		Reverse Sweep Snare	15
	Valence	Sawtooth wave	3
		Noisy	9
		Bitcrush	12
Modal	Arousal	Low Tones and Minor Key	2
	Valence	Upbeat and Major Key	3
			7
		Slow tempo and Minor Key	8
		Slow chord progression that feature vocal or brass choirs	11

For example, the distinction between rhythm, timbres, and patterns are often blurred to human perception when elements such as speed come into play. However, auditory analysis revealed that the model picked up some distinctively rhythmic features (see Table C.4). Specifically, an auditory analysis of the top 10 excerpts of (a) Arousal_Feature_1 highlighted a pumping like rhythmic quality, (b) Arousal_Filter_10 seemed to capture a driving rhythm, and (c) Valence_Filter_15 appeared to capture a firm rhythmic quality. As such, these features all seem to capture an explainable and manipulable rhythmic quality.

Contrary to the relatively energetic rhythmic features presented above, Arousal_Feature_9 and Valence_Filter_6 seem to capture the temporal quality of sparsity (see Table C.5). Interestingly, while these two filters have a significant positive correlation with each other ($r = .43$, $p < .05$), they only highlight a few of the same excerpts. This indicates that while they both appear to capture similar qualities (i.e. from a human's perceptual capability), they are not synonymous emotional features.

Other rhythmic features seem to be related to specific musical patterns (see Table C.6). For example, in the arousal model, Arousal_Feature_4 seems to capture a quality characterized by static, repetitive rhythms with a long decay, while Arousal_Feature_3 picks up specific harp and pizzicato like ostinatos. Furthermore, in the valence model, Valence_Feature_4 captures minor ostinatos, Valence_Feature_10 captures rising stepwise and arpeggio like patterns, and Valence_Feature_14 is related to rapid harp-like arpeggios and musical trills.

Auditory analysis also revealed that several features also appear related to drones (see Table C.7). In the arousal model three features seemed to directly relate to drones: Arousal_Feature_7 (drones), Arousal_Feature_8 (long release - drone), and Arousal_Feature_14 (buzzy - squarewaves). In the valence model, drones seemed to be captured by Valence_Feature_1 (drones), Valence_Feature_5 (low range and buzzy), and Valence_Feature_13 (low rhythmic drones).

Some features capture specific instrument sounds (see Table C.8). In the arousal condition, Arousal_Feature_5 captures taiko and bass drums, Arousal_Feature_6 captures mid to high range sustained sounds, Arousal_Feature_11 captures brass sounds, Arousal_Feature_12 relates to strings, Arousal_Feature_13 captures light pianos and pizzicatos, and Arousal_Feature_15 identifies the effect of reverse percussion sweeps (often used as an effect to relay suspense). Valence_Feature_2 captures smooth sawtooth-like sounds, Valence_Feature_9 captures noisy instruments, Valence_Feature_12 captures a bit crushing effect (heavy metal-like).

There are also several features that appear to relate to mode (i.e. major or minor). The excerpts that make up the top list for Arousal_Feature_2 are characterized low tones in a minor key. Valence_Feature_3 seems to capture upbeat excerpts in a major key and Valence_Feature_7 similarly. Valence_Feature_8 highlights slow excerpts in minor keys. Mixing time, timbre and mode, Valence_Filter_11 highlights several excerpts that exhibit slow chord progression and feature vocal or brass choirs. Mode is more difficult to manipulate in loops, because in music, modal changes typically require some type of modulation. This challenge is however overcome by having loop libraries with loops designed to transition into each other. An example of this can be seen in Table C.9, where the EL_VOL02 library has excerpts activating several different modal features, meaning a system can compose within that library and easily navigate to different modes.

5.6.3 Within family manipulations

As outlined in the previous section, certain musical features appear to be directly related to emotional responses. This shows that it is theoretically possible to manipulate emotion through changing the musical features of an excerpt, so in this section I set out to explore a few concrete examples of how changing musical features might affect emotion. The goal of this section is not to perform a comprehensive analysis of all the features that can be manipulate to alter emotional responses, but to show the usefulness of loop families for AAC systems. As discussed in Section 5.3, I can investigate this by randomly dropping

voices, with the assumption that dropping voices from a loop will inherently result in the removal of a random selection of features as well. Thus, by comparing several variations of the loop (i.e. excerpts with differing voices dropped out) we should be able to get an indication of how certain musical features affect emotion.

Table 5.10 presents an example family of excerpts, consisting of Drums, Piano, Strings, Sub, and Synth components. The table outlines four variations of excerpt for this example (one is the full excerpt, and the other three excerpts have random voices dropped out). For this example, as seen in Figure 5.13. A column graph showing the arousal feature-CRNN activations for the group "Dys_90_D"., the feature that appears to cause the greatest difference in the arousal space is Arousal_Feature_8, which is related to a drone-like sounds. That sound is the synth sound, and when it is introduced the average arousal increases. As this is the sole difference between the full excerpt, "DYS_90_D_Full_SP_01.mp3" (Arousal M = 6.02, Valence M = 6.09) and "DYS_90_D_Time 98_1_1_1_1_0_.mp3" (Arousal M = 5.93, Valence M = 5.44), it can therefore be assumed to be responsible for the observed difference in arousal. This drone-like quality of the synth is also captured by Valence_Feature_1, so is also clearly a contributor to the difference in valence (see Figure 5.13). Another example within the valence space is the positive difference in valence that becomes apparent with the introduction of the piano and drums. The feature breakdown provided Figure 5.13, shows that this likely stems from the upbeat quality captured by Valence_Feature_7.

Table 5.4 Example Family 1: Manipulating emotion by changing musical features

Library	Group Name	Instruments	Arousal (average)	Valence (average)
Dystopian	DYS_90_D_Full_SP_01.mp3	Drums, Piano, Strings, Sub, Synth	6.02	6.09
	DYS_90_D_Time 98_0_0_1_1_0_.mp3	Strings, Sub	4.96	3.75
	DYS_90_D_Time 98_0_0_1_1_1_.mp3	Strings, Sub, Synth	5.46	5
	DYS_90_D_Time 98_1_1_1_1_0_.mp3	Drums, Piano, Strings, Sub	5.93	5.44

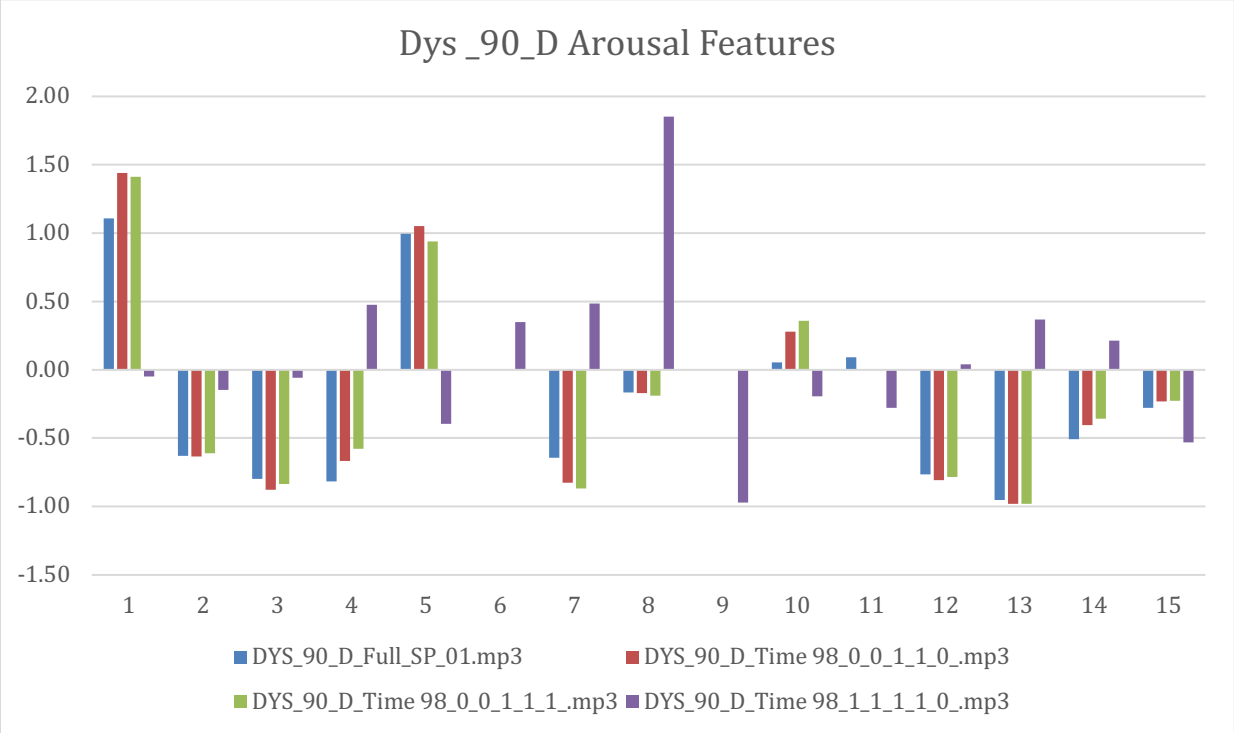

Figure 5.13. A column graph showing the arousal feature-CRNN activations for the group "Dys_90_D". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

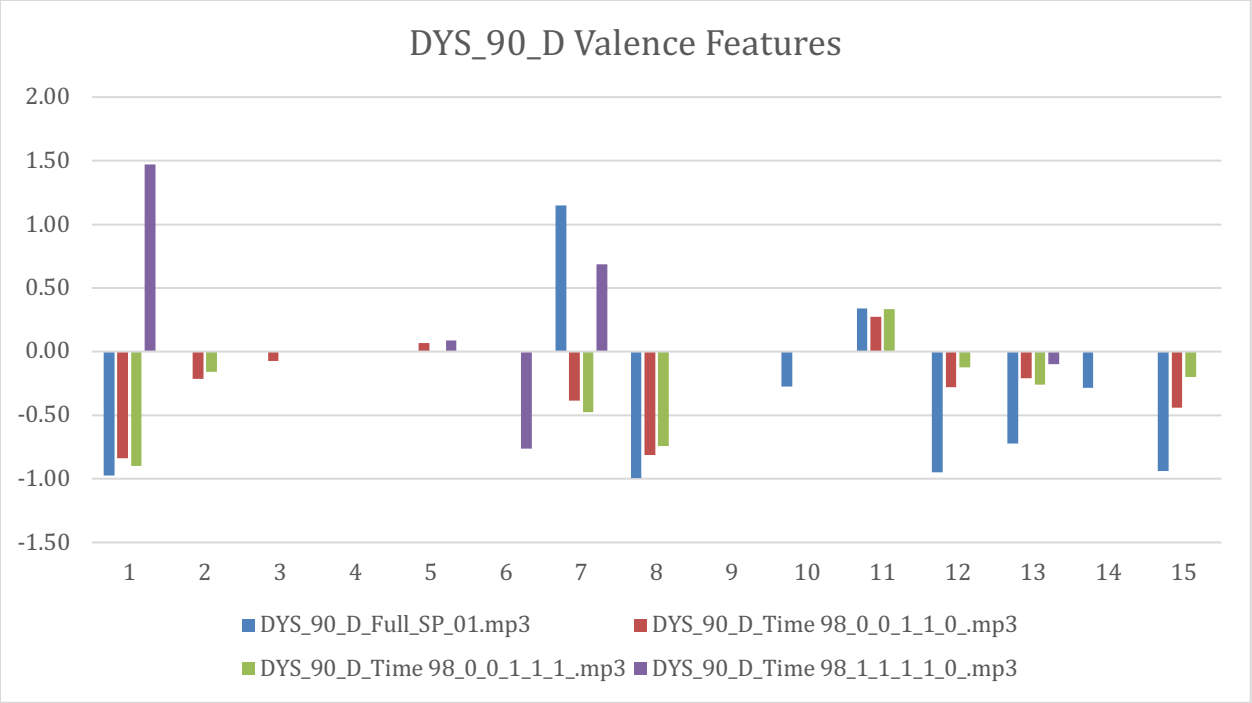

Figure 5.14. A column graph showing the valence feature-CRNN activations for the group "Dys_ 90_ D". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

The second example family I examined consists of Choir, ChoirStrings, DrumPercBrass, Harp, Piano, TunedPercussion, and UprightBass (see Table 5.5). The feature-CRNN activations for this group are presented in Figure 5.15 and Figure 5.16, for arousal and valence respectively. The introduction of the choir seems to have a suppressing effect on the tuned percussion, as realised in Arousal_Feature_4 (Static, Repetitive, Long Decay), Arousal_Feature_3 (Harp and Pizzicato like Ostinatos), and Arousal_Feature_13 (Light Piano and Pizzicato). This suppression results in a increase in arousal for the family. The tuned percussion has positive effect on valence, activating Valence_Feature_3 (Upbeat Major Key) and Valence_Feature_7 (Upbeat Major Key), increasing positive valence when it is introduced. An AAC would exploit the parameterisation of this loop to intentionally manipulate emotion.

Table 5.5 Example Family 2: Manipulating emotion by changing musical features

Library	Group_Name	Instruments	Arousal	Valence
MusicForMedia	MFM_100_F#_01_0_0_0_1_1_1_1_.mp3	Harp, Piano, TunedPercussion, UprightBass	4.06	6.95
	MFM_100_F#_01_1_0_0_1_1_0_1_.mp3	Choir, Harp, Piano, UprightBass	4.43	5.32
	MFM_100_F#_01_1_1_0_1_1_0_0_.mp3	Choir, ChoirStrings, Harp, Piano	4.63	5.40
	MFM_100_F#_Full_SP.mp3	Choir, ChoirStrings, DrumPercBrass, Harp, Piano, TunedPercussion, UprightBass	5.09	7.32

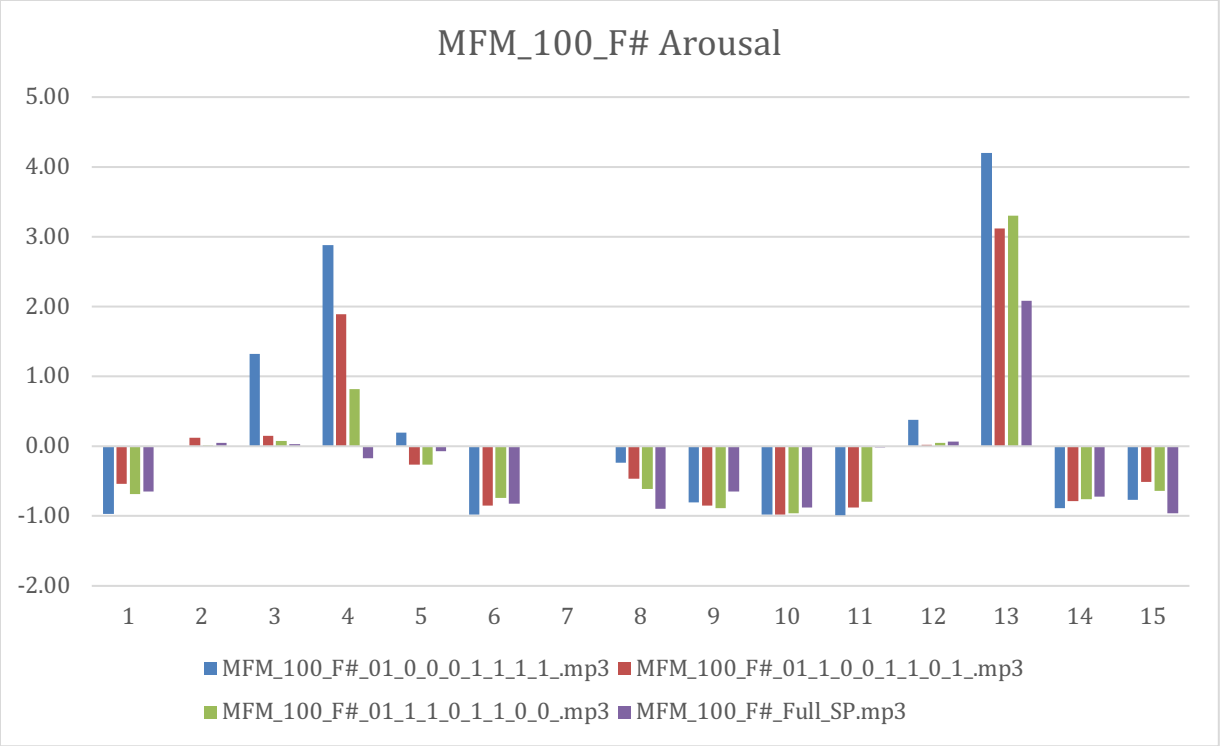

Figure 5.15. Arousal feature-CRNN activations for the group "MFM_100_F#". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

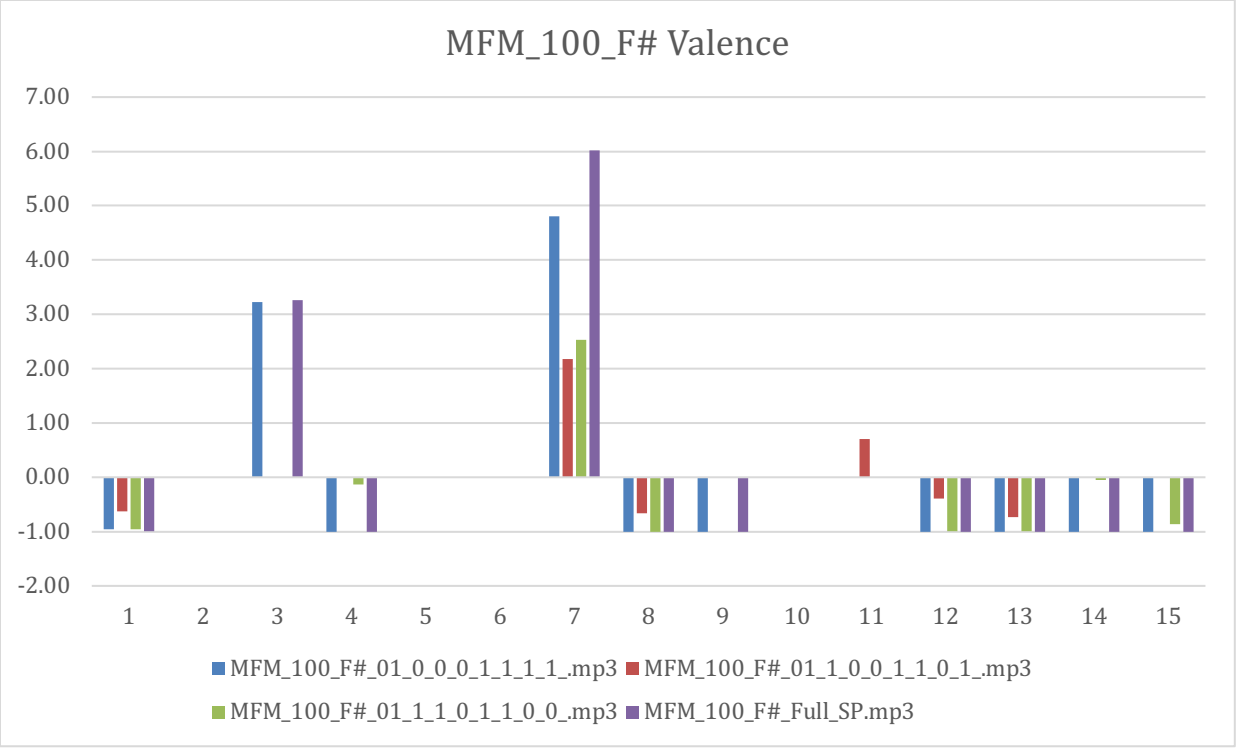

Figure 5.16. Valence feature-CRNN activations for the group "MFM_100_F#". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

The final group I examine consists of Basses, Choir, PercussionA, PercussionB, Piano, ViolasViolins, and Vocal (Table 5.6). The feature-CRNN activation for the arousal and valence models are shown in Figure 5.17 and Figure 5.18 respectively. In this group, the ostinato in the strings is captured by Arousal Feature 3 and the static repetitive nature is captured by Arousal Feature 4, which together lead to an increase in average felt arousal. The choir creates positive valence with its introduction, activating valence feature 11 with its slow vocal chord progression, but the effect is somewhat masked with the addition of percussions. An AAC based on the personalised emotion prediction system could learn this relationship from the features to automatically manipulate the voices and induce targeted emotional responses.

Table 5.6 Example Family 3: Manipulating emotion by changing musical features

Library	Group_Name	Instruments	Arousal	Valence
MusicForMedia	MFM_120_C_01_1_0_0_1_0_0_1_.mp3	Percussion B, Piano, ViolasViolins, Vocal	4.79	4.47
	MFM_120_C_01_1_1_0_0_0_0_0_.mp3	Basses, Percussion B, Piano, Vocal	4.55	6.07
	MFM_120_C_01_1_1_0_0_1_1_0_.mp3	Basses, Choir, Percussion B, Piano	5.34	6.93
	MFM_120_C_Full_SP.mp3	Basses, Choir, Percussion A, Percussion B, Piano, ViolasViolins, Vocal	5.24	6.56

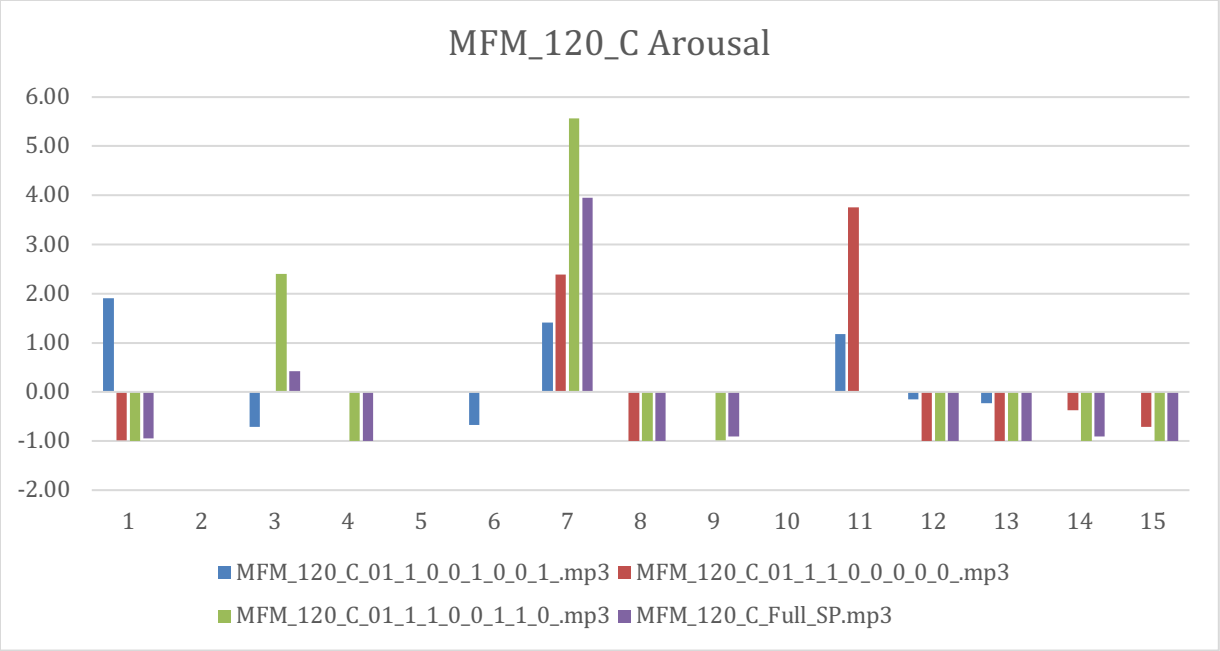

Figure 5.17. Arousal feature-CRNN activations for the group "MFM_120_C". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

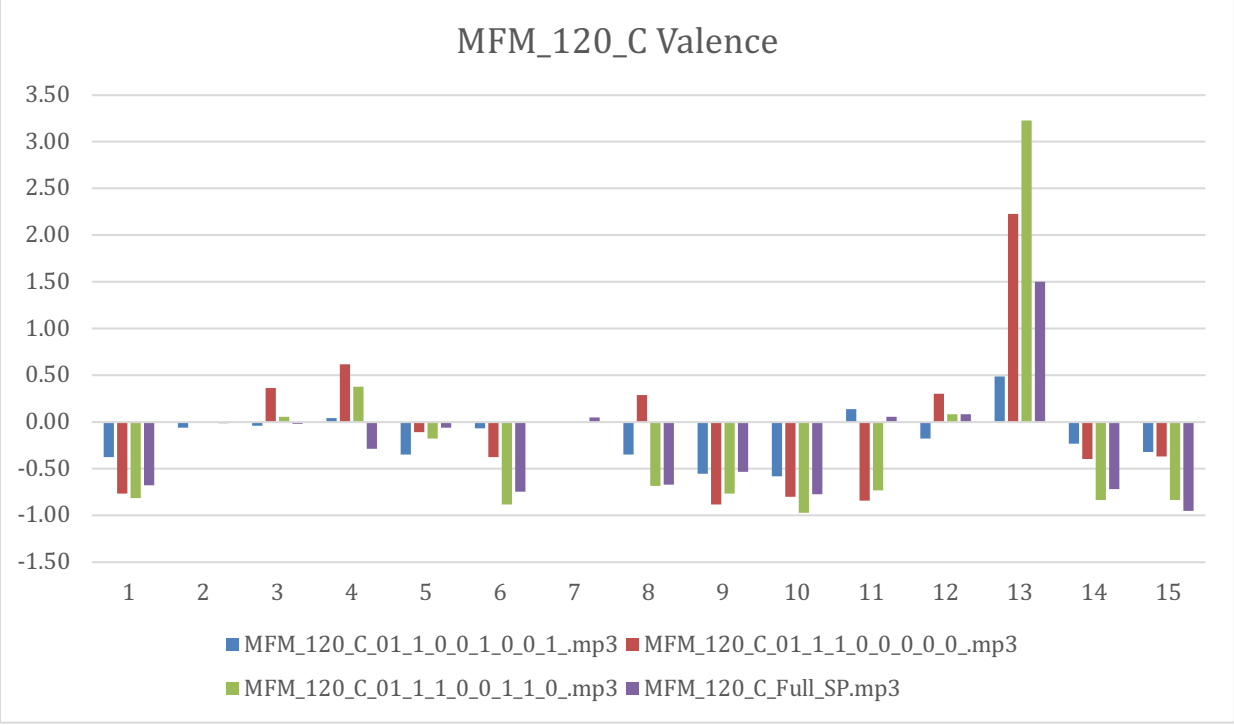

Figure 5.18 Valence feature-CRNN activations for the group "MFM_120_C". The y-axis is the activation of the feature, the x-axis is the feature number, and the colours represent the different variants of the loop family.

5.7 Chapter Conclusion

The goal of this chapter was to first validate that the personalised emotion prediction systems, presented in Chapter 4, are better suited for the task of emotion prediction than non-personalised emotion prediction. Many individual differences factor into an individual's emotional experience when listening to an excerpt of music and failure to account for these differences can result in high variability and contrast in emotional response. In this chapter, I made three modifications to the previous study. First, I increase the number of participants to 1943 from 120 previously. This adjustment allows for the examination of a larger number of individual responses to music, while also increasing the chances of identifying more closely aligned participants to base predictions on. Second, I increase the number of excerpts that are evaluated to 1307 from 120 previously. This allows participants to rate many more examples and provides a wider range of musical features for the content-based filtering techniques to learn. Third, I reduce the musical difference between excerpts, allowing the content-based filtering techniques to learn at a more resolute level what manipulations of musical features result in changes in individuals' emotional responses.

This chapter shows that collaborative filtering is a useful approach when a database of ratings is available for a defined stimulus set. The collaborative filtering approaches, user-based and item-based, outperform all other personalised and non-personalised techniques in predicting emotional responses. They are limited however, in that they are constrained to their pre-rated stimulus set. The content-based approaches, CB-CRNN and distance-weighted knn, also outperform non-personalised recommendation, and because they rely on musical features rather than other ratings exclusively, they can be used to predict a participant's ratings on novel items. The results of these systems validate the use of personalised music emotion prediction systems as better suited than non-personalised emotion prediction systems at accounting for individual differences in emotional responses. These studies show the potential of personalised emotion prediction

systems to reduce the variability in emotional research studies, and more reliably induce intended affect in individuals.

A secondary goal was to show that emotion could be manipulated within a musical composition by changing the musical features within an excerpt. Music manipulation could be useful for applications such as film scoring, gaming, and musical therapy. Furthermore, AAC systems could learn to manipulate emotional responses by changing the musical features of a musical excerpt. In this chapter, I used loops to construct musical excerpts – dropping different instrumental voices in and out of musical compositions. Through analysis of the emotional musical features, and of groups of loops, I showed that it is indeed possible to manipulate emotional responses to music using loops – associating loops within a group with musical features that had a direct impact on emotional response. These results are promising for researchers and composers who may want to prepare stimulus pre-induction or understand the musical factors contributing to emotional responses. The results are also promising for composers who work in industries like gaming, and must compose emotional music for different scenarios.

Chapter 6 Conclusion

Music is tremendously effective in its ability to induce emotional responses in people, and thus is amongst the most commonly used tools in both emotion induction research and applications such as film and videogame scoring where music is used to set cinematic moods or manipulate levels of tension (Baumgartner et al., 2006; Juslin & Laukka, 2004; Kenealy, 1988; Zentner et al., 2008; Zhang et al., 2014). However, very little research has been dedicated to identifying and/or accounting for the factors that lead to individual differences in peoples' emotional responses (Frieler et al., 2013; Juslin & Västfjäll, 2008). An individual's emotional response can be influenced by a wide range of factors: including, to name just a few, their personality (Vuoskoski & Eerola, 2011b, 2011a; Vuoskoski et al., 2012), episodic memories and social influences (Evans & Schubert, 2008; Juslin & Västfjäll, 2008), and their musical background (Belcher & Haridakis, 2013; Juslin & Laukka, 2004). In practice, the lack of understanding of these factors has made it difficult to both (a) reliably create emotion inducing stimulus sets, and (b) forecast how an individual will respond to a given musical stimulus. In this thesis, I rectify these two issues with the development of a personalised music affect induction system that accounts for individual differences in emotional response.

6.1 Achievements

There is a clear gap in research in which a more effective and standardized process is required for the selection of music stimuli. A comprehensive review of 250 music and emotion studies (Eerola & Vuoskoski, 2013) revealed at least three major pitfalls in the selection of music stimuli:

1. Forty-eight percent of researchers used familiar classical music excerpts for emotion induction.

2. In terms of who chose the stimuli:
    a. Thirty-three percent of researchers arbitrarily chose their stimulus sets.
    b. Thirty-nine percent of researchers did not reveal their selection method at all. [10]

The best music for emotion induction should not only effectively induce affect, but also minimise the role of subjective musical preference, which is often associated with familiarity and genre preferences. For example, emotional responses to familiar music can be affected by factors such as episodic memories (Evans & Schubert, 2008). A researcher may select a seemingly happy song with the intention of inducing a positive emotional response, however some individuals may associate the familiar music with a contradictory memory (e.g. a sad event), leading to an emotional response that diverges from the research intentions. Furthermore, a researcher's own perception of and emotional response to, music is subject to their individual experiences and preferences, and does not necessarily reflect how others may respond to the same musical stimuli. A more rigorous process is needed to guide researchers' selection of stimulus sets for emotional induction. In Chapter 3, I introduce a preprocessing step to the stimulus selection method, by which a committee of music experts first select music excerpts which they believe will induce the intended emotion, before a broad base of participants provides emotion ratings, on the arousal and valence dimensional scales, to each excerpt to confirm the experts' initial classifications. Furthermore, to alleviate the effect of genre preferences and maximize effectiveness in emotional response, I limit the stimulus set to film music because it is designed with the express intent of inducing emotion in wide and diverse audiences.

Another concern addressed in Chapter 3 is that not much research has explored the effects of modern (say electronic-based) music on emotion. Researchers have recognized that the ability to manipulate the components of musical compositions can be

[10] Another 8% percent used a Pilot study, 9% used a previous study, 6% used an Expert panel, and 4% used Participants for selection.

used to determine the causal relationship between musical features and emotional responses, but often lack the ability to develop such compositions themselves (Eerola et al., 2013). The benefit of exploring electronic music for the personalised emotion prediction system is that the music can be much more amenable than orchestral music to the manipulation of musical parameters. I address this concern in Chapter 3 by employing an electronic-music based stimulus set, which I use in three ways. First, I use the stimulus set to validate that modern electronic-based music can be as effective in inducing emotion responses as orchestral music. This provides both a new validated stimulus set for researchers to use, and opens a new paradigm of study for the use of modern electronic-based music in psychological research as an alternative to familiar classical music. Second, I identify music components within the electronic-based music that contribute to averaged emotional responses, a domain that has rarely been explored in modern electronic-based music. Finally, I show that certain features of the electronic music could be used to predict emotional responses in the arousal and valence dimension. This research shows promise in forming predictive emotional models based on musical features, and provides high-level insight about the types of features that could be manipulated in electronic-based music to influence emotional responses.

After exploring electronic music in Chapter 3, by (a) showing the promise of modern electronic music's ability to predict emotional responses to music, (b) validating that electronic-based music can be used to effectively induce emotions across the range of the arousal and valence scales, and (c) showing that the features of the music can form predictive models of emotion responses, Chapter 4 takes this research further, to account for individual differences in emotion responses. In practice, failing to account for individual differences in emotional responses leads to inconsistent results, failures to replicate in psychological studies (Frieler et al., 2013; Juslin & Västfjäll, 2008), and the inability to systematically control musical variables (Juslin & Sloboda, 2011). In Chapter 4, I use clustering techniques as an exploratory step to show, at a coarse level, that within the

averaged emotional ratings there exist several cohorts of individuals with differing emotional responses to musical features. In fact, a further examination of these cohorts and their emotional responses to musical features show that in many cases people can have contrasting emotional responses to the same music (see Section 4.3.3). This step further validates the need for a more personalised emotion prediction system to account for individual differences in emotion responses to music.

I present several solutions to personalised music affect induction with the final study in Chapter 4, based on techniques used in modern recommendation engines. I first explore collaborative filtering techniques – user-based collaborative filtering techniques rely on the ratings of other similar users to make predictions about how an individual will rate a novel item, while item-based collaborative filtering techniques rely on similar items (as determined by the ratings of other people) to predict how an individual will rate a novel item. These techniques essentially capture latent information that could be the result of factors such as social contagion and cultural preferences without having to explicitly code them. In Chapter 4, collaborative-filtering techniques outperform all other techniques, including non-personalised, in predicting individuals' emotion responses to music. The limitation of these techniques is that they require a pre-rated database of music stimuli to base their predictions on and it isn't easy to add new items. This could be a challenge if, for example, the researchers wanted to introduce new music or use completely different stimulus sets, but is the most effective technique in settings where this is not the case.

The second approach to personalised emotion prediction I implement in Chapter 4 is content-based filtering techniques – an approach that utilises the individuals' ratings on other similar musical excerpts (i.e. determined by the musical features), to predict their emotional responses to novel music excerpts. In the first case, I used a standard distance-weighted knn approach with the musical features extracted in Chapter 3 to calculate an individual's ratings on novel music, showing that this approach to emotional response

prediction significantly outperformed non-personalised prediction. I also created a novel Siamese Convolutional-Recurrent Neural Network model that learned to both encode musical features from MFCC representations, and predict an individual's emotional response to novel musical excerpts based on a short history of their previous emotional responses. This approach further personalised the extraction of relevant emotional musical features, as the process of feature engineering was directly connected to the emotional responses they created rather than retrofitted to capture perceptive musical components. This approach significantly outperformed both the distance-weighted knn technique and the non-personalised approach to emotional response prediction. While the content-based techniques did not perform as well as the collaborative-filtering techniques, their benefit lies in the fact that they are not constrained to the dataset, nor the people they were trained on. This gives researchers the ability to extend the stimulus set, freely adding music excerpts to the existing stimulus set or fully replacing it. This could be useful if, for example, the excerpts become too familiar, the researcher wants to extend the study, or a composer wants to use the system on completely novel music.

With Chapters 3 and 4, I have (a) introduced a rigorous method for developing stimulus sets for music emotion induction, (b) shown that it is possible to use modern electronic-based music to reliably induce emotion across the range of the arousal and valence dimensions, and (c) implemented and developed techniques to account for individual differences and better predict emotional responses to music stimuli. In Chapter 5, I further extend my contributions by validating these techniques on a larger population of individuals, and with a larger set of musical stimuli. A larger and more musically diverse stimulus set afforded the content-based filters a broader range of musical features, testing their ability to generalize and therefore improving their robustness to potentially novel stimuli. A larger number of participants offered the collaborative-filtering techniques a higher probability of "tightly aligned" neighbours on which to predict a participant's emotional responses, and the content-based filters more data to train on. The performance

of both the collaborative and content-based filtering approaches was improved in Chapter 5.

Finally, in Chapter 5, I show how musical features can be used to manipulate emotional responses, by introducing the use of loop libraries. Loops are designed to be added or subtracted from a musical composition freely, so with an understanding of emotional music features we can use loops as tools to manipulate emotional response. By dropping different loops in and out of musical compositions and analysing the average emotional responses to these modified excerpts, I show that certain musical features captured in loops can indeed be used to manipulate emotional responses. The benefits of this research go beyond the psychological research applications, and extend to areas such as music therapy, marketing, film, and video game development, all of which rely on manipulating music to achieve affective goals (Juslin & Sloboda, 2013).

6.2 Future works

In this thesis, I have created robust systems for personalised emotion prediction that can be used to more reliably induce emotions in individuals. There are at least three research domains that will benefit from these systems. The first is basic emotion research - researchers can use the personalised emotion prediction system to more reliably induce and manipulate affect for the purpose of studying emotional responses. The second is in psychology of music and musicology research, in which researchers can use this system to better understand the components and structures of music that contribute to emotional responses. The third includes applied research, in which practitioners in music therapy, consumer marketing, film and video game making, and many more could use this system to create music that better helps them achieve their emotional goals. Given these areas of application, I foresee several logical extensions to the research I have developed, specifically relating to applications of different emotional models, the development of

affective algorithmic composition systems, and studies exploring emotion in longer form musical compositions.

While the focus in this thesis is on predicting dimensional emotion (arousal and valence), a logical extension to this research would be to explore discrete emotion, such as happiness, anger, surprise, fear, disgust, and sadness (Ekman, 1992). I chose to study dimensional models for multiple reasons:

a) Research has thus far ruled out the idea of one-to-one mappings between discrete emotions and specific regions in the brain (Hamann, 2012).
b) Discrete models have been shown to perform poorer in characterizing emotionally ambiguous music (Eerola & Vuoskoski, 2010).
c) Dimensional models provide a higher resolution of emotional responses than discrete models.

However, these reasons neither rule out the importance of discrete models (Harmon-Jones, Harmon-Jones, & Summerell, 2017), nor the potential practical applications of these models. For example, practitioners such as film and video game composers are often tasked with creating music to induce specific emotional responses, say fear or surprise, in order to match the accompanying storyline or scene. Personalised emotion prediction systems trained to predict discrete emotional states would afford this scenario with more ease than systems trained to predict dimensional responses. The development of stimulus sets and models for discrete emotions could be an interesting area to further explore with these techniques.

Another logical extension of this research relates to the development of affective algorithmic composition. In Section 5.6.3 I demonstrate how different loop configurations within a group affected average arousal and valence responses and explained the auditory analysis of the features in Section 5.6.2 to explain these changes. As a logical next step,

it would be interesting to develop an AI system that automatically maps the musical features of a loop library to changes in emotional responses and utilises this knowledge to automatically produce sequence-based compositions. This development could be accomplished using both the models and the stimulus set I have developed in my research.

Finally, it may be of interest to explore longer forms of music composition. In my research, I focus on the emotional responses to excerpts and short phrases of musical compositions. A logical extension to this work would be to examine longer form musical compositions, to determine if there are emotional effects related to contextual factors within a music itself. For example, if the music is mostly negative and then positive phrases are introduced, how do factors such as recent memory effect the current emotional response? A theoretical advantage of the Convolutional-Recurrent Neural Network, introduced in Section 4.5, is that it would capture and account for such temporal dependencies in its predictions. This would be an interesting area for researchers to further explore.

6.3 Chapter Conclusion

In this thesis, I have created a robust systems for personalised emotion prediction that will allow researchers and practitioners to more reliably induce emotions in individuals. In Chapter 3 I validate, for the first time, that modern electronic-based music stimuli can be just as effective in inducing affect as the popular orchestral stimuli that has traditionally dominated the domain of music emotion induction (Eerola & Vuoskoski, 2013). I also validate the usefulness of this development in Chapter 5 by, for the first time, showing that loop-based music stimuli sets, in which loops can be easily added to or subtracted from a musical composition, can be used as a tool to induce changes people's emotional state. These novel contributions will allow researchers and practitioners to a) create more amenable datasets that can be used to manipulate emotional responses and

to b) counteract the emotional effects of familiarity that are inherently present in the popular classical music that dominates current music emotion induction stimulus sets.

Furthermore, in Chapter 4 and Chapter 5, for the first time, I implement and validate deep learning and modern recommender system techniques as solutions for, accounting for the latent variables that lead to individual differences in emotional responses to music. This novel contribution to the research domain will enable researchers and practitioners to more precisely select emotional stimuli for individuals and to therefore more reliably achieve their emotional induction goals. The connection between musical features and emotional responses, learned for example by the novel Convolution-Recurrent Neural Network model, can be further utilised by affective algorithmic compositions systems, to automatically develop stimuli from the loop-based stimulus set in order to induce specific emotional responses in individuals.

Overall, I have provided the domain of emotion induction with a strong and robust system and methodology for selecting music stimuli for reliable and effective emotion induction in individuals. The novel contributions of this thesis open up innumerable possibilities to researchers and practitioners. It has the potential to form the basis for innovation in many more works in the domains of basic emotion research, psychology of music and musicology, and applied research fields such as music therapy, consumer marketing, film and video game making, and many more.

Appendix A

Table A.1 Electronic Music Excerpt Ratings.

Excerpt	Emotion (expert committee)	Album	Track	mm:ss	Ar (F)	Ar (P)	Va (F)	Va (P)	Like	Familiar
101	High Arousal	The Dark Knight	11	04:08-04:38	7.1	6.6	7.1	5.9	6.3	5.0
21	High Arousal	Far Cry 4	3	00:00-00:30	6.7	6.6	5.7	6.4	5.4	4.5
5	High Arousal	The Girl with the Dragon Tattoo	28	02:14-02:44	5.7	5.9	4.9	4.9	3.7	3.4
37	High Arousal	The Social Network	3	00:41-01:11	6.6	6.6	6.3	5.2	4.3	4.1
38	High Arousal	The Dark Knight	10	00:12-00:42	7.5	6.8	7.0	5.9	6.4	5.8
149	High Arousal	Spring Break	1	00:00-00:28	6.8	7.8	6.7	6.3	5.9	5.4
150	High Arousal	The Girl with the Dragon Tattoo	28	02:44-03:05	6.8	6.4	4.7	4.1	4.1	3.7
69	High Arousal	Spring Break	10	00:32-00:59	6.3	7.5	5.5	6.3	4.3	3.7
53	High Arousal	Far Cry 4	12	01:10-01:40	6.6	7.8	4.2	4.3	4.5	4.4
117	High Arousal	The Social Network	12	01:53-02:20	6.8	7.2	6.1	5.7	6.0	6.3
102	High Arousal	The Girl with the Dragon Tattoo	16	02:24-02:54	5.6	5.7	6.0	5.1	5.2	3.9
151	High Arousal	Spring Break	10	00:04-00:31	6.1	6.5	6.2	5.2	5.2	4.1
118	High Arousal	Far Cry 4	19	00:30-01:00	6.5	6.9	6.3	5.3	5.1	4.0
85	High Arousal	Far Cry 4	24	01:50-02:13	5.9	6.5	4.7	4.1	4.6	3.7
6	High Arousal	The Girl with the Dragon Tattoo	4	00:44-01:03	5.3	7.0	6.0	5.5	5.0	4.4

39	High Arousal	A Series of Unfortunate Events	29	00:32-01:01	5.3	5.9	6.1	6.0	5.7	4.5
7	High Arousal	The Road to Perdition	20	00:15-00:39	7.5	7.6	6.0	5.3	5.5	4.9
133	High Arousal	The Girl with the Dragon Tattoo	36	05:30-06:00	7.4	6.9	5.0	3.9	4.6	4.0
119	High Arousal	Gone Girl	18	00:12-00:32	4.8	5.2	3.5	3.5	3.6	2.9
70	High Arousal	Far Cry 4	14	00:00-00:30	6.4	6.9	5.4	4.8	5.6	4.4
86	High Arousal	Three	1	00:00-00:19	6.5	6.7	5.6	5.0	5.2	4.3
71	High Arousal	Far Cry 4	26	00:16-00:46	7.5	7.5	4.7	4.8	4.7	4.3
134	High Arousal	Three	17	00:48-01:18	7.2	7.1	5.8	4.8	5.7	5.2
54	High Arousal	A Series of Unfortunate Events	11	01:10-01:40	4.7	4.7	4.7	4.5	5.2	4.5
103	High Arousal	The Girl with the Dragon Tattoo	17	01:33-01:55	5.3	5.9	4.0	2.6	3.4	3.2
22	High Arousal	Three	5	02:03-02:27	7.4	7.3	4.6	3.7	4.4	4.9
23	High Arousal	Far Cry 4	5	00:01-00:30	6.0	6.8	5.5	5.7	4.9	4.4
135	High Arousal	Three	15	01:30-01:55	7.4	7.2	5.7	4.3	5.3	4.3
55	High Arousal	The Lego Movie	5	00:20-00:50	6.8	7.1	5.9	5.9	4.8	4.4
87	High Arousal	The Social Network	5	00:00-00:28	5.3	5.6	6.1	5.5	5.4	4.3
40	Low Arousal	The Social Network	14	00:00-00:20	5.2	4.1	4.5	3.1	3.9	3.6
136	Low Arousal	Gone Girl	11	01:00-01:30	5.6	3.6	4.7	3.8	4.5	3.9
8	Low Arousal	Gone Girl	3	00:23-00:53	3.3	3.2	4.1	4.3	4.8	4.2
152	Low Arousal	The Knick	7	00:15-00:35	5.6	4.2	6.3	6.1	5.7	4.4

56	Low Arousal	The Knick	18	00:39-01:09	4.0	4.2	5.9	6.1	5.7	4.3
120	Low Arousal	The Road to Perdition	3	00:18-00:48	4.5	2.6	5.4	5.1	5.7	4.3
104	Low Arousal	Gone Girl	12	01:50-02:20	5.0	3.8	6.2	4.9	5.7	4.5
72	Low Arousal	The Knick	16	00:09-00:27	5.1	3.5	3.9	4.0	4.9	4.2
24	Low Arousal	American Beauty	14	00:00-00:30	2.7	3.1	4.4	3.4	4.7	4.3
88	Low Arousal	The Girl with the Dragon Tattoo	22	00:46-01:16	4.1	4.1	6.0	4.8	5.5	3.9
89	Low Arousal	Far Cry 4	9	01:55-02:15	4.9	4.5	6.5	5.7	5.3	3.9
73	Low Arousal	A Series of Unfortunate Events	24	02:08-02:28	5.1	4.5	4.0	3.3	4.3	3.9
137	Low Arousal	The Knick	4	00:00-00:23	5.7	3.8	5.7	4.6	5.3	4.3
41	Low Arousal	The Girl with the Dragon Tattoo	13	03:34-03:57	5.4	5.0	4.7	4.4	4.9	4.3
121	Low Arousal	The Knick	18	01:18-01:37	4.8	3.5	4.8	5.3	5.4	4.2
138	Low Arousal	Far Cry 4	30	02:00-02:30	4.8	3.6	5.6	5.3	5.8	4.6
105	Low Arousal	The Girl with the Dragon Tattoo	5	00:00-00:30	4.1	4.1	2.9	3.1	2.9	2.9
74	Low Arousal	Gone Girl	2	00:00-00:32	3.3	2.9	5.8	5.7	6.1	4.8
9	Low Arousal	Gone Girl	23	00:30-00:53	3.5	3.9	4.9	4.1	4.9	4.1
57	Low Arousal	Batman	1	00:00-00:30	4.6	4.3	4.4	3.7	4.6	3.6
106	Low Arousal	The Knick	19	00:10-00:31	3.8	3.7	4.7	4.5	4.7	3.9
153	Low Arousal	The Lego Movie	2	00:40-01:00	6.6	6.6	4.7	4.0	4.4	3.7
58	Low Arousal	The Social Network	12	00:20-00:30	4.3	4.7	6.2	6.0	5.7	6.2

122	Low Arousal	American Beauty	1	00:00-00:30	5.3	4.1	6.8	5.9	5.8	5.6
10	Low Arousal	The Lego Movie	4	00:00-00:30	4.0	3.8	6.6	6.8	6.0	5.1
154	Low Arousal	American Beauty	2	00:00-00:30	5.5	3.7	6.3	6.1	6.2	4.7
90	Low Arousal	The Dark Knight	7	00:14-00:40	5.2	5.1	3.7	2.6	3.9	3.8
42	Low Arousal	The Girl with the Dragon Tattoo	24	03:05-03:32	5.4	5.5	5.3	5.7	5.2	4.0
25	Low Arousal	The Road to Perdition	1	00:05-00:35	2.3	2.9	3.9	4.1	4.0	3.7
26	Low Arousal	American Beauty	3	00:00-00:30	4.0	4.1	5.8	5.7	5.4	4.3
59	Negative Valence	The Girl with the Dragon Tattoo	5	01:42-02:05	5.1	4.2	3.3	3.0	3.7	3.7
27	Negative Valence	The Road to Perdition	5	01:00-01:22	4.0	3.6	3.6	3.6	3.7	3.7
11	Negative Valence	American Beauty	14	00:00-00:30	2.6	2.9	4.0	4.5	4.7	4.2
12	Negative Valence	Gone Girl	4	00:16-00:37	3.6	4.8	4.0	2.3	3.2	3.4
107	Negative Valence	Gone Girl	7	00:00-00:32	4.1	3.5	4.9	4.4	5.2	3.9
139	Negative Valence	Far Cry 4	5	00:00-00:30	6.9	6.4	6.1	4.6	5.5	4.4
13	Negative Valence	The Girl with the Dragon Tattoo	27	01:47-02:14	5.8	4.8	3.6	4.3	4.1	3.9
140	Negative Valence	Far Cry 4	1	02:14-02:38	5.5	4.2	5.2	4.3	4.6	4.0
60	Negative Valence	Cinderella	6	00:00-00:30	4.9	4.1	2.9	2.3	3.5	3.3
75	Negative Valence	The Lego Movie	23	01:18-01:48	4.7	4.3	6.5	6.1	7.0	5.5
123	Negative Valence	Far Cry 4	20	00:00-00:30	5.7	4.5	4.9	4.6	5.2	4.6
124	Negative Valence	Three	3	01:24-01:51	6.0	4.6	4.3	3.9	4.4	3.9
28	Negative Valence	Batman	8	02:30-03:00	6.2	5.9	4.7	3.9	4.6	4.6

61	Negative Valence	The Girl with the Dragon Tattoo	21	00:56-01:23	4.4	3.2	3.9	2.9	4.7	3.9
76	Negative Valence	The Social Network	7	00:30-01:00	4.5	4.9	4.1	3.5	3.9	3.5
43	Negative Valence	The Girl with the Dragon Tattoo	6	00:04-00:34	3.4	3.6	6.3	6.1	6.8	5.3
44	Negative Valence	The Knick	8	00:46-01:10	4.3	5.2	5.5	5.2	4.6	3.6
155	Negative Valence	Batman	8	00:30-01:00	6.6	5.8	5.8	5.1	6.1	5.1
45	Negative Valence	Far Cry 4	28	00:00-00:25	6.9	7.4	4.5	4.7	4.2	3.9
91	Negative Valence	Three	5	01:18-01:43	5.9	5.1	3.9	3.7	4.6	4.1
77	Negative Valence	The Social Network	1	01:15-01:40	4.3	3.8	3.4	3.5	3.8	3.6
108	Negative Valence	The Social Network	18	00:53-01:16	4.4	3.9	3.5	3.7	4.4	3.7
125	Negative Valence	Gone Girl	16	01:30-01:55	6.4	5.0	4.1	4.2	4.7	3.9
92	Negative Valence	Gone Girl	19	01:00-01:20	5.1	5.2	3.1	3.5	4.5	3.6
93	Negative Valence	The Road to Perdition	5	00:11-00:31	4.6	3.9	3.6	3.0	4.2	3.9
156	Negative Valence	Far Cry 4	7	00:00-00:30	7.5	7.7	5.9	4.5	5.1	4.6
109	Negative Valence	The Social Network	7	00:23-00:53	4.5	4.7	2.8	3.1	3.4	2.8
157	Negative Valence	Three	18	00:07-00:30	6.2	5.8	4.0	3.8	4.6	3.9
29	Negative Valence	Nemo	34	00:36-01:06	6.4	6.5	5.3	4.7	4.6	4.2
141	Negative Valence	The Dark Knight	12	01:30-02:00	6.9	7.0	6.1	5.0	6.3	5.1
158	Positive Valence	A Series of Unfortunate Events	29	02:00-02:22	5.0	5.5	6.4	5.5	5.7	4.3
78	Positive Valence	The Lego Movie	13	00:15-00:45	7.6	7.4	6.4	6.4	4.6	4.4

126	Positive Valence	The Lego Movie	24	00:36-01:06	5.8	5.9	7.1	7.1	6.4	5.4
62	Positive Valence	Nemo	1	00:52-01:09	3.8	5.1	6.7	7.3	5.6	4.7
63	Positive Valence	The Girl with the Dragon Tattoo	6	02:44-03:15	5.6	5.4	5.7	5.7	5.9	4.6
46	Positive Valence	Nemo	3	00:00-00:30	3.2	2.8	6.3	6.2	6.2	5.1
14	Positive Valence	Nemo	28	01:06-01:23	4.3	4.1	6.7	6.1	6.2	5.1
94	Positive Valence	The Knick	1	01:46-02:16	6.1	5.7	6.9	5.6	5.8	4.4
15	Positive Valence	Far Cry 4	10	01:10-01:39	4.7	4.4	5.5	5.9	5.4	4.7
127	Positive Valence	Spring Break	3	00:49-01:19	4.6	3.8	6.4	5.4	5.4	4.4
142	Positive Valence	Nemo	3	00:12-00:35	5.1	2.7	6.1	6.6	6.2	5.2
30	Positive Valence	Far Cry 4	18	00:00-00:28	6.2	5.9	5.5	5.3	5.1	4.3
16	Positive Valence	Spring Break	4	00:00-00:30	6.0	6.3	5.5	6.9	4.9	5.0
110	Positive Valence	The Lego Movie	13	00:30-01:00	6.8	6.7	7.1	6.9	5.6	4.8
143	Positive Valence	Far Cry 4	25	00:00-00:30	6.3	5.6	5.9	4.3	5.2	4.3
79	Positive Valence	Spring Break	18	00:30-01:00	3.7	3.7	6.2	6.1	6.1	5.1
31	Positive Valence	The Lego Movie	7	00:00-00:30	6.3	6.0	6.8	6.3	6.2	5.8
95	Positive Valence	The Dark Knight	12	01:06-01:26	6.6	6.4	7.8	6.4	6.5	5.7
80	Positive Valence	The Knick	14	00:00-00:30	4.7	5.3	5.6	6.1	5.3	4.3
47	Positive Valence	The Lego Movie	3	01:31-01:55	7.1	7.4	7.2	6.9	5.5	5.0
159	Positive Valence	Spring Break	7	01:30-01:50	5.9	4.6	7.0	6.4	5.8	4.6
96	Positive Valence	American Beauty	4	00:30-01:00	5.4	4.0	7.1	6.2	6.4	4.4

48	Positive Valence	A Series of Unfortunate Events	5	01:10-01:40	3.2	3.1	6.9	6.5	6.7	5.2
144	Positive Valence	The Road to Perdition	26	02:30-02:53	4.7	3.7	5.5	6.8	6.3	5.1
32	Positive Valence	Nemo	5	00:26-00:42	5.2	4.4	7.0	7.1	6.2	5.4
111	Positive Valence	The Girl with the Dragon Tattoo	6	00:25-00:44	4.7	3.1	5.7	4.9	6.4	4.6
128	Positive Valence	A Series of Unfortunate Events	5	01:21-01:41	4.8	3.0	5.8	5.6	6.0	4.8
64	Positive Valence	A Series of Unfortunate Events	21	00:05-00:34	3.4	3.5	6.5	6.3	5.7	4.3
112	Positive Valence	The Social Network	8	01:25-01:52	5.5	4.9	6.9	5.3	5.6	4.2
160	Positive Valence	Nemo	21	01:36-01:56	6.2	5.9	6.6	5.8	5.5	4.6

Table A.2 Music Excerpts from Eerola & Vuoskoski (2010)

Excerpt	**Emotion**	**Album**	**Track**	**mm:ss**	**Ar(F)**	**Ar (P)**	**Va (F)**	**Va (P)**	**Like**	**Familiar**
17	High Energy	Batman	18	00:55-01:15	6.7	6.5	7.7	7.0	6.5	6.1
65	High Energy	Man of Galilee CD1	2	03:02-03:18	5.7	7.1	6.8	7.9	6.1	5.7
113	High Energy	The Untouchables	6	01:50-02:05	6.4	5.2	7.3	7.4	6.0	5.4
129	High Energy	Shine	5	02:00-02:16	5.2	6.3	7.2	6.7	6.2	5.0
145	High Energy	Shine	15	01:00-01:19	4.4	5.1	6.9	6.6	5.7	4.4
1	High Energy	Juha	2	00:07-00:18	4.2	4.2	6.2	6.3	5.1	4.7
33	High Energy	Lethal Weapon 3	4	01:40-02:00	3.7	4.4	5.5	6.9	5.2	4.6
49	High Energy	Crouching Tiger, Hidden Dragon	13	01:52-02:10	5.0	5.7	6.1	5.3	6.3	5.2

81	High Energy	Oliver Twist	7	01:30-01:46	5.5	3.6	7.5	6.2	6.6	5.7
97	High Energy	Batman	4	02:31-02:51	6.1	5.9	7.3	7.3	6.3	6.2
2	Low Energy	Blanc	16	00:00-00:15	3.2	3.5	3.5	4.1	4.6	4.2
34	Low Energy	Road to Perdition	16	00:17-00:32	4.0	4.2	5.3	4.7	5.3	4.5
50	Low Energy	Blanc	10	00:13-00:31	2.6	3.5	5.4	4.3	5.6	5.0
66	Low Energy	Batman Returns	12	00:57-01:14	3.5	2.7	4.8	4.7	4.9	4.2
114	Low Energy	Running Scared	15	02:06-02:27	4.9	4.2	4.8	3.8	5.1	4.2
18	Low Energy	Blanc	18	00:00-00:16	3.6	2.9	4.9	4.3	5.1	5.1
82	Low Energy	Big Fish	15	00:55-01:11	4.2	4.1	5.6	4.8	5.6	4.3
98	Low Energy	Big Fish	11	01:26-01:40	5.2	3.9	6.3	5.9	6.3	5.0
130	Low Energy	Oliver Twist	6	00:51-01:07	5.1	3.1	4.9	4.4	5.3	4.3
146	Low Energy	Juha	16	00:00-00:15	5.4	3.4	5.6	5.8	5.9	5.0
19	Negative Valence	Road to Perdition	6	00:34-00:49	5.4	4.6	3.3	3.1	3.6	3.4
67	Negative Valence	Grizzly Man	16	01:05-01:32	5.6	5.4	2.9	2.7	3.6	3.8
83	Negative Valence	Lethal Weapon 3	7	00:00-00:16	4.6	4.1	4.3	3.9	4.9	3.6
99	Negative Valence	The English Patient	8	01:35-01:57	5.5	5.6	4.3	3.0	4.2	4.4
115	Negative Valence	Hellraiser	5	00:00-00:15	5.5	6.1	4.1	3.8	4.9	4.6
3	Negative Valence	Batman	9	00:57-01:16	3.3	3.1	4.2	4.3	4.4	4.1
35	Negative Valence	Shakespeare in Love	11	00:21-00:36	3.7	4.4	3.8	4.1	4.3	4.1
51	Negative Valence	The Fifth Element	9	00:00-00:18	4.5	4.3	4.1	3.7	4.6	4.1

131	Negative Valence	Big Fish	15	00:15-00:30	5.4	4.2	5.6	4.3	5.2	4.1
147	Negative Valence	Juha	18	02:30-02:46	6.1	5.3	6.2	5.9	5.8	4.4
4	Positive Valence	Gladiator	17	00:14-00:27	4.2	4.3	6.2	6.5	5.6	4.7
68	Positive Valence	Juha	10	00:20-00:38	4.1	4.1	7.3	7.9	6.8	5.6
84	Positive Valence	Dances with Wolves	10	00:28-00:46	5.6	5.0	7.7	6.5	6.2	5.1
132	Positive Valence	Blanc	12	00:51-01:06	4.4	3.9	6.7	6.3	6.4	5.0
148	Positive Valence	Pride & Prejudice	9	00:01-00:21	4.4	2.9	6.1	5.8	5.9	4.4
20	Positive Valence	Vertigo OST	6	04:42-04:57	6.2	6.1	6.3	5.9	5.9	6.1
36	Positive Valence	Vertigo OST	6	02:02-02:17	3.8	4.4	4.7	5.2	5.4	4.7
52	Positive Valence	Man of Galilee CD1	2	00:19-00:42	4.7	6.2	6.7	7.2	6.1	5.7
100	Positive Valence	Shakespeare in Love	21	00:03-00:21	7.4	5.4	7.6	6.2	6.2	5.1
116	Positive Valence	Outbreak	6	00:16-00:31	5.6	4.5	6.6	5.5	5.7	4.8

Appendix B

Table B.3 Factor Loadings for each feature

Feature	MR 2	MR 1	MR 4	MR1 4	MR1 1	Factor
barkbands crest (dmean)		0.68				MR1
barkbands crest (dmean2)		0.67				MR1
barkbands crest (dvar)		0.70				MR1
barkbands crest (dvar2)		0.77				MR1
barkbands spread (dmean)		0.64				MR1
barkbands spread (dmean2)		0.67				MR1
hfc (dmean)		0.66				MR1
hfc (dmean2)		0.70				MR1
melbands crest (dmean)		0.60				MR1
melbands crest (dmean2)		0.58				MR1
melbands crest (dvar)		0.63				MR1
melbands crest (dvar2)		0.66				MR1
melbands spread (dmean)		0.55				MR1
melbands spread (dmean2)		0.61				MR1
pitch salience (dmean)		0.59				MR1
pitch salience (dmean2)		0.61				MR1
pitch salience (dvar)		0.60				MR1
pitch salience (dvar2)		0.67				MR1
spectral centroid (dmean)		0.58				MR1
spectral centroid (dmean2)		0.63				MR1
spectral complexity (dmean)		0.74				MR1
spectral complexity (dmean2)		0.76				MR1
spectral complexity (dvar)		0.84				MR1
spectral complexity (dvar2)		0.86				MR1
spectral complexity (var)		0.53				MR1
spectral energyband high (dmean)		0.61				MR1
spectral energyband high (dmean2)		0.65				MR1
spectral energyband high (median)		0.58				MR1
spectral rolloff (dmean)		0.77				MR1
spectral rolloff (dmean2)		0.78				MR1
spectral rolloff (dvar)		0.57				MR1

spectral rolloff (dvar2)		0.67				MR1
zerocrossingrate (dmean)		0.51				MR1
zerocrossingrate (dmean2)		0.58				MR1
spectral contrast coeffs 6 (dmean)					0.50	MR11
spectral contrast coeffs 5 (dvar)					0.56	MR11
spectral contrast coeffs 6 (dvar)					0.63	MR11
spectral contrast coeffs 5 (dvar2)					0.57	MR11
spectral contrast coeffs 6 (dvar2)					0.63	MR11
spectral contrast coeffs 6 (var)					0.59	MR11
dynamic complexity				-0.69		MR14
hfc (median)				0.67		MR14
silence rate 60dB (dmean)				-0.77		MR14
silence rate 60dB (dmean2)				-0.77		MR14
silence rate 60dB (dvar)				-0.78		MR14
silence rate 60dB (dvar2)				-0.78		MR14
silence rate 60dB (mean)				-0.68		MR14
silence rate 60dB (var)				-0.74		MR14
spectral decrease (mean)				-0.58		MR14
spectral decrease (median)				-0.75		MR14
spectral energy (mean)				0.59		MR14
spectral energy (median)				0.75		MR14
spectral energyband low (median)				0.57		MR14
spectral energyband middle low (mean)				0.54		MR14
spectral energyband middle low (median)				0.68		MR14
spectral flux (median)				0.53		MR14
spectral rms (mean)				0.74		MR14
spectral rms (median)				0.81		MR14
barkbands flatness db (mean)	0.81					MR2
barkbands flatness db (median)	0.80					MR2
dissonance (mean)	-0.50					MR2
erbbands flatness db (mean)	0.77					MR2
erbbands flatness db (median)	0.75					MR2
melbands flatness db (mean)	0.80					MR2
melbands flatness db (median)	0.80					MR2
spectral entropy (mean)	-0.66					MR2
spectral entropy (median)	-0.65					MR2
spectral kurtosis (dmean)	0.78					MR2
spectral kurtosis (dmean2)	0.77					MR2

spectral kurtosis (mean)	0.92					MR2
spectral kurtosis (median)	0.92					MR2
spectral skewness (dmean)	0.78					MR2
spectral skewness (dmean2)	0.79					MR2
spectral skewness (dvar)	0.57					MR2
spectral skewness (dvar2)	0.56					MR2
spectral skewness (mean)	0.90					MR2
spectral skewness (median)	0.90					MR2
spectral skewness (var)	0.67					MR2
spectral spread (mean)	-0.61					MR2
spectral spread (median)	-0.61					MR2
mfcc 1	-0.60					MR2
mfcc 2	0.54					MR2
mfcc 3	0.54					MR2
spectral contrast coeffs 6 (mean)	-0.61					MR2
spectral contrast coeffs 6 (median)	-0.61					MR2
spectral contrast valleys 4 (mean)	-0.56					MR2
spectral contrast valleys 5 (mean)	-0.73					MR2
spectral contrast valleys 6 (mean)	-0.67					MR2
spectral contrast valleys 4 (median)	-0.56					MR2
spectral contrast valleys 5 (median)	-0.72					MR2
spectral contrast valleys 6 (median)	-0.65					MR2
spectral contrast coeffs 1 (dmean)			0.68			MR4
spectral contrast coeffs 2 (dmean)			0.67			MR4
spectral contrast coeffs 3 (dmean)			0.78			MR4
spectral contrast coeffs 4 (dmean)			0.77			MR4
spectral contrast coeffs 1 (dmean2)			0.64			MR4
spectral contrast coeffs 2 (dmean2)			0.62			MR4
spectral contrast coeffs 3 (dmean2)			0.75			MR4
spectral contrast coeffs 4 (dmean2)			0.74			MR4
spectral contrast coeffs 1 (dvar)			0.65			MR4
spectral contrast coeffs 2 (dvar)			0.68			MR4
spectral contrast coeffs 3 (dvar)			0.77			MR4
spectral contrast coeffs 4 (dvar)			0.78			MR4
spectral contrast coeffs 1 (dvar2)			0.61			MR4
spectral contrast coeffs 2 (dvar2)			0.64			MR4
spectral contrast coeffs 3 (dvar2)			0.73			MR4
spectral contrast coeffs 4 (dvar2)			0.74			MR4
spectral contrast valleys 1 (dmean)			0.59			MR4

spectral contrast valleys 2 (dmean)			0.63			MR4
spectral contrast valleys 1 (dmean2)			0.58			MR4
spectral contrast valleys 2 (dmean2)			0.63			MR4

Appendix C

Table C.4 Rhythmic features and the top 10 music excerpts that activate that feature

Arousal_Feature_1 (Pumping Rhythm)	BZ_11-00-00-00-00.mp3
	BZ_00-00-06-28-74.mp3
	CM_TechnoWorm_01.mp3
	BZ_00-03-00-48-37.mp3
	BZ_11-16-00-00-00.mp3
	BZ_09-26-00-00-37.mp3
	BZ_11-00-00-00-58.mp3
	CM_Chopper_01.mp3
	CM_Chopper_02.mp3
	BZ_13-00-00-43-00.mp3
Arousal_Feature_10 (Driving Rhythm)	BZ_00-00-00-00-56.mp3
	BZ_11-16-00-00-00.mp3
	BZ_31-23-00-00-70.mp3
	FR_Em_2_1_1.mp3
	BZ_11-00-00-00-58.mp3
	BZ_18-00-00-49-23.mp3
	BZ_11-00-00-00-00.mp3
	BZ_11-15-00-00-04.mp3
	FR_Em_3_1_1.mp3
	BZ_10-00-00-01-00.mp3
Valence_Feature_15 (Firm Rhythm)	CM_MetalTron_01.mp3
	BZ_00-00-08-00-37.mp3
	EL_VOL3_02_143_Dm_Part1_1_0_1_0_1_1_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part3_0_0_0_1_0_0_0_0_1_0_.mp3
	EL_VOL2_02_141_D5_to_Ab_Part1_1_1_0_0_1_1_1_0_1_.mp3
	EL_VOL2_02_141_D5_to_Ab_Part1_0_1_0_1_1_1_1_0_1_.mp3
	EL_VOL2_05_75_Dm_to_Ebm_Part4_0_0_0_1_0_0_1_0_1_1_1_0_.mp3
	FR_Dm_0_4.mp3
	FR_D#m_0_0_3_0.mp3

	FR_Am_0_1_0_0.mp3

Table C.5 Rhythmic sparsity features and the top 10 excerpts that activate that feature

Arousal_Feature_9 (Rhythmic sparsity)	EL_VOL1_04_158_Cm_Part1_1_1_1_0_1_0_1_0_1_0_.mp3
	EL_02_141_Var_MIX_pt1.mp3
	CM_FallenHero_02.mp3
	DYS_118_G_Haze_1_0_1_1_1_.mp3
	CM_MinorMelody_01.mp3
	CM_FallenHero_01.mp3
	CM_HappyIrishWhistle_01.mp3
	BZ_00-00-00-00-02.mp3
	BZ_00-00-00-00-04.mp3
	BZ_00-00-00-00-28.mp3
Valence_Feature_6 (Rhythmic sparsity)	FR_Dm_0_4.mp3
	FR_Am_0_1_0_0.mp3
	EL_02_141_Var_MIX_pt1.mp3
	EL_VOL2_02_141_D5_to_Ab_Part1_1_0_1_1_1_1_1_0_1_.mp3
	EL_05_75_Cm_MIX_pt1.mp3
	EL_02_141_Ab:G#m_MIX_pt2.mp3
	EL_04_158_Cm_MIX_pt2.mp3
	EL_VOL2_02_141_D5_to_Ab_Part2_0_0_0_1_0_1_0_1_0_0_1_.mp3
	EL_03_91_Ebm_MIX_pt2.mp3
	EL_03_110_D_MIX_pt1.mp3

Table C.6 Musical pattern features and top 10 excerpts that activate that feature

Arousal_Feature_3 (Harp and Pizzicato like Ostinatos)	EL_VOL1_05_78_A_to_Bb_Part1_0_1_0_0_0_.mp3
	EL_VOL1_05_78_A_to_Bb_Part1_0_1_1_1_0_.mp3
	EL_05_78_A_MIX_pt1.mp3
	EL_VOL1_01_70_Dm_to_Em_Part2_0_1_1_.mp3
	EL_VOL1_02_100_G7_to_Gm_Part1_0_1_1_1_0_1_.mp3
	MFM_100_F#_01_0_0_0_1_1_1_1_.mp3

	EL_VOL2_04_98_Am_Part2_0_1_0_0_1_0_0_.mp3
	EL_VOL3_01_115_F#_to_Eb_Part1_0_1_0_0_0_0_0_1_.mp3
	EL_VOL2_03_110_D_Part2_1_0_0_1_1_1_.mp3
	EL_VOL1_02_100_G7_to_Gm_Part1_1_1_1_1_0_0_.mp3
Arousal_Feature_4 (Static, Repetitive, Long Decay)	MFM_100_F#_01_0_0_0_1_1_1_1_.mp3
	MFM_130_C_01_1_0_0_1_0_1_1_.mp3
	MFM_125_Fm_01_1_0_1_1_0_1_0_.mp3
	MFM_100_F#_01_1_0_0_1_1_0_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part1_0_1_1_1_0_.mp3
	EL_VOL1_05_78_A_to_Bb_Part1_0_1_0_0_0_.mp3
	MFM_120_G#_01_0_1_0_1_0_1_.mp3
	EL_05_78_A_MIX_pt1.mp3
	MFM_130_C_01_0_0_0_1_1_0_0_.mp3
	DYS_100_Am_Blissfull_1_0_1_0_0_.mp3
Valence_Feature_4 (minor ostinatos)	EL_04_158_Cm_MIX_pt2.mp3
	EL_VOL2_02_141_D5_to_Ab_Part2_0_0_0_1_0_1_0_1_0_0_1_.mp3
	EL_VOL3_05_75_Cm_to_Gm_Part2_0_0_0_1_1_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part3_1_0_0_0_0_1_0_1_1_0_.mp3
	EL_04_98_Am_MIX_pt1.mp3
	EL_03_110_D_MIX_pt3.mp3
	EL_VOL1_02_100_G7_to_Gm_Part2_1_0_0_1_1_1_0_.mp3
	EL_VOL3_01_115_F#_to_Eb_Part1_0_1_1_0_1_1_1_1_.mp3
	EL_05_78_Bb_MIX_pt3.mp3
	EL_04_98_Am_MIX_pt2.mp3
Valence_Feature_10 (stepwise and arpeggio type patterns)	EL_05_75_Cm_MIX_pt1.mp3
	EL_02_141_Ab:G#m_MIX_pt2.mp3
	EL_04_158_Cm_MIX_pt2.mp3
	EL_VOL2_02_141_D5_to_Ab_Part2_0_0_0_1_0_1_0_1_0_0_1_.mp3

	EL_03_91_Ebm_MIX_pt2.mp3
	EL_03_110_D_MIX_pt1.mp3
	EL_VOL3_05_75_Cm_to_Gm_Part2_0_0_0_1_1_1_.mp3
	EL_VOL2_03_110_D_Part2_1_0_0_0_1_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part3_1_0_0_0_0_1_0_1_1_0_.mp3
	EL_04_98_Am_MIX_pt1.mp3
Valence_Feature_14 (rapid harp-like arpeggios and musical trills)	CM_Euphoria_01.mp3
	MFM_120_G#_02_0_0_1_0_0_1_.mp3
	CC_Intro_18_Full.mp3
	EL_02_141_Var_MIX_pt4.mp3
	FR_Dm_0_4.mp3
	FR_Am_0_1_0_0.mp3
	EL_02_141_Var_MIX_pt1.mp3
	MFM_120_Am_01_1_0_0_1_1_1_.mp3
	EL_VOL2_03_110_D_Part3_1_0_0_0_0_0_0_.mp3
	DYS_90_C#_Forget_0_1_0_0_1_.mp3

Table C.7 Drone features and the top 10 excerpts that activate them

Arousal_Feature_7 (drones)	DYS_90_D_Time 98_1_1_1_1_0_.mp3
	DYS_100_D#_Dark Fog_1_1_0_1_1_1_.mp3
	FR_Gm_4_0_0_1.mp3
	FR_Gm_0_0_1_0.mp3
	CM_Purgatory_01.mp3
	MFM_130_C_02_0_1_1_1_1_0_.mp3
	DYS_132_F_Full_SP_01.mp3
	MFM_130_C_Full_SP.mp3
	MFM_120_Dm_01_1_1_1_0_1_0_0_.mp3
	EL_VOL2_03_110_D_Part1_0_0_1_.mp3
Arousal_Feature_8 (long release - drone)	DYS_100_D#_Dark Fog_1_1_0_1_1_1_.mp3
	DYS_90_D_Time 98_1_1_1_1_0_.mp3
	FR_Gm_0_1_1_1.mp3
	CT_01_0_0_0_0_1_.mp3
	FR_Gm_0_0_1_0.mp3
	FR_D#m_1_0_1_0.mp3
	FR_Gm_4_0_0_1.mp3

	MFM_120_C_01_1_1_0_0_0_0_0_.mp3
	EL_VOL3_02_143_Dm_Part1_1_0_1_0_1_1_1_.mp3
	EL_VOL3_02_143_Dm_Part1_0_0_0_0_1_0_1_.mp3
Arousal_Feature_14 (buzzy - squarewaves)	DYS_100_D#_Dark Fog_1_1_0_1_1_1_.mp3
	DYS_90_C#_Forget_1_0_1_0_1_.mp3
	DYS_90_D_Time 98_1_1_1_1_0_.mp3
	FR_Cm_0_0_2.mp3
	DYS_70_E_Omnis_0_1_0_0_1_0_.mp3
	FR_Cm_0_3_5.mp3
	FR_G#m_2_0_0_0.mp3
	FR_Dm_0_2.mp3
	FR_Cm_0_2_0.mp3
	FR_D#m_0_1_1_11.mp3
Valence_Feature_1 (drones)	BZ_00-00-07-00-00.mp3
	CT_07_0_1_0_0_.mp3
	BZ_00-00-06-00-00.mp3
	FR_Cm_3_0_0.mp3
	FR_Gm_4_4_3_1.mp3
	FR_D#m_0_1_0_0.mp3
	FR_Fm_0_1_0_0.mp3
	FR_Bm_0_3_0.mp3
	FR_D#m_0_0_4_0.mp3
	FR_Fm_2_1_0_0.mp3
Valence_Feature_5 (low range - buzzy)	FR_Fm_2_0_2_0.mp3
	FR_Fm_1_1_0_0.mp3
	FR_Fm_2_1_1_1.mp3
	FR_Fm_2_1_2_0.mp3
	FR_Am_1_0_1_0.mp3
	FR_Fm_2_1_0_0.mp3
	FR_D#m_0_0_3_0.mp3
	FR_Fm_2_1_2_1.mp3
	FR_Fm_0_1_1_0.mp3
	FR_Gm_4_6_2_6.mp3
Valence_Feature_13 (low rhythmic drones)	CC_Intro_20_Full.mp3
	BZ_00-00-08-00-37.mp3
	FR_D#m_0_0_3_0.mp3
	FR_Am_1_0_1_0.mp3
	FR_G#m_0_1_0_0.mp3
	FR_Am_0_0_1_0.mp3

	BZ_00-00-04-00-00.mp3
	FR_Cm_1_0_0.mp3
	DYS_100_Am_Blissfull_1_0_1_0_0_.mp3
	FR_Cm_7_4_3.mp3

Table C.8 Instrument features and the top 10 excerpts that activate them

Arousal_Feature_5 (Taiko and Bass Drums)	EL_04_140_Gm_MIX_pt1.mp3
	EL_04_140_Gm_MIX_pt2.mp3
	EL_03_113_Am_MIX_pt3.mp3
	EL_02_141_G#m_MIX_pt3.mp3
	EL_VOL2_02_141_D5_to_Ab_Part5_0_1_0_0_0_1_0_0_.mp3
	EL_03_110_D_MIX_pt3.mp3
	EL_VOL2_02_141_D5_to_Ab_Part5_1_1_0_1_1_1_0_0_.mp3
	EL_VOL1_03_91_Ebm_Part2_0_1_0_0_1_1_1_1_0_.mp3
	EL_04_98_Am_MIX_pt2.mp3
	EL_02_141_Var_MIX_pt5.mp3
Arousal_Feature_6 (mid to high range sustained sounds)	EL_VOL1_04_158_Cm_Part1_1_1_1_0_1_0_1_0_1_0_.mp3
	FR_Gm_0_1_1_1.mp3
	DYS_100_D#_Dark Fog_1_1_0_1_1_1_.mp3
	DYS_90_D_Time 98_1_1_1_1_0_.mp3
	EL_02_141_Var_MIX_pt1.mp3
	FR_F#m_0_2_0_0.mp3
	DYS_70_F_Dub Commissar_1_0_0_.mp3
	FR_Gm_0_0_1_0.mp3
	FR_Dm_0_4.mp3
	CM_NobleSacrifice_01.mp3
Arousal_Feature_11 (Brass)	EL_VOL1_04_158_Cm_Part1_1_1_1_0_1_0_1_0_1_0_.mp3
	CM_EvilRising_01.mp3
	CM_FormallySad_01.mp3
	CM_March_01.mp3
	CM_BrassSteps_01.mp3
	FR_Gm_0_1_1_1.mp3
	DYS_90_C#_Full_SP_01.mp3
	CM_AdventureTime_01.mp3

	CM_LighterMoment_01.mp3
	CM_HopeRises_01.mp3
Arousal_Feature_12 (Strings)	EL_VOL2_04_98_Am_Part3_0_0_1_0_1_1_0_0_1_1_1_.mp3
	EL_VOL2_04_98_Am_Part3_0_1_0_1_1_1_0_0_0_0_1_.mp3
	EL_05_78_A_MIX_pt1.mp3
	EL_02_100_G7_MIX_pt1.mp3
	EL_VOL1_05_78_A_to_Bb_Part1_0_1_1_1_0_.mp3
	EL_VOL2_04_98_Am_Part1_0_0_0_0_1_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part1_0_1_0_0_0_.mp3
	EL_05_78_A_MIX_pt2.mp3
	EL_VOL1_05_78_A_to_Bb_Part2_0_1_0_0_1_1_1_.mp3
	EL_VOL2_04_98_Am_Part1_1_1_1_0_0_1_.mp3
Arousal_Feature_13 (Light Piano and Pizzicato)	MFM_100_F#_01_0_0_0_1_1_1_1_.mp3
	MFM_130_C_01_0_0_0_1_1_0_0_.mp3
	MFM_130_C_01_1_0_0_1_0_1_1_.mp3
	MFM_120_G#_01_0_1_0_0_1_0_.mp3
	MFM_120_G#_01_0_1_0_1_0_1_.mp3
	MFM_100_Gm_01_0_1_1_0_1_1_0_.mp3
	MFM_100_F#_01_1_1_0_1_1_0_0_.mp3
	EL_02_100_G7_MIX_pt1.mp3
	MFM_120_C_01_1_1_0_0_1_1_0_.mp3
	MFM_120_G#_01_0_0_0_0_1_0_.mp3
Arousal_Feature_15 (Reverse Sweep Snare)	FR_Gm_0_1_1_1.mp3
	EL_04_140_Gm_MIX_pt2.mp3
	EL_01_79_Dm_MIX_pt3.mp3
	EL_VOL2_02_141_D5_to_Ab_Part5_1_1_0_1_1_1_0_0_.mp3
	FR_Am_0_1_0_0.mp3
	FR_Gm_0_9_3_0.mp3
	FR_F#m_0_2_0_0.mp3
	FR_Cm_9_0_2.mp3
	FR_F#m_5_3_0_0.mp3
	EL_03_110_D_MIX_pt3.mp3
Valence_Feature_2 (sawtooth)	CM_MetalTron_01.mp3
	FR_Am_0_0_1_1.mp3

	EL_VOL2_03_110_D_Part2_1_0_0_0_1_1_.mp3
	EL_05_78_Bb_MIX_pt3.mp3
	EL_VOL2_04_98_Am_Part3_0_0_1_0_1_1_0_0_1_1_1_.mp3
	EL_VOL2_04_98_Am_Part3_0_1_0_1_1_1_0_0_0_0_1_.mp3
	EL_VOL1_05_78_A_to_Bb_Part3_0_0_0_1_0_0_0_0_1_0_.mp3
	EL_05_75_Dm_Ebm_MIX_pt4.mp3
	EL_VOL2_01_79_Dm_to_Gm_Part3_0_0_1_0_0_1_1_1_0_1_1_.mp3
	EL_03_91_Ebm_MIX_pt1.mp3
Valence_Feature_9 (noisy)	FR_Dm_0_4.mp3
	FR_D#m_0_0_3_0.mp3
	FR_Am_0_1_0_0.mp3
	FR_Am_1_0_1_0.mp3
	FR_G#m_0_1_0_0.mp3
	FR_Am_0_0_1_0.mp3
	BZ_00-00-04-00-00.mp3
	CT_Action5-Cue_G_125.mp3
	CC_Intro_14_Full.mp3
	MFM_130_A_01_0_0_0_1_0_0_.mp3
Valence_Feature_12 (bitcrush)	BZ_00-00-08-00-37.mp3
	FR_Am_0_0_1_1.mp3
	CM_MetalTron_01.mp3
	BZ_00-00-03-00-43.mp3
	FR_F#m_5_3_1_1.mp3
	BZ_11-32-03-03-40.mp3
	BZ_41-00-00-41-07.mp3
	FR_Dm_0_4.mp3
	BZ_22-30-04-03-65.mp3
	BZ_00-31-00-00-11.mp3

Table C.9 Modal features and the top 10 excerpts that activate them

Arousal_Feature_2 (low-tones and minor key)	FR_Gm_0_1_1_1.mp3
	DYS_100_Am_Blissfull_1_0_1_0_0_.mp3
	MFM_130_A_01_1_1_0_1_1_1_.mp3
	DYS_90_C#_Full_SP_01.mp3
	MFM_100_D#_01_1_1_0_1_1_0_.mp3

	MFM_115_Em_Full_SP.mp3
	MFM_100_Em_01_1_1_0_1_0_.mp3
	MFM_100_D#_01_1_1_0_0_0_1_.mp3
	MFM_125_Fm_01_1_0_0_0_1_0_1_.mp3
	MFM_110_Em_01_1_0_0_.mp3
Valence_Feature_3 (Upbeat Major Key)	CM_TooHappy_01.mp3
	MFM_100_D#_Full_SP.mp3
	EL_03_91_Ebm_MIX_pt2.mp3
	EL_VOL2_03_110_D_Part2_1_0_0_1_1_1_.mp3
	EL_03_110_D_MIX_pt1.mp3
	MFM_115_Em_Full_SP.mp3
	MFM_110_F_Full_SP.mp3
	MFM_115_Bm_Full_SP.mp3
	MFM_115_Am_Full_SP.mp3
	EL_VOL2_03_110_D_Part2_0_1_1_1_0_1_.mp3
Valence_Feature_7 (Upbeat Major Key)	MFM_110_F_Full_SP.mp3
	CM_TooHappy_01.mp3
	MFM_115_C_Full_SP.mp3
	MFM_115_D_Full_SP.mp3
	MFM_100_F#_Full_SP.mp3
	MFM_125_D#_01_0_1_1_1_0_1_1_.mp3
	MFM_100_D#_Full_SP.mp3
	MFM_120_Dm_Full_SP.mp3
	MFM_115_Bm_Full_SP.mp3
	CM_Polite_01.mp3
Valence_Feature_8 (slow minor)	EL_VOL3_04_140_Gm_to_G#m_Part3_0_0_1_1_.mp3
	EL_VOL3_01_115_F#_to_Eb_Part1_0_1_0_0_0_0_0_1_.mp3
	EL_VOL3_01_115_F#_to_Eb_Part2_1_1_0_0_0_1_0_0_1_1_1_1_1_0_0_.mp3
	EL_VOL1_03_91_Ebm_Part1_0_0_1_1_1_1_1_.mp3
	EL_VOL1_03_91_Ebm_Part1_1_1_1_1_0_0_1_.mp3
	EL_VOL2_03_110_D_Part1_0_0_1_.mp3
	EL_VOL1_01_70_Dm_to_Em_Part1_0_0_0_1_.mp3

	EL_VOL3_01_115_F#_to_Eb_Part2_0_1_1_1_0_0_1_0_0_0_1_0_0_0_0_.mp3
	EL_VOL2_02_141_D5_to_Ab_Part1_0_1_0_1_1_1_1_0_1_.mp3
	EL_02_141_Var_MIX_pt1.mp3
Valence_Feature_11 **(slow chord progression that feature vocal or brass choirs.)**	CM_Prayer_01.mp3
	CM_FanfareTheme_01.mp3
	CM_TenseChoir_01.mp3
	CM_MinorMelody_01.mp3
	CM_Haunted_01.mp3
	DYS_128_F#_Full_SP_01.mp3
	MFM_120_Am_01_1_0_0_0_0_1_.mp3
	CM_MemoryDance_02.mp3
	MFM_120_C_01_1_1_0_0_0_0_0_.mp3
	EL_VOL3_04_140_Gm_to_G#m_Part3_1_1_0_1_.mp3

References

Bai, J., Peng, J., Shi, J., Tang, D., Wu, Y., Li, J., & Luo, K. (2016). Dimensional music emotion recognition by valence-arousal regression. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2016 IEEE 15th International Conference on* (pp. 42–49). IEEE.

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychophysiology*, *60*(1), 34–43.

Belcher, J. D., & Haridakis, P. (2013). The Role of Background Characteristics, Music-Listening Motives, and Music Selection on Music Discussion. *Communication Quarterly*, *61*(4), 375–396.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546–2554).

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference* (pp. 13–20). Citeseer.

Blood, A. J., Zatorre, R. J., Bermudez, P., & Evans, A. C. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature Neuroscience*, *2*(4), 382.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... others. (2013). Essentia: An Audio Analysis Library for Music Information Retrieval. In *ISMIR*

(pp. 493–498). Retrieved from http://www.academia.edu/download/45760627/ESSENTIA_an_Audio_Analysis_Library_for_M20160518-6376-1ecqtul.pdf

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems* (pp. 737–744). Retrieved from http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf

Bruner, G. C. (1990). Music, mood, and marketing. *The Journal of Marketing*, 94–104.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(Jul), 2079–2107.

Chen, H.-C., & Chen, A. L. (2001). A music recommendation system based on music data grouping and user interests. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 231–238). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=502625

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *ArXiv Preprint ArXiv:1409.1259*. Retrieved from http://arxiv.org/abs/1409.1259

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *ArXiv Preprint ArXiv:1606.00298*.

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 2392–2396). IEEE.

Chollet, F. (2015). *Keras*.

Christopher, K. R., Kapur, A., Carnegie, D. A., & Grimshaw, G. M. (2014). A History of Emerging Paradigms in EEG for Music. In *ICMC*.

Chung, J., & Vercoe, G. S. (2006). The affective remixer: Personalized music arranging. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 393–398). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1125535

Collins, K. (2008). *Game sound: an introduction to the history, theory, and practice of video game music and sound design*. Mit Press.

Collins, K. (2017). *From Pac-Man to pop music: interactive audio in games and new media*. Routledge.

Collins, K., Kapralos, B., & Tessler, H. (2014). *The oxford handbook of interactive audio*. Oxford University Press.

Cook, N., & Dibben, N. (2010). Emotion in Culture and History. *Handbook of Music and Emotion: Theory, Research, Applications*, 45–72.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671.

Dakhel, G. M., & Mahdavi, M. (2011). A new collaborative filtering algorithm using K-means clustering and neighbors' voting. In *Hybrid Intelligent Systems (HIS), 2011 11th International Conference on* (pp. 179–184). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6122101

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., ... others. (2010). The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 293–296). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1864770

Dörfler, M., Bammer, R., & Grill, T. (2017). Inside the spectrogram: Convolutional Neural Networks in audio processing. In *Sampling Theory and Applications (SampTA), 2017 International Conference on* (pp. 152–155). IEEE.

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726864/

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *ISMIR* (pp. 621–626).

Eerola, T., & Vuoskoski, J. K. (2010). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*. Retrieved from http://pom.sagepub.com/content/early/2010/08/23/0305735610362821.abstract

Eerola, T., & Vuoskoski, J. K. (2013). A review of music and emotion studies: approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, *30*(3), 307–340.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553.

Enns, M. (2015). Game scoring: Towards a broader theory.

Evans, P., & Schubert, E. (2008). Relationships between expressed and felt emotions in music. *Musicae Scientiae*, *12*(1), 75–99.

Fernández, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, 513–582.

Friberg, A., & Sundströöm, A. (2002). Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception: An Interdisciplinary Journal*, *19*(3), 333–349.

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae*, *17*(3), 265–276.

Gabrielsson, A. (2002). Perceived emotion and felt emotion: same or different? *Musicae Scientiae*, *6*(1; SPI), 123–148.

Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grimshaw, G. M. (2017). Affective neuroscience: A primer with implications for forensic psychology. *Psychology, Crime & Law*, (just-accepted), 1–39.

Gundlach, R. H. (1935). Factors determining the characterization of musical phrases. *The American Journal of Psychology*, 624–643.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, *16*(9), 458–466.

Harmon-Jones, E., Harmon-Jones, C., & Summerell, E. (2017). On the Importance of Both Dimensional and Discrete Models of Emotion. *Behavioral Sciences*, *7*(4), 66.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Hevner, K. (1935). The affective character of the major and minor modes in music. *The American Journal of Psychology*, 103–118.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 246–268.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 1–17.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238.

Juslin, P. N., & Sloboda, J. (2011). *Handbook of music and emotion: Theory, research, applications*. OUP Oxford.

Juslin, P. N., & Sloboda, J. A. (2013). Music and emotion. In *The Psychology of Music (Third Edition)* (pp. 583–645). Elsevier.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(05), 559–575.

Kenealy, P. (1988). Validation of a music mood induction procedure: Some preliminary findings. *Cognition & Emotion*, *2*(1), 41–48.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010a). Music emotion recognition: A state of the art review (pp. 255–266). Citeseer.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010b). Music emotion recognition: A state of the art review. In *Proc. ISMIR* (pp. 255–266). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.7740&rep=rep1&type=pdf

Kleinginna Jr, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*(4), 345–379.

Koelsch, S., Fritz, T., v. Cramon, D. Y., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: an fMRI study. *Human Brain Mapping*, *27*(3), 239–250.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145). Montreal, Canada.

Kuo, F.-F., & Shan, M.-K. (2002). A personalized music filtering system based on melody style classification. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* (pp. 649–652). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1184020

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 688–693). IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lin, D., & Jayarathna, S. (2018). Automated Playlist Generation from Personal Music Libraries. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 217–224). IEEE.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80.

Lindsay, K. A., & Nordquist, P. R. (2007). More Than a Feeling: Some Technical Details of Swing Rhythm in Music. *Acoustics Today*, *3*(3), 31–42.

London, J. (2012). *Hearing in time: Psychological aspects of musical meter*. Oxford University Press.

Lu, L., Liu, D., & Zhang, H.-J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 5–18.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

McFee, B., Barrington, L., & Lanckriet, G. (2012). Learning content similarity for music recommendation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(8), 2207–2218.

Nierhaus, G. (2009). *Algorithmic composition: paradigms of automated music generation*. Springer Science & Business Media.

Ortony, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Panda, R., Rocha, B., & Paiva, R. P. (2015). Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, *29*(4), 313–334.

Pauws, S., & Eggen, B. (2002). PATS: Realization and user evaluation of an automatic playlist generator. In *ISMIR*. Retrieved from http://www.ismir2002.ismir.net/proceedings/02-FP07-4.pdf

Phillips, W. (2014). *A composer's guide to game music*. MIT Press.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, *1*(3), 3–33.

Rink, J. (2005). *The practice of performance: studies in musical interpretation*. Cambridge University Press.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, *5*(1), 35–53.

Rowe, R. (1992). *Interactive music systems: machine listening and composing*. MIT press. Retrieved from http://dl.acm.org/citation.cfm?id=530519

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Schedl, M., Knees, P., McFee, B., Bogdanov, D., & Kaminskas, M. (2015). Music recommender systems. In *Recommender Systems Handbook* (pp. 453–492). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4899-7637-6_13

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schubert, E. (2013). Emotion felt by the listener and expressed by the music: literature review and theoretical perspectives. *Frontiers in Psychology*, *4*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3865445/

Shan, M.-K., Kuo, F.-F., Chiang, M.-F., & Lee, S.-Y. (2009). Emotion-based music recommendation by affinity discovery from film music. *Expert Systems with Applications*, *36*(4), 7666–7674.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth." In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 210–217). ACM Press/Addison-Wesley Publishing Co. Retrieved from http://dl.acm.org/citation.cfm?id=223931

Shukla, S., & Banka, H. (2018). An Automatic Chord Progression Generator Based On Reinforcement Learning. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 55–59). IEEE.

Song, Y., Dixon, S., & Pearce, M. (2012). A survey of music recommendation systems and future perspectives. In *9th International Symposium on Computer Music Modeling and Retrieval*. Retrieved from https://www.researchgate.net/profile/Yading_Song/publication/277714802_A_Survey_of_Music_Recommendation_Systems_and_Future_Perspectives/links/5571726608aef8e8dc633517.pdf

Thammasan, N., Fukui, K., & Numao, M. (2017). Multimodal Fusion of EEG and Musical Features in Music-Emotion Recognition. In *AAAI* (pp. 4991–4992).

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford University Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tkalcic, M., Kosir, A., & Tasic, J. (2011). Affective recommender systems: the role of emotions in recommender systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems* (pp. 9–13). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.369.8712&rep=rep1&type=pdf

Ungar, L. H., & Foster, D. P. (1998). Clustering methods for collaborative filtering. In *AAAI workshop on recommendation systems* (Vol. 1, pp. 114–129). Retrieved from http://www.aaai.org/Papers/Workshops/1998/WS-98-08/WS98-08-029.pdf

Västfjäll, D. (2002). Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, *5*(1 suppl), 173–211.

Vercoe, G. S. (2006). *Moodtrack: practical methods for assembling emotion-driven music*. Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/39923

Vuoskoski, J. K., & Eerola, T. (2011a). Measuring music-induced emotion A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, *15*(2), 159–173.

Vuoskoski, J. K., & Eerola, T. (2011b). The role of mood and personality in the perception of emotions represented by music. *Cortex*, *47*(9), 1099–1106.

Vuoskoski, J. K., Thompson, W. F., McIlwain, D., & Eerola, T. (2012). Who enjoys listening to sad music and why? *Music Perception: An Interdisciplinary Journal*, *29*(3), 311–317.

Weinberg, G., Raman, A., & Mallikarjuna, T. (2009). Interactive jamming with Shimon: a social robotic musician. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 233–234). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1514152

Williams, D., Kirke, A., Miranda, E. R., Roesch, E., Daly, I., & Nasuto, S. (2014). Investigating affect in algorithmic composition systems. *Psychology of Music*, 0305735614543282.

Yang, Y.-H., & Chen, H. H. (2011). *Music emotion recognition*. CRC Press.

Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(3), 40.

Yoon, K., Lee, J., & Kim, M.-U. (2012). Music recommendation system using emotion triggering low-level features. *Consumer Electronics, IEEE Transactions On*, *58*(2), 612–618.

Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2006). Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences. In *ISMIR* (Vol. 6, p. 7th). Retrieved from http://130.54.20.150/members/yoshii/papers/ismir-2006-yoshii.pdf

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, *8*(4), 494.

Zhang, X., Hui, W. Y., & Barrett, L. F. (2014). How does this make you feel? A comparison of four affect induction procedures. *Frontiers in Psychology*, *5*.