USING MUSIC AND EMOTION TO ENABLE EFFECTIVE AFFECTIVE COMPUTING

BRENNON CHRISTOPHER BORTZ

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

R. Benjamin Knapp, Chair
D. Scott McCrickard
Steve R. Harrison
Eric D. Lyon
Isabel S. Bradburn
Javier E. Jaimovich

15 May 2019

Keywords: Affective Computing, Music, Psychophysiology, Databases
Using Music and Emotion to Enable Effective Affective Computing
Brennon Christopher Bortz

The computing devices with which we interact daily continue to become ever smaller, intelligent, and pervasive. Not only are they becoming more intelligent, but some are developing awareness of a user’s affective state. Affective computing—computing that in some way senses, expresses, or modifies affect—is still a field very much in its youth. While progress has been made, the field is still limited by the need for larger sets of diverse, naturalistic, and multimodal data.

This work first considers effective strategies for designing psychophysiological studies that permit the assembly of very large samples that cross numerous demographic boundaries, data collection in naturalistic environments, distributed study locations, rapid iterations on study designs, and the simultaneous investigation of multiple research questions. It then explores how commodity hardware and general-purpose software tools can be used to record, represent, store, and disseminate such data. As a realization of these strategies, this work presents a new database from the Emotion in Motion (EiM) study of human psychophysiological response to musical affective stimuli comprising over 23,000 participants and nearly 67,000 psychophysiological responses.

Because music presents an excellent tool for the investigation of human response to affective stimuli, this work uses this wealth of data to explore how to design more effective affective computing systems by characterizing the strongest responses to musical stimuli used in EiM. This work identifies and characterizes the strongest of these responses, with a focus on modeling the characteristics of listeners that make them more or less prone to demonstrating strong physiological responses to music stimuli.

This dissertation contributes the findings from a number of explorations of the relationships between strong reactions to music and the characteristics and self-reported affect of listeners. It demonstrates not only that such relationships do exist, but takes steps toward automatically predicting whether or not a listener will exhibit such exceptional responses. Second, this work contributes a flexible strategy and functional system for both successfully executing large-scale, distributed studies of psychophysiology and affect; and for synthesizing, managing, and disseminating the data.
collected through such efforts. Finally, and most importantly, this work presents the EiM database itself.
Using Music and Emotion to Enable Effective Affective Computing
Brennon Christopher Bortz

The computing devices with which we interact daily continue to become ever smaller, intelligent, and pervasive. Not only are they becoming more intelligent, but some are developing awareness of a user’s affective state. Affective computing—computing that in some way senses, expresses, or modifies affect—is still a field very much in its youth. While progress has been made, the field is still limited by the need for larger sets of diverse, naturalistic, and multimodal data.

This dissertation contributes the findings from a number of explorations of the relationships between strong reactions to music and the characteristics and self-reported affect of listeners. It demonstrates not only that such relationships do exist, but takes steps toward automatically predicting whether or not a listener will exhibit such exceptional responses. Second, this work contributes a flexible strategy and functional system for both successfully executing large-scale, distributed studies of psychophysiology and affect; and for synthesizing, managing, and disseminating the data collected through such efforts. Finally, and most importantly, this work presents the Emotion in Motion (EiM) (a study of human affective/psychophysiological response to musical stimuli) database comprising over 23,000 participants and nearly 67,000 psychophysiological responses.
For Whit, Bug, Nugget, and whoever is next...
I think music is the greatest art form that exists, and I think people listen to music for different reasons, and it serves different purposes. Some of it is background music, and some of it is things that might affect a person’s day, if not their life, or change an attitude. The best songs are the ones that make you feel something.

— Eddie Vedder

ACKNOWLEDGMENTS

First, I’m indebted to my advisor, Ben Knapp, with whom I’ve had the pleasure of working for more than a decade now. Ben, for your continual patience with me and encouragement throughout this work you have my sincere gratitude. There have been times, certainly, when I’ve been nothing short of exasperating. Thank you for spurring me on toward the goal. Your insight, knowledge, and creativity have been indispensable during this work.

These more than ten years working with Ben started on an entirely different continent. We began working together at the Sonic Arts Research Centre at Queen’s University Belfast—a place that will always remain very special to me. There, I was fortunate to work with and alongside a number of exceptional people, many of whom remain friends to this day. In particular, my coworkers in the Music, Sensors, and Emotion research group made my time in Northern Ireland a wonderful one. Miguel Ortiz Pérez, Niall Coghlan, Cavan Fyans, Adnan Marquez-Borbon, Michael Gurevich, Paul Stapleton, and Nick Gillian, you all are tremendous people. Michael Alcorn, I’m thankful for meeting you many years ago, which kicked off this wild adventure.

In particular, one of the finest people that I have been privileged to have the opportunity to know from my time there is Javier Jaimovich. Our work that began there has continued to today as a fruitful collaboration. Javier, your friendship, diligence, and creativity have been an inspiration for me. I’m very thankful to call you a friend and to have worked with you, and I look forward to continuing to do so.

This dissertation reaches for inspiration across a number of disciplinary boundaries. I’m grateful to my committee members who come from equally diverse disciplines: Ben Knapp, Javier Jaimovich, Isabel Bradburn, Steve Harrison, Scott McCrickard, and Eric Lyon. I appreciate the insights and wisdom that you have brought to this effort.

What began in Belfast, Northern Ireland ends here, in Blacksburg, Virginia, where I came to work as the “inaugural” doctoral student in the Institute for Creativity, Arts, and Technology at Virginia Tech. Resetting
progress on my Ph.D. after already having worked on it for several years made for some unnerving times, but the faculty, staff, and my fellow students at ICAT helped prove it was the right decision, in the end. Melissa, Holly, Deba, Tanner, Zach, Liz, and so many others, you are all fantastic.

To my mother, father, and brother, I know that many of my choices in life have given you reason for pause over the years. Thank you for your continual encouragement, and for the example of what an incredible family is. Ryan, your perseverance, especially, is a constant motivation for me.

My deepest gratitude is owed to my two girls, Aisling and Kella, and my wife, Whitney. I’ve missed too many important moments in your childhoods, girls. Don’t worry, Daddy’s “stupid dissertation” is done. Whitney, you’ve been a motivation and inspiration to me since all the way back at the swing in front of Darnell Hall. Thank you for your tireless encouragement, for being an incredible mother to our children, for your long-suffering patience, and in general, for putting up with my nonsense. I love you.
## Contents

4.1.5 Summary of Similar Systems 54

4.2 Issues with the Original EiM System 54
4.2.1 Issues with the Study Apparatus 55
4.2.2 Issues with Data Synthesis, Management, and Dissemination 57
4.2.3 Specific Issues with and Changes to the EiM Study Design 58

4.3 Requirements for a Flexible System 60
4.3.1 Requirements for the Data Collection System 60
4.3.2 Requirements for the Data Synthesis, Management, and Dissemination System 62

4.4 Realizing the Flexible System 65
4.4.1 Signal Capture and Processing 66
4.4.2 Database 67
4.4.3 Web Application 74

4.5 Wider Contributions of This Work 77
4.5.1 Open Database and Supporting Software Libraries 78

4.6 Summary 79

5 Strong Physiological Responses 80
5.1 Background 80
5.1.1 What are Strong Reactions to Music? 80

5.2 Strong Psychophysiological Responses in EiM 84
5.2.1 Hallelujah 84
5.2.2 Raining Blood 85
5.2.3 Into Dust 87

5.3 Modeling Strong Psychophysiological Reactions 88
5.3.1 Identifying Reactions Deterministically 88
5.3.2 Exploring Reaction Groups 89
5.3.3 Ordinal Logistic Regression 97
5.3.4 Principal Component Analysis 100
5.3.5 Neural Networks 103

5.4 Summary 113

6 Discussion 114
6.1 RQ1 Reflection 114
6.1.1 Documenting the Issues with EiM 114
6.1.2 Requirements for a Better Data Collection Strategy 116
6.1.3 Continued Improvements to the Strategy and System 116

6.2 RQ2 Reflection 120
6.2.1 Continued Improvements to the Strategy and System 122
6.2.2 “Isn’t this all just common sense?” 124
### CONTENTS

6.3 RQ3 Reflection 125  
6.3.1 Statistical Analysis 125  
6.3.2 Modeling Approaches 127  
6.4 Notes on Affect, Arousal, Music, and Future Work 129  
6.5 Final Remarks 132

**REFERENCES** 133  
A **APPENDIX A: EIM SCREEN SHOTS** 151  
B **APPENDIX B: EXAMPLE DATA FILES** 159  
B.1 trial Document 159  
B.2 signal Document 160  
B.3 Raw signal File Document Metadata 161  
B.4 Processed signal File Document Metadata 161  
B.5 media Document 162  
B.6 Study Design Document 162  
Associated Publications 164
### List of Figures

| Figure 2.1 | Russell’s circumplex model of emotion constructed from subject’s ratings of emotion words, and from subject’s self-reported emotion [144, p. 1167–1168]. Reprinted with permission. |
| Figure 2.2 | Watson and Tellegen’s two-factor structure of affect [175, p. 221]. Reprinted with permission. |
| Figure 2.3 | Hevner’s circumplex model of musical affect [62, p. 249]. (From American Journal of Psychology. Copyright ©1936 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.) |
| Figure 3.1 | EiM conceptual diagram. |
| Figure 3.2 | Typical EiM experimental apparatus configuration. |
| Figure 4.1 | EiM system block diagram. |
| Figure 5.1 | Mean electrodermal activity (EDA) signals from participants who listened to Jeff Buckley’s Hallelujah in Dublin, Taipei City, and Taichung City. |
| Figure 5.2 | Mean EDA signals from participants who listened to Slayer’s Raining Blood in Dublin, Taipei City, and Taichung City. |
| Figure 5.3 | Mean EDA signals from participants who listened to Mazzy Star’s Into Dust in Dublin, Taipei City, and Taichung City. |
| Figure 5.4 | Demographics for participants who heard Hallelujah in Dublin, Taiwan City, and Taichung City: (a) distributions of ages by reaction group; (b) counts of sexes by reaction group; (c) counts of nationalities; (d) counts of languages. |
| Figure 5.5 | Univariate distributions of participant characteristics by reaction group for participants who heard Hallelujah in Dublin, Taiwan City, and Taichung City: (a) age; (b) concentration; (c) musical expertise. |
| Figure 5.6 | Distributions of activity and familiarity by reaction group for participants who heard Hallelujah in Dublin: (a) Activity (felt arousal); (b) Familiarity. |
Figure 5.7 Example bivariate distributions involving artistic
between groups for participants who heard Hallelu-
jah in Taiwan. 98
Figure 5.8 Component loadings for first five components from
Dublin and Taiwan Hallelujah principal component
analysis (PCA). 102
Figure 5.9 Example pairwise plots of component values from
PCA-transformed Dublin and Taiwan Hallelujah data. 103
Figure 5.10 Feature weights after least absolute shrinkage and
selection operator (LASSO) regression on Dublin and
Taiwan Hallelujah data. 109
Figure 5.11 Changes in weighted $F_{0.5}$ scores in response to num-
ber of input features in the models of Dublin and
Taiwan Hallelujah data. 110
Figure 6.1 “Code Quality” Reprinted from https://xkcd.com/
1695/ with permission. 124
Figure A.1 Screen shot of welcome screen. 152
Figure A.2 Screen shot of implicit consent form. 153
Figure A.3 Screen shot of demographics questionnaire. 154
Figure A.4 Screen shot of abbreviated Big Five personality bat-
tery questionnaire. 155
Figure A.5 Screen shot of questionnaire following media ex-
cerpt playback. 156
Figure A.6 Screen shot of musical background questionnaire. 157
Figure A.7 Screen shot of final questionnaire. 158
# List of Tables

<table>
<thead>
<tr>
<th>Table 2.1</th>
<th>Definitions of key terms from Juslin and Sloboda’s <em>Handbook of Music and Emotion: Theory, Research, Applications</em> [91, p. 10].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2</td>
<td>Summary of survey from Rigg [141, p. 430–434]. Emphasized characteristics of musical features indicate common findings across surveyed studies.</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Lower-level musical features as they relate to basic emotions [51].</td>
</tr>
<tr>
<td>Table 3.1</td>
<td><em>EiM</em> iterations, locations, and start and end dates.</td>
</tr>
<tr>
<td>Table 3.2</td>
<td><em>EiM</em> sample sizes by location.</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Demographic variables.</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Measures of musical training and experience.</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Big Five Inventory-10 (10-item) (<em>BFI-10</em>) measures of personality traits.</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Self-reported measures of affective response.</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>Common data from all <em>EiM</em> locations.</td>
</tr>
<tr>
<td>Table 3.8</td>
<td>Musical stimuli used in <em>EiM</em> iterations.</td>
</tr>
<tr>
<td>Table 3.9</td>
<td>Media pool size and experimental stimuli count per iteration.</td>
</tr>
<tr>
<td>Table 3.10</td>
<td>Common excerpts from all <em>EiM</em> locations.</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Comparisons of affective/musical responses between reaction groups from participants who heard <em>Hallelujah</em> in Dublin.</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Comparisons of personality items between reaction groups for participants who heard <em>Hallelujah</em> in Taiwan (n = 188).</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Comparisons of personality traits between reaction groups for participants who heard <em>Hallelujah</em> in Taiwan (n = 188).</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Test results from comparisons of bivariate joint distributions between groups for participants who heard <em>Hallelujah</em> in Dublin and Taiwan (where Bonferroni-corrected p-values are &lt; .05).</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>Ordinal logistic regression goodness of fit for <em>Hallelujah</em> reaction in Taiwan.</td>
</tr>
<tr>
<td>Table 5.6</td>
<td>Ordinal logistic regression overall model fit for <em>Hallelujah</em> reaction in Taiwan.</td>
</tr>
</tbody>
</table>
Table 5.7  Ordinal logistic regression parameter estimates for *Hallelujah* reaction in Taiwan.  100
Table 5.8  Variance explained as a result of Dublin *Hallelujah* PCA.  102
Table 5.9  Available features for prototype neural network models of reaction vs. no-reaction groups described in Section 5.3.5.1.  104
Table 5.10 Confusion matrix for baseline classifier that predicts no-reaction for all examples for Dublin and Taiwan *Hallelujah* validation dataset.  105
Table 5.11 Confusion matrix for wide-and-deep classifier using all basic features for Dublin and Taiwan *Hallelujah* data.  107
Table 5.12 Confusion matrix for basic neural network classifier using all basic features for Dublin and Taiwan *Hallelujah* data.  108
Table 5.13 Results for classifiers with successively larger numbers of input features as determined by LASSO regression on Dublin and Taiwan *Hallelujah* data.  110
Table 5.14 Available features for neural network models of reaction vs. no-reaction groups discussed in Section 5.3.5.2.  111
Table 5.15 Tests of significance for improvement in refined custom model classification performance over baseline classifiers for *Hallelujah*, *Raining Blood*, and *Into Dust*.  113
Table 6.1  Summary of differences between *Hallelujah* reaction and no-reaction groups.  126
# LIST OF LISTINGS

<table>
<thead>
<tr>
<th>Listing</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Excerpt of example participant answers file from iteration 5.</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Excerpt of example participant answers file from iteration 9.</td>
<td>57</td>
</tr>
<tr>
<td>4.3</td>
<td>Example media document.</td>
<td>71</td>
</tr>
<tr>
<td>4.4</td>
<td>Example questionnaire view.</td>
<td>76</td>
</tr>
<tr>
<td>B.1</td>
<td>Example trial document.</td>
<td>159</td>
</tr>
<tr>
<td>B.2</td>
<td>Example signal document demonstrating document relationships.</td>
<td>160</td>
</tr>
<tr>
<td>B.3</td>
<td>Example raw signal file document metadata.</td>
<td>161</td>
</tr>
<tr>
<td>B.4</td>
<td>Example processed EDA signal document metadata.</td>
<td>161</td>
</tr>
<tr>
<td>B.5</td>
<td>Example media document.</td>
<td>162</td>
</tr>
<tr>
<td>B.6</td>
<td>Example study design document (excerpt).</td>
<td>162</td>
</tr>
</tbody>
</table>
**LIST OF ACRONYMS**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANS</td>
<td>autonomic nervous system</td>
</tr>
<tr>
<td>AUC-PR</td>
<td>area under the precision-recall curve</td>
</tr>
<tr>
<td>BFI-10</td>
<td>Big Five Inventory-10 (10-item)</td>
</tr>
<tr>
<td>BFI-44</td>
<td>Big Five Inventory-44 (44-item)</td>
</tr>
<tr>
<td>BSON</td>
<td>binary JSON</td>
</tr>
<tr>
<td>CNN</td>
<td>convolutional neural network</td>
</tr>
<tr>
<td>CNS</td>
<td>central nervous system</td>
</tr>
<tr>
<td>CSV</td>
<td>comma-separated values</td>
</tr>
<tr>
<td>ECG</td>
<td>electrocardiogram</td>
</tr>
<tr>
<td>EDA</td>
<td>electrodermal activity</td>
</tr>
<tr>
<td>EDR</td>
<td>electrodermal response</td>
</tr>
<tr>
<td>EEG</td>
<td>electroencephalogram</td>
</tr>
<tr>
<td>EiM</td>
<td>Emotion in Motion</td>
</tr>
<tr>
<td>EMG</td>
<td>electromyogram</td>
</tr>
<tr>
<td>EOG</td>
<td>electrooculogram</td>
</tr>
<tr>
<td>FACS</td>
<td>Facial Action Coding System</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional magnetic resonance imaging</td>
</tr>
<tr>
<td>FORCE</td>
<td>Future of Research Communications and e-Scholarship</td>
</tr>
<tr>
<td>GEMS</td>
<td>Geneva Emotional Music Scale</td>
</tr>
<tr>
<td>HR</td>
<td>heart rate</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>LASSO</td>
<td>least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>MVC</td>
<td>Model-View-Controller</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>MIR</td>
<td>music information retrieval</td>
</tr>
<tr>
<td>MuSE</td>
<td>Music, Sensors, and Emotion</td>
</tr>
<tr>
<td>NSF</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>OSC</td>
<td>Open Sound Control</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PET</td>
<td>positron emission tomography</td>
</tr>
<tr>
<td>PNS</td>
<td>peripheral nervous system</td>
</tr>
<tr>
<td>POX</td>
<td>pulse oximetry</td>
</tr>
<tr>
<td>RDBMS</td>
<td>relational database management system</td>
</tr>
<tr>
<td>RQ1</td>
<td>Research Question 1</td>
</tr>
<tr>
<td>RQ2</td>
<td>Research Question 2</td>
</tr>
<tr>
<td>RQ3</td>
<td>Research Question 3</td>
</tr>
<tr>
<td>SAM</td>
<td>Self-Assessment Manikin</td>
</tr>
<tr>
<td>SEM</td>
<td>Strong Experiences with Music</td>
</tr>
<tr>
<td>SNS</td>
<td>somatic nervous system</td>
</tr>
<tr>
<td>SPA</td>
<td>single-page application</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
“Rhythm and melody supply imitations of anger and gentleness, and also of courage and temperance, and of all the qualities contrary to these, and of the other qualities of character, which hardly fall short of the actual affections, as we know from our own experience, for in listening to such strains our souls undergo a change.” [4, p. 172]

Aristotle

1.1 THE POWER OF MUSIC

Since at least the time of the ancient Greek philosophers, the power of music to both express and manipulate emotion has been acknowledged. In The Republic, Plato describes the potential of music to even shape one’s character: “[M]usical training is a more potent instrument than any other, because rhythm and harmony find their way into the inward places of the soul, on which they mightily fasten, imparting grace, and making the soul of him who is rightly educated graceful, or of him who is ill-educated ungraceful...” [133, p. 127] Across the span of human history, music has been employed to stir the emotions of humans for good and ill. Less than eighty years ago, Joseph Goebbels—Adolf Hitler’s Reich Minister of Public “Enlightenment” and Propaganda—utilized music heavily in his propagandizing, stating, “Music affects the heart and emotions more than the intellect. Where then could the heart of a nation beat stronger than in the huge masses, in which the heart of a nation has found its true home?” [64, p. 5] The powerful positive effects of music, on the other hand, are acknowledged in the United Nations’ construction and release of the “world’s happiest playlist” and Secretary General Ban Ki-moon’s comments: “On this day we are using the universal language of music to show solidarity with the millions of people around the world suffering from poverty, human rights abuses, humanitarian crises and the effects of environmental degradation and climate change.” [26] Confucius described not only the power of music to produce pleasure or displeasure, and maintain order or disorder; but also its necessity:

1 A wonderful example of music’s incredible capacity to immediately evoke a positive affective response is that of Henry Dreher, an Alzheimer’s patient, chronicled in the 2014 documentary film, Alive Inside [143]. (https://www.youtube.com/watch?v=EgNLLelQYwI)
Now music produces pleasure—what the nature of man cannot be without. That pleasure must arise from the modulation of the sounds, and have its embodiment in the movements (of the body)—such is the rule of humanity. These modulations and movements are the changes required by the nature, and they are found complete in music. Thus men will not be without the ministration of pleasure, and pleasure will not be without its embodiment, but if that embodiment be not suitably conducted, it is impossible that disorder should not arise. [30, p. 127]

1.2 THE POTENTIAL OF AFFECTIVE COMPUTING

When humans interact with one another, the presence, awareness, and communication of emotion directly impact the quality of their interactions [163]. Moreover, the ability to recognize in and convey emotions to those with whom one interacts permits interaction and communication in a manner far more effective and intelligent than one devoid of emotion. Indeed, recent studies have clearly demonstrated that “the ability to recognize and express emotions [is] essential for natural communication” [132, p. 2]. For example, it has been shown that the neural circuitry central to emotion is key to perception, cognition, and attention [33], [72]. Moreover, and perhaps counter-intuitively, emotion is the “secret ingredient” that allows human rationality by which they can make choices drawn from a possibly infinite search space of choices, where computers often fail miserably [37].

Unquestionably, interaction with computing systems has become a central activity in our daily lives. Often, one communicates and interacts with computers more regularly, and just as socially, as any other agent in his or her life. To demonstrate this, Reeves and Nass staged a study where participants went through a tutoring session with a virtual agent on a computer. Following the session, the participant was asked to provide feedback about the performance of their virtual tutor either using pen and paper, using the same computer, or using a different computer. Surprisingly, participants provided significantly more positive feedback when they were doing so using the computer that had tutored them [136]. This suggests that when one interacts with a computer, she treats it as if it were emotionally sensitive to her interaction; on some level, people already behave as if computing systems can feel. Human-computer interaction is, at its heart, the interactive communication of humans with computers. If one expects (or at least hopes that) the humans with whom he interacts to be able to recognize, communicate, and reciprocate emotion in order to communicate efficiently and effectively, why would he not expect just the same from the computers with which he interacts? Certainly, computing systems equipped with the ability to recognize, communicate, and reciprocate—at
least synthetically—emotion would permit far more efficient, effective, and naturalistic interaction.

Many have dreamed of futuristic worlds in which pervasive, invisible computing systems respond to their every wish and whim. It would be nothing short of magical for one’s house to adjust the lights, background music, and prepare a favorite dinner automatically in response to its owner’s difficult day. Perhaps in such a future cinemas automatically manipulate a film’s visuals and score to create the most enjoyable experience for each unique audience. One’s computer automatically letting a significant other know that he or she is in need of encouragement, or that their day has gone swimmingly, might even hold the potential for saving a relationship or two. Closer to current day homes than the Jetsons’, though, affective computing—computing that relates to, arises from, or deliberately influences emotions [131, p. 1]—does hold the very real promise of improving interactions with computers, and people’s lives on the whole. For instance, initial work on tutoring systems that respond intelligently to the affective state of the student have shown impressive progress and results [34], [35].

1.3 THE RELATIONSHIP BETWEEN MUSIC AND AFFECT

The relationship between music and affect is a very active area of current research, and many debates around this relationship remain far from settled. However, recent research has highlighted a number of consistent themes.

That emotional intent can be perceived by (as opposed to elicited within) a listener has been accepted for millennia. Evidence from a number of studies suggest that not only can the perceived emotion in a selection of music be reliably determined from the characteristics of the features of the music itself [172], but also that judgments of these perceived emotions are reasonably consistent across musical cultures [7], and the ability to make such judgments develops early in childhood [1], [157]. In one of the earliest survey papers of results of experiments exploring listener judgments of the expressive content of selections of music, Hevner concluded that carefully designed experiments would clarify what seemed then to be a set of regular, systematic relationships between the high-level features of music and its expressive content [61]. Since then, while it has become clearer that there may be no hard and fast rules for just how to express particular emotions through certain musical configurations, it has become clear that listeners regularly agree well on “moodstates” that music intends to express [167], that one can do a reasonably good job of targeting certain emotions with precomposed melodies [170], and that listeners also agree well on the expressed emotion in improvised music [8]. In addition, these
findings seem to often hold across age groups [98], [166], [167], and cultures [7], [49].

On the other hand, is music capable of actually evoking emotion in a listener, or does music only express emotion that is then perceived by a listener? Numerous recent studies provide strong evidence for the claim that music does indeed evoke emotions in listeners. Examining both physiological responses as well as continuous self-reports of emotion, Krumhansl found that at least for musical emotions (emotions that generally correspond to basic emotions such as happiness, sadness, fear, and anger), these responses were distinguishable between emotions [101]. In the same year, Nyklícek et al. were able to also identify serenity and agitation elicited by music using only cardiorespiratory measures [122]. They were further able to differentiate significantly between serenity, happiness, sadness, and agitation. Later, Rickard used skin conductance measures to demonstrate that not only does music elicit genuine emotional responses that are no different than real, “in the wild” emotions, but that these musically induced emotional responses varied similarly to real emotions in that the intensities of these responses were correlated with the intensities of the stimuli [140]. Evidence now suggests that music can and does regularly evoke real emotional responses in listeners that are not unlike the emotions that one experiences in everyday life, as these and numerous other complementary studies suggest [13], [70], [76], [95], [109], [140]. Thus, music presents a useful tool not only for the volitional modulation of one’s affect, but for the study of human response to affective stimuli, in general [122].

1.4 PSYCHOPHYSIOLOGY

Psychophysiology (defined by Andreassi as “the study of relations between psychological manipulations and resulting physiological responses, measured in the living organism, to promote understanding of the relation between mental and bodily processes” [3, p. 2]) is a frequently used approach for exploring responses to musical stimuli, as well as affective stimuli of other sorts.

At the highest level, the human nervous system comprises two parts: the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS (comprising the brain and spinal cord) and the PNS (all nervous system components apart from the brain and spinal cord) work together through two-way communication. The PNS is further subdivided into the autonomic nervous system (ANS) and somatic nervous system (SNS). The ANS controls glands and smooth (involuntary) muscles, while the SNS controls skeletal (voluntary) muscles. Finally, the ANS is again subdivided into sympathetic and parasympathetic branches: the former preparing the body to act, and the latter promoting restoration. Psychophysiology is
concerned with studying psychological processes by examining responses across the nervous system (and the ways in which it interacts with other parts of the body) in response to psychological manipulations.

Psychophysiology is useful because self-reports of psychological state, especially as they concern affect, tend to be skewed by the subject [103]. Moreover, while there are some aspects of affect that people can reliably communicate based on introspection, there are other aspects of affect that people cannot reliably report—self-reporting dominance on Russell’s circumplex model is difficult, for instance (Section 2.1.2.2). Furthermore, self-report does not always directly reflect the physiological and behavioral aspects of emotion [18]. Psychophysiology is thus a useful method for examining psychological phenomena apart from as well as alongside self-report.

While psychophysiology is a useful tool, many studies highlight interindividual and intraindividual differences in psychophysiological responses. In particular, ANS activity has been shown to vary with differences in age [76], [83], [108], [123], [145], [171], [183], sex [25], [108], [145], [171], [180], race and ethnicity [83], [108], [114], [127], personality [24], [32], [121], [125], [142], [159], [179], and the time of day [47], [66], [171], [179], among other factors. Many of these studies also demonstrate significant interactions between these variables.

Concerning studies of music and affect through psychophysiology, it is important to consider several points. First, though general agreement has often been shown when groups of listeners rate the overall affective thrust of a selection of music, extra-musical variables (e.g., personality, culture, and context) may influence an individual listener’s experience of music [42], [157], [158], [169]. On the other hand, one person for whom a particular song evokes a memory of a painful experience in life will have a very different affective response upon hearing the song than another person for whom the song is completely unfamiliar [88]. Similarly, it is not inconceivable that a song in one’s own language may move them in a way very different than the same song sung in a foreign language. It has also become clearer that the mapping between music and affect isn’t as simple as one might like to believe: music cannot “influence emotional state in the same kind of reliable way that a drug such as caffeine affects arousal,” as Sloboda and Juslin put it [157, p. 87]. There is also evidence that there may be distinct patterns of psychophysiological responses—or, “protoresponses”—to musical stimuli between listeners [67]. Effective studies of music and emotion through psychophysiology must then either be designed to carefully control for each of these variables—a difficult task, indeed—or must work with very large sample sizes. It seems likely that only extremely large samples, or focused studies informed by insights gained through broader studies with large samples, have the potential to

“I like beautiful melodies telling me terrible things.”
—Tom Waits
produce widely generalizable conclusions about the relationships between music, emotion, and psychophysiology.

Compounding all of this is the questionable external validity of studying psychological or physiological phenomena in a laboratory [22]. Research has shown, for example, that heart rate and blood pressure measurements are affected simply by the presence of a doctor [110]. At a minimum, it is at least reasonable to assume that having measurements of one’s physiology taken while listening to music and sitting in a laboratory might produce different affective responses than if one were listening to the same music in a more naturalistic setting.

Because of these issues, it is logical that to be positioned to draw generalizable conclusions about natural psychophysiological responses to music, two things are important. First, a very large sample that crosses numerous demographic boundaries is useful, if not essential. Only with a large sample size will one be able to collect data from diverse populations, avoid fanatical control of all study variables, and maintain respectable statistical power. Second, it is beneficial to collect these data in as naturalistic a setting as is feasible. According to Schuller “multimodal affect data recorded…outside of the lab…[that] feature a higher diversity of participants, labeling, contained cultural aspects, languages, and situational context…[are] desperately needed” [150, p. 331]. Furthermore, as opportunities to collect such large and unique samples are extraordinarily rare, the means by which data are collected should be constructed in such as way as to permit collection in multiple locations, rapid on-the-fly iteration of study design, and with the end goal of the investigation of simultaneous and perhaps diverging lines of research inquiry.

1.5 Research Aims and Questions

A deepened understanding of autonomic psychophysiological response to emotional stimuli will serve to enable more efficacious affective computing interactions. Understanding the relationships between affective stimuli and psychophysiological responses can inform the interpretation of psychophysiological signals in general, as well as the use of music within an interaction to guide user affect. Given that music produces affective responses in humans not unlike those that affective stimuli in real-world interactions elicit, greater knowledge of psychophysiological responses to music might serve as a useful vector into the question of how to design more effective affective computing interactions. Concretely, the aim of the present study is to contribute to affective computing a deeper understanding of human psychophysiological response to musical stimuli and proven approaches for effectively gathering data in this research area.
1.5.1 **Gaps in Knowledge**

There are a number of gaps in knowledge that the research questions for this work aim to address. First, this need for large databases of naturalistic and multimodal affect data collected from diverse populations is a gap in knowledge in and of itself. In other efforts where a sample size of 20 or 50 may suffice, investigators are able to collect and analyze their own data. When, however, the task of amassing the data needed to investigate a number of a field at-large’s research questions is itself prohibitive, the collection, synthesis, and dissemination of such data stands alone as a contribution to the field. In fact, the Future of Research Communications and e-Scholarship (FORCE) group states that “data should be considered legitimate, citable products of research” and that “data citations should be accorded the same importance in the scholarly record as citations of other research” [39], and the National Science Foundation includes in their Grant Proposal Guide [120, p. II-12] acceptable research products as those that are “citable and accessible [products] including but not limited to publications, data sets, software, patents, and copyrights.”

In addition, large scale, human participant-facing studies are relatively rare in practice, and little is known about the best strategies to implement, deploy, and execute these studies. Furthermore, once large amounts of naturalistic, multimodal affect data, are amassed, strategies for synthesizing such data, structuring and representing them, and efficiently sharing the resulting databases are lacking, as well. For these reasons, the database and software framework presented in this work alone represent a substantial and unrivaled contribution to the corpus of affective computing scholarship.

Next, there is a dearth of studies that successfully model human psychophysiological response to salient moments in musical stimuli with a focus on ambulatory psychophysiological measures. Ambulatory psychophysiological measures are measures that can be taken while a subject is freely walking. Specifically, this lack of studies includes research focused on modeling such responses in terms only of characteristics and self-reported responses of the person, not in terms of the psychophysiology itself. In other words, little (if any) research seeks to answer the question, knowing the characteristics of a person, whether or not it is possible to predict how they will respond to a particular moment in a selection of music, for example.

In response to these gaps in knowledge, this dissertation answers the research questions described in the following sections.

---

2 For instance, numerous research awards are granted by the National Science Foundation (NSF) wherein database assembly and dissemination is at least the primary, if not only, research activity. See, for instance, the following NSF award numbers: 1539129, 0643887, 1023939, 1459715, 1355381, and 0517910.
1.5.2 Research Question 1 (RQ1)

What is an effective strategy for psychophysiological studies that simultaneously permits:

- The assembly of very large samples that cross numerous demographic boundaries,
- Data collection in naturalistic environments,
- Distributed and/or virtual study locations,
- Rapid iterations on study designs, and,
- The simultaneous investigation of multiple lines of research inquiry?

A database that meets the demands of the research problem defined here (see specifically Research Question 3 (RQ3) in Section 1.5.4) and, more generally, the ongoing needs of others in this research community, is supremely difficult to build in the best of circumstances, and altogether infeasible in others. It naturally calls for the collection of data over a long period of time and in multiple disparate locations. Further, and precisely because of the necessity of a distributed data collection endeavor, this work must be orchestrated in such a fashion as to allow for rapid iteration on and adjustments to study designs. Finally, a database that meets the needs of not only the present work but also the work of fellow scholars in the research community must be diverse enough to allow for the investigation of multiple lines of research inquiry. The effort of designing and deploying such a data collection effort is complicated by the relatively minimal guidance provided by previous work in this or other fields. The challenges in executing Emotion in Motion (EiM), a study begun in 2010 by the Music, Sensors, and Emotion (MuSE) research group, were the original impetus for this research question. It is largely one of “meta-research”—how does one successfully develop and execute research efforts not only at scale, but also to meet these numerous requirements of the data?

1.5.2.1 RQ1 Objectives

1. Document the challenges and shortcomings of the original EiM study execution strategy.

2. Develop requirements for a revised strategy that addresses these shortcomings. Additionally, the newly-developed strategy should be flexible enough to:

   a) Be quickly deployed to new remote or virtual study locations,
b) Be executed outside of the laboratory,

c) Be executed independent of investigator input or control,

d) Allow flexible and/or multiple study designs,

e) Permit instantaneous updates to these designs, and,

f) Be extensible to various formats and modalities of stimuli.

3. Realize a functioning software framework that satisfies these requirements.

4. Deploy studies using this framework to study locations in Taipei City and Taichung City, Taiwan, and Houston, Texas.

5. Evaluate the effectiveness of the redeveloped strategy following these deployments, developing recommendations for improvements to this strategy.

The research aim here is to explore and document the most effective ways for collecting such data, synthesizing these multimodal data into a coherent representation, and actively sharing artifacts of this kind with peers in the research community. The result of this work is the EiM database that is likely the largest database in existence of self-reported affective and physiological response to musical excerpts. EiM records several psychophysiological signals while subjects listen to randomly selected musical excerpts, as well as collects responses to a number of other qualitative and quantitative measures, including self-reported affective response, demographic information, musical background, and information about subject personality. This massive study has been underway for the last nine years, and continues to run in various locations around the world. It is a growing database of more than 67,000 signal recordings, from more than 23,000 participants. The EiM database, as well as the strategies developed in line with Research Question 1 (RQ1) and Research Question 2 (RQ2) form the bulk of the contribution of this work.

1.5.3 Research Question 2 (RQ2)

How can commodity hardware and general purpose software tools be used to record, represent, store, and disseminate very large, multimodal databases gathered from psychophysiological studies that use varying study designs?

Such databases as those described in RQ1 only realize their full potential when leveraged by peers within the research community. More immediate to the purposes of this work, the task of analysis of such large amounts of multimodal data is cumbersome, at best. To complicate matters, the few available tools and strategies for recording, representing, storing,
and disseminating large-scale multimodal data are either outdated or not flexible enough to permit all of these tasks with heterogeneous data (e.g., physiological signal time series and questionnaire responses) in any sort of harmonized fashion. Considering specific software solutions, yes, certain platforms for storing time series data are available, but these are often specifically geared toward recording and analyzing continuous rolling measurements (e.g., daily rainfall measures), not for high sample rate data paired with data in other forms. At a strategic level, how does one represent and store these heterogeneous data collected through varied study designs in a way that permits analysis between commonalities in the data while also respecting and clearly delineating differences in designs? Finally, how does one implement such a strategy with readily available tools such that dissemination amongst and collaboration between peer researchers is efficient and effective?

1.5.3.1 RQ2 Objectives

1. Develop requirements for a data synthesis, management, and access strategy capable of effectively:
   a) Unifying, representing, and storing heterogeneous data (e.g., high-fidelity time series and questionnaire responses),
   b) Capturing and communicating variations in the circumstances of data collection (e.g., varying study designs and populations),
   c) Facilitating access to commonalities in data between study designs,
   d) Allowing access to external investigators.

2. Given the unavailability of task-specific tools that can be used to realize this strategy, develop a method for using commodity hardware and general-purpose software to do so.

3. Use these tools to implement this strategy with the database developed in RQ1.

4. Evaluate the effectiveness of using these tools for this data synthesis, management, and access strategy.

1.5.4 Research Question 3 (RQ3)

How can the strongest psychophysiological responses present in the EiM database be characterized?

Specifically, how are the groups of participants that do or do not exhibit such responses characterized, and can these relationships be effectively captured in a computational model?
As previously cited, other work has demonstrated both within and outside of the psychophysiological paradigm that certain music can elicit reliable affective responses in listeners. As a test of the work performed for RQ1 and RQ2, Chapter 5 will describe several stimuli in the EiM study that regularly elicited strong physiological responses from large groups of participants spanning all collected demographics, personality traits, and other collected participant characteristics. RQ3 deals with responses to these stimuli specifically, with the end goal of determining whether or not it is possible to determine the likelihood of observing such responses from a given participant without knowledge of their psychophysiological response characteristics.

1.5.4.1 RQ3 Objectives

1. Develop an approach for automatic extraction of strong psychophysiological reactions to musical stimuli.

2. Use this method to partition observations in the EiM database for stimuli that elicit these responses into groups of participants who demonstrate such responses and those who do not.

3. Characterize the strong reaction-eliciting stimuli.

4. Model the differences between response/no-response groups both within and between musical stimuli.

1.6 SUMMARY

This chapter introduces the work accomplished in this dissertation and how it attends to gaps in related research on affective computing. As demonstrated here, it is evident that improved affective computing holds great promise, and that music is unquestionably an effective means of eliciting affect within a listener. Marrying these two statements, this dissertation will explore ways in which a better understanding of affective response to musical stimuli stands to inform affective computing as a whole. To do so, this work will describe the EiM study and database that can be leveraged in order to accomplish this (Chapter 3). Next, it will explore improved methods for gathering, synthesizing, and disseminating databases like the EiM database (Chapter 4). Finally, the potential of the EiM database will be explored for opportunities to model strong psychophysiological responses in terms of the characteristics of a given person (Chapter 5). In closing, Chapter 6 will provide an overview of all the work presented in this dissertation, and provide suggestions for continued research in the field. By the conclusion of this dissertation, it should be evident that the
database and system it contributes can serve as a way forward for carrying out this continued research.
The focus of this dissertation now turns to an exploration of other related work that has served to inform this research. Affect—the collection of emotions, moods, and preferences—is considered first. After laying a foundation for understanding affect, its relationship with music is considered. Affect and music have long been studied together, though efforts to study them with a psychophysiological approach with the aim of informing affective computing are relatively recent. Thus, an understanding of the interplay of affect, music, and psychophysiology will be important in framing the contributions provided by this work that will be presented in subsequent chapters.

2.1 AFFECT

Affect is the central topic that intersects the two larger subjects across which this dissertation is situated: music and psychophysiology. Thus, this review must necessarily take a discussion of affect, and emotion (an affective phenomenon) in particular, as its starting point. While humans have been theorizing on emotion since at least the time of the Ancient Greeks, the study of emotion within psychology as an independent discipline only began as recently as the nineteenth century. Nevertheless, the body of literature on emotion is voluminous. The aim of this section is to provide a brief background on the theories of emotion that have been the most influential, with greater attention given to those that have been chosen for use in the work described in this dissertation. Before doing so, however, it is necessary to establish a working definition for emotion itself.

2.1.1 What is an Emotion?

The Oxford English Dictionary defines emotion as “any strong mental or instinctive feeling, as pleasure, grief, hope, fear, etc., deriving especially from one’s circumstances, mood, or relationship with others” [128]. While this definition does accurately communicate the intense and object-directed nature of emotions, it does not clearly delineate emotions from other affective phenomena. In the second edition of their landmark text on music and emotion [91], Juslin and Sloboda identified a troubling trend among scholars in the affective sciences (and specifically researchers of music...
and emotion) of confusing terms when referring to affective phenomena—at times using them interchangeably, and at others simply using terms according to a researcher’s own nuanced definitions. In response to this, Juslin and Sloboda set forth definitions of a number of key terms related to affect, and required that all contributors to their work adhere to these definitions in order to facilitate communication and the integration of work. In line with this, this dissertation follows these definitions, as well. The subset of terms that are used throughout this dissertation that have been borrowed from [91] are reproduced in Table 2.1.

From the definitions in Table 2.1, it is clear that various affective phenomena are primarily differentiated by their duration, intensity, and object-directedness. The theories presented in Section 2.1.2 are concerned primarily with emotions: those brief and intense reactions to a specific object, where an object might refer to another person, an actual object, a set of circumstances, or something else.

2.1.2 Theories of Emotion

With this definition of emotion in hand, various theories of emotion are now considered. A number of reviews of the landscape of theories of emotion have been presented in the past. For a systematic introductory overview, readers are directed to Moors’s survey article [116]. Scherer’s survey with an eye toward different theories’ utility within affective computing is equally useful [148]. Moors’s survey contrasts differing theories of emotion through the lens of process description, paying attention to the ways in which different theories propose answers to the following questions at the functional, algorithmic, and implementational levels:

- What elicits an emotion, and how are emotions elicited?
- What determines the intensity of an emotion?
- How are emotions differentiated from one another?

Here, basic emotion theories (Section 2.1.2.1), constructivist theories (Section 2.1.2.2), and appraisal theories (Section 2.1.2.3) of emotion are reviewed. Where possible, the proposed answers that these classes of emotion theories provide to the above questions are provided in order to maintain coherence.

2.1.2.1 Basic Emotion Theories

Basic emotion theories (called affect program theories by Moors [116]) are rooted in Darwin’s observation of eight unique families of emotion across primates [38]. In [45] and [46] Ekman proposed and later refined his theory
Table 2.1: Definitions of key terms from Juslin and Sloboda’s *Handbook of Music and Emotion: Theory, Research, Applications* [91, p. 10].

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affect</strong></td>
<td>This is used as an umbrella term that covers all evaluative—or “valenced” (positive/negative)—states (e.g., emotion, mood, preference). The term denotes such phenomena in general. If that is not intended, a more precise term (e.g., mood, emotion, preference) is used instead.</td>
</tr>
<tr>
<td><strong>Arousal</strong></td>
<td>This term is used to refer to physical activation of the autonomic nervous system. Physiological arousal is one of the components of an emotional response, but could also occur in the absence of emotion (e.g., due to exercise). Arousal is often reflected in the “feeling” component (i.e., the subjective experience).</td>
</tr>
<tr>
<td><strong>Emotion</strong></td>
<td>This term is used to refer to quite a brief but intense affective reaction that usually involves a number of sub-components—subjective feeling, physiological arousal, expression, action tendency, and regulation—that are more or less “synchronized”. Emotions focus on specific “objects” and last minutes to a few hours (e.g., happiness, sadness).</td>
</tr>
<tr>
<td><strong>Feeling</strong></td>
<td>This term is used to refer to the subjective experience of emotions or moods. Feeling is one component of an emotion that is typically measured via verbal self-report.</td>
</tr>
<tr>
<td><strong>Mood</strong></td>
<td>This term is used to denote such affective states that are lower in intensity than emotions, that do not have a clear “object”, and that are much longer lasting than emotions (i.e., several hours to days). Moods do not involve a synchronized response in components like expression and physiology (e.g., gloomy).</td>
</tr>
<tr>
<td><strong>Preference</strong></td>
<td>This term is used to refer to longer-term affective evaluations of objects or persons with a low intensity (e.g., liking a particular type of music).</td>
</tr>
<tr>
<td><strong>Personality trait</strong></td>
<td>This term is used to refer to relatively stable affective dispositions, which are characterized by low intensity, and a behavioral impact, which is usually the result of an interaction with situational factors (e.g., a neurotic personality).</td>
</tr>
</tbody>
</table>
of basic emotions. Those emotions that qualified for Ekman as “basic” are fear, anger, disgust, sadness, contempt, amusement, pride in achievement, satisfaction, relief, and contentment. Ekman argued that eleven characteristics distinguish these from one another (e.g., facial expressions, distinctive physiology, automatic appraisal, and distinctive universals in antecedent events), and that these characteristics are distinct universal signals. Furthermore, basic emotions are set apart from other affective phenomena by these additional characteristics: distinctive appearance in development, presence in other primates, quick onsets and brief durations, their unbidden occurrence, association with distinctive thoughts and memories, and distinctive subjective experiences [46].

To Ekman, basic emotions are a product of evolution, useful in adapting to and performing fundamental life tasks. Thus, these fundamental life tasks (Ekman’s antecedent events) serve as the drivers or elicitors of emotion. When this occurs, an affect program is activated. When the program for a basic emotion is activated, the distinct configurations Ekman described (a unique physiological response, universal and distinct expression, and automatic appraisal) present themselves. The specifics of any affect program have evolved in service to the fundamental life task for which they are triggered (e.g., the fight-or-flight response when one is fearful).

Izard, another proponent of basic emotion, differs only slightly from Ekman in his conceptualization of emotion [71], [73], [75]. While Izard espoused the concept of basic emotions, his list of basic emotions differs slightly: interest, joy/happiness, sadness, anger, disgust, and fear. A notable difference between Izard’s Differential Emotions Theory and Ekman’s theory of basic emotions is the stress Izard placed on emotion schemas. Emotion schemas (or, “nonbasic emotions”) “consist of an evolved feeling [from a basic emotion] plus learned labels and concepts” [73, p. 265]. These emerge in early development, but following the emergence of basic emotions, and in step with one’s cognitive development. Most emotion schemas frequently involve some sort of appraisal process, and in fact, emotion schemas are roughly equivalent to the “emotions” in the parlance of appraisal theories (Section 2.1.2.3).

For both of these and other representative models of basic emotion theories, a basic emotion is a product of the evolutionary process that has a common set of components (e.g., neural circuitry, outward expression, physiological response, and feelings) that work together to serve a specific adaptive function of either motivation, organization, or regulation. These emotions are elicited by the presentation of a fundamental life task or situation. It is the cognitive processing or appraisal of an emotion that determines its intensity. Basic emotions are differentiated from one another according to the components or characteristics of an emotion as described by Ekman—unique expression, physiology, and antecedents.
In recent years a number of studies have presented evidence for the specificity of physiological responses to basic emotions (see [20], [21], [99] for reviews). The problem of specificity when it came to the physiological responses elicited by emotional experiences was one of their primary criticisms.

### 2.1.2.2 Constructivist Theories

**James’s Theory of Emotion** Perhaps the first comprehensive theory of emotion generation and perception was the feeling theory of emotion proposed by William James in 1884. James suggested that “bodily changes follow directly the perception of [an] exciting fact, and that our feeling of the same changes as they occur is the emotion” [82, p. 189–190]. In other words, according to James, emotions are those sensations that one feels as a result of physiological changes in the body that follow the perception of some source of “excitation”. He further described the multitudinous arrangements of physiological responses that uniquely denote different emotions: “no shade of emotion, however slight, should be without a bodily reverberation as unique, when taken in its totality, as is the mental mood itself” [82, p. 192]. Furthermore, the feeling of these bodily changes occurs instantaneously when such changes occur. If it were possible to divorce all bodily feelings of an emotion from the emotion itself, James argued that a “cold and neutral state of intellectual perception is all that remains” [82, p. 193]. Indeed, James goes so far as to say that human emotion divorced from its physical manifestation is a “nonentity”. In James’s model, some object or event acts upon a sense organ (e.g., the eyes, skin, or ears). Neural impulses travel to the sensory cortex to communicate as much where a neural activation pattern manifests—the correlate of the perception of the object (or event). Following this perception, further neural impulses travel to the muscles and heart, for example, where they produce such emotion-specific patterns of physiological changes. Again, these changes are registered by the surrounding sensory organs and relayed back to the sensory cortex. Here, another neural activation pattern manifests that is the correlate of the emotion itself. For James, emotions are these patterns of bodily sensations themselves; only physiological reactions are necessary for emotions. James considered the fact that volitionally bringing about certain physiological patterns of emotion can elicit the emotion itself (“working” oneself into a rage or moping about all day to evoke sadness) as evidence of this. In his theory, emotions are the effects, rather than the causes, of emotional behaviors—emotions in and of themselves serve no useful functions.
Schachter and Singer’s two-factor theory  In contrast to James, Schachter and Singer, noting difficulty in distinguishing different emotional states based on differences in physiological patterns alone, suggested that cognition must play an important role in the process by which one characterizes and interprets different emotional states. Given a pattern of physiological changes, one interprets them within their current context, and only then is the emotional state differentiated from other emotional states that may share a similar physiological response pattern. By this, “an emotional state may be considered a function of a state of physiological arousal and of a cognition appropriate to this state of arousal” [147, p. 398]. They hypothesized that a person whose physiology had been chemically aroused might reliably interpret their arousal as, for instance, anger, when placed in an environment suggestive of anger. Their initial study lent some support to this hypothesis, and prompted a great deal of research interest. These further investigations instead suggested that physiological arousal is not a necessary antecedent of emotion. However, Schachter and Singer’s two-factor theory, and the work that followed it, did convince many theorists of the equal importance of the cognitive and non-cognitive components of the emotional experience [137].

Dimensional theories  Whereas theories that are built on basic emotions model them as categorical (or mixtures of categorical emotions), dimensional theories conceptualize emotion as being described by combinations of different dimensions of affect—the most prevalent among these dimensions being valence, arousal, and dominance. Valence corresponds to the pleasantness of an emotion, arousal to activation or intensity, and dominance to the potency of an emotion. Four exemplar dimensional theories of emotion are those of Russell [144], Watson and Tellegen [175], Larsen and Diener [104], and Thayer [168].

Russell proposed that the best dimensional representation of affect is around the perimeter of a circle (circumplex) in two-dimensional bipolar space (Figures 2.1a and 2.1b) [144]. In his model, pleasantness–unpleasantness (valence) and arousal–sleep are the two bipolar dimensions. In these, his original figures, valence is represented along the horizontal axis, and arousal along the vertical axis. Emotions fall around the perimeter of the circumplex as combinations of the two dimensions. In addition, he posited that these terms for emotion represented fuzzy sets, rather than discrete categories. In other words, moving around the circumplex slowly passes between emotion terms, where partial membership in one or more of the terms is possible. Finally, Russell noted from several studies that in his model, emotions tend not to cluster around the axes, but tend to be distributed around the perimeter of the circumplex. This led Russell to believe that affective space lacks what he and Watson and Tellegen (below)
referred to as “simple structure”. Importantly, at times, even proponents of basic emotions use a dimensional framework in their approach because they permit for far easier measurement of the characteristics of emotion than other frameworks allow [74].

Watson and Tellegen, following a meta-analysis of studies of “transient” self-reported affect, found that the dimensions of positive affect and negative affect account for up to three quarters of variance in emotion ratings [175]. Their circumplex model is built around these two orthogonal, independent, and uncorrelated dimensions, where they note that affective space is most dense around high positive affect and high negative affect. Watson and Tellegen noted that their model is not at odds with other circumplex models—Russell’s model can, for instance, be drawn from theirs by taking the pleasantness-unpleasantness “axis” as valence and the strong engagement-disengagement axis as arousal (see also [122] for empirical support of this claim). Nevertheless, Watson and Tellegen argued that their rotation of these two basic dimensions is most appropriate for capturing relationships between affective descriptors, as these two dimensions reflect “fundamentally different processes”.

Moods, for Thayer [168], are closely related to emotions and can also be conceptualized in dimensional terms. However, where Watson and Tellegen considered positive affect and negative affect to be the most useful dimensions on which to model affective space, Thayer instead considered energy and tension to be the principal components of mood, as he noted that the terms that Watson and Tellegen grouped under positive affect mostly seemed to be related to energy, while those that Watson and Tellegen grouped under negative affect all seemed to be related to tension. From this, he derived four basic moods—calm-energy, calm-tiredness, tense-energy, and tense-tiredness—and described how humans naturally move between these states in regular daily patterns, and how these movements impact emotion.

In spite of the apparent clash between dimensional and discrete theories of emotion, the two approaches are not necessarily at odds with one another, but can be used individually or in a complementary fashion for modeling emotion [22]. While the underlying mechanisms responsible for emotion generation and perception differ wildly between dimensional, discrete (basic), and appraisal theories (Section 2.1.2.3), all of the approaches are compatible when it comes to the meaning of emotion terms (e.g., happy, sad, tense, or relaxed)—those with which this dissertation works. With respect to emotion terms, these different approaches are essentially different but compatible perspectives on how to situate such terms with respect to one another [48]. This dissertation makes no claims about how emotions are generated, or by which processes one may perceive and understand them. Instead, it is concerned with what one may experience,
for instance, when listening to a particular song. The need here is the ability to accurately represent an emotion or affective state, changes between
such states, and relationships between different states. On this point, these different theories of emotion are compatible. The present work primarily uses Russell’s circumplex to represent emotion. Not only is it often used in many other affective computing studies ([164] and [152], for example), but it lends itself particularly well to computational representation. In addition, of the available choices for modeling affect, the straightforward basic constructs of this circumplex model are most easily understood by study participants, particularly those for whom English is not their first language.

### 2.1.2.3 Appraisal Theories

As James regularly repeated, his theory of emotion was “pretty sure to be met with immediate disbelief,” and it did, in fact, receive a great deal of criticism.¹ Many arguments leveled at James’s theory took issue with the fact that the *appraisal* of an object or event played no role in emotion generation; rather, emotions were simply responses to sensory perceptions [138]. In the 1960s, Arnold [5] and Lazarus [105] presented their own (similar) theories of emotion generation. These are considered here together as *appraisal theory*.

In stark contrast to James’s suggestion that one is fearful *because* they tremble, appraisal theory posits that one is fearful because of *something*, and that *thing* is the object toward which the emotion is directed. Arnold suggested that emotion generation presupposes cognition of two types.

---
¹ Gardiner provided an overview of a number of these criticisms in [53].
First, one must make factual cognitions (e.g., the object is, in fact, real). Second, an appraisal of whether or the object is good or bad for the person is required. Lazarus concretized this appraisal stage further by stating that the object is deemed positive if it is congruent with one’s goals (e.g., to accomplish a task, to feel pleasure, or to survive), or deemed negative if it is incongruent with one’s goals [106]. In slightly different terms, the two proposed that emotions are differentiated by their characteristics in several dimensions: goodness vs. badness, presence vs. absence, and the potential one has of coping with the object or event. Taken together, these two cognitive processes serve as the driver in the generation of emotions.

In the years that followed, numerous researchers presented their own refinements to the appraisal theory of emotion. Ortony et al.’s model is useful for representing emotion in affective computing applications, but dimensional models are far easier to use in reporting emotion. Ortony et al.’s is perhaps the best known and most widely used of these theories; this is especially true in affective computing, as one of their stated goals was to construct a computationally tractable model of emotion. At the foundation of their model is the assumption that “emotions arise as a result of the way in which the situations that initiate them are construed by the experiencer” [126, p. 1]. Using evidence drawn from the language of emotions and self-report, they break emotions into valenced reactions of three different types: those to events, agents, and objects. These three types of reactions in turn elicit three basic classes of emotions: pleased–displeased (brought about by a reaction to an event), approving–disapproving (brought about by a reaction to an agent), and liking–disliking (brought about by a reaction to an object). One’s appraisal of an event, agent, or object is made with respect to one of three central variables: desirability, praiseworthiness, or appealingness, respectively. These, in turn, are evaluated in terms of the person’s goals, standards, and attitudes, respectively. Finally, the intensity of the emotion that is induced is proportional to the values of a number of variables (e.g., the sense of reality, proximity, unexpectedness, and arousal).

After presenting their theory, Ortony et al. went on to dissect twenty-two different emotions (both positively and negatively valenced) within the framework of their model, and described how other emotions could be similarly deconstructed.

Since Arnold and Lazarus originally presented their theories, appraisal theories have come to be widely accepted throughout psychology as the most likely explanation of the generation of emotions [117], evinced by expansions to the theory and its use for modeling emotion beyond the realm of psychology alone. Reisenzein attributed the success of appraisal theories to their ability to explain most of the complex peculiarities of emotion [138]. Among these are the facts that there are manifold different emotions, different people can react differently to the same objects (or events, or agents), the same emotion may be elicited by very different objects, learning of an object in different ways can produce the same
emotion, and that a person’s emotion may change in concert with their appraisal of an object.

This dissertation does not purport to offer new knowledge about the nature of emotions or how they are elicited. Rather, this brief review is provided in order to situate the choice made in this dissertation of how to model and represent emotion—and more generally, affect. Specifically, this work used dimensional models of emotion in order to solicit self-response from users in the study. While the underlying mechanisms of emotion vary from theory to theory, it is possible to draw connections between the qualities or “names” of emotions from one theory to another. A dimensional model is more easily measured through self-report questionnaires, and thus, this was the construct used to solicit such responses in the study this dissertation will describe.

2.2 MUSIC AND AFFECT

The relationship between music and affect is far more complex than was once thought. Slow music in a minor mode may indeed evoke sadness in a listener, and fast music in a major mode may evoke happiness, but such simplistic rules are not representative of the true complexity of the interaction between music and affect. This section provides an overview of the current understanding of the relationship between music and affect.

2.2.1 Reliable Perception of Expressed Musical Emotion

It is important to first address the difference between emotions that one perceives when listening to music, and emotions that music evokes within a listener. In the first case, music simply expresses emotion that a listener can decode and perceive. This is quite different from the second case, wherein music actually elicits an emotion within a listener that he or she experiences as their own. The next section addresses the question of whether or not music is capable of evoking emotion. This section focuses on the phenomenon of the perception of emotion that music expresses. While the present study is concerned with the genuine modulation of affect by music, the distinction between the two is nevertheless important, because of the implications of this distinction, particularly with respect to experimental design.

Empirical investigation of the perception of emotion expressed by music began in the 1930s with Hevner’s landmark work [61], which was introduced in Chapter 1. Much more recently, in 2002, Västfjäll reviewed a number of studies that used music to induce specific affective states [172]. Although the purpose of his review was to explore support for
the idea that music can reliably elicit certain affective responses, Västfjäll drew several wider conclusions about musically-expressed mood from the musical selections that various researchers used in their individual studies. While many musical selections were used in only a single study, a number of selections were used very frequently in multiple studies for the expression of particular target moods. For instance, Prokofiev’s *Russia under the Mongolian Yoke* was selected by researchers in 13 of the 41 studies to induce a “negative mood”. That independent researchers would gravitate toward certain musical selections as expressive and evocative of particular moods indicates that these selections (for reasons that will be reviewed in Section 2.2.3) are reliable in being perceived as expressive of these moods.

Juslin and Laukka presented not only similar evidence for the reliable perception of musically-expressed emotion, but also evolutionary support for why this perception is so reliable [89]. Comparing data from 104 studies of vocal expression and 41 studies of music performance, they found that the same acoustic cues that enable human perception of affective signals in communication are at work when searching for affective expression in music. For example, where one may use a slow rate of speech, little voice intensity variability, and the low fundamental frequency of one’s voice to perceive sadness, the same also occurs in the tempo, sound level, and pitch level of music that expresses sadness. Juslin and Laukka also found tentative support across these studies that our ability to perceive affective expression from these acoustic cues may develop as early as infancy. In an equally wide-ranging review the following year, the authors considered a number of the same questions [90]. In considering the question of whether or not music can express specific emotions, they found evidence from over 100 studies that not only can music express specific emotions, but also that listener agreement on what emotions a particular piece of music expresses is usually quite high for broad categories of emotion. While agreement on finer gradations of emotion within these categories tended to break down, at broader levels even musical training, age, and gender had largely inconsequential effects on listener agreement.

Not only do humans tend to perceive emotion expressed by music consistently, there is evidence that they do so across cultures and musical backgrounds, and that this ability develops early in life. While Balkwill and Thompson acknowledged that people do experience music through a filter of culture-specific knowledge (or naïvete), the psychophysical dimensions of music—aspects of the music that are perceptible without musical expertise (e.g., tempo, melodic complexity, rhythmic complexity, and pitch range)—are equally important in conveying musical emotion [7]). They hypothesized that listeners from one tonal system would be able to accurately perceive the emotion expressed by music from another tonal system.
by leveraging these psychophysical dimensions of music. In their studies, they found that subjects unfamiliar with Hindustani ragas were able to differentiate between the target emotions of joy, sadness, anger, and peace (though subjects were less sensitive to peace); that participants’ ratings of a number of psychophysical musical dimensions were significantly correlated with their perceptions of emotion; and that participants’ ratings of joy and sadness, in particular, were strongly correlated with expert ratings of the same. Adachi et al. presented recordings of children singing in English to Canadian children and adults, as well as Japanese children and adults, with the hypotheses that children could perceive affective content in music across musical cultures, and even better than adults [1]. Indeed, their results provided support for both of these hypotheses, and further revealed that not only do adults rely on these psychophysical musical dimensions for perceiving affective content in music from their own and other musical cultures, but children do, as well.

The studies reviewed here are only a small selection of numerous studies (for more, see [49], [92], [166], [167]) that provide evidence for the claim that humans generally tend to make similar judgments of the affective content of music irrespective of age, gender, culture, or education. In fact, the consistency with which humans perceive emotion expressed by music has permitted the automated recognition of musically-expressed emotion, based on music features alone, to become a burgeoning field of interest. In 2012, Yang and Chen reviewed the state of the art in automated music emotion recognition, presenting numerous already successful efforts to do so, though there are still a number of open problems in the field [181].

### 2.2.2 Perceived versus Felt Emotion

It is clear, then, that not only does music express affective content, but it does so in ways that are reliably and consistently perceived by listeners. Is the affective capacity of music limited to only the expression of affect, though? Also central to research into music and emotion is the question of whether music has the capacity to actually evoke a genuine affective response in the listener.

#### 2.2.2.1 Evidence for the Emotivists

One of the cognitivists’ (those who argue that music does not evoke emotion, but only expresses it) primary arguments against the possibility that music elicits real emotional responses was the supposed dearth of any such evidence. Kivy, in particular, claimed that listeners did not exhibit any behavior that would indicate that they experienced genuine emotion [96]. Similarly, Meyer [115], claimed that listeners may report feeling
a particular emotion, but when this occurs, it is possibly a conflation of perceived emotion with felt emotion. He argued—in the absence of evidence to the contrary at the time—that any physiological responses to heard music were different from other types of emotional experiences. The emotivist view also conflicts with appraisal theories of emotion, in that a “real” object to be appraised is absent (e.g., fearful music does not actually present the listener with a tangible thing to be feared). If listener claims of emotional changes in response to music are not to be trusted, then the emotivist camp is left to find evidence of the elicitation of emotion elsewhere, and psychophysiology has proven to be a fruitful area of research in this regard.

Intense emotional responses to some music seems to be an experience that many can recall. In one of the earliest studies to explore the possibility that music reliably evokes real emotional responses, Sloboda considered psychophysiological phenomena of which listeners could be aware on their own, including, for example, shivers, laughter, a lump in the throat, crying, and goose pimples \cite{154}, \cite{155}. Respondents to Sloboda’s survey indicated what music evoked these responses, the consistency with which they were evoked, and when in the music these responses occurred. Based on these responses, which all came from music performers, Sloboda was able to determine that melodic appoggiaturas and sequences along with descending harmonic progressions to the tonic through the circle of fifths elicited tears, that sudden harmonic shifts elicited shivers, and that acceleration and syncopation elicited increases in heart rate. Due to the population from which his survey drew, he was unable to make clear causal claims about the repeatability of his study with non-musicians, but this study nevertheless reinvigorated investigation through psychophysiology into the possibility that music does elicit genuine affective change in listeners.

The psychophysiological evidence of affective responses to music is measurable at far finer levels of detail than shivers and crying, however. If Meyer’s grounds for dismissing psychophysiological responses to music are sound, then two things must be true. First, no patterns of psychophysiological response to music would be evident across listeners when they hear a piece of music. Second, if multiple listeners do exhibit common psychophysiological response patterns, then these must all be the result of all listeners holding the same beliefs about how any given musical stimulus will affect them physiologically—a tenuous assumption at best, if such patterns do present themselves. In 1997, Krumhansl published the results of a sweeping study refuting Meyer’s claims and beginning to provide evidence of responses at these finer levels of detail \cite{101}, \cite{102}. Here, a range of cardiovascular, electrodermal, and respiratory measures were continuously recorded while participants listened to excerpts selected to elicit sadness, fear, and happiness. In one experiment, subjects gave
continuous ratings of one of felt sadness, fear, happiness, or tension while listening to the music, as well as provided self-reported ratings after hearing the excerpt of a number of different emotions they may have felt. In a second experiment, physiological signals were recorded while subjects listened to excerpts, and these subjects completed the same self-report questionnaire. In both experiments, subjects responded very similarly with regard to the emotions they experienced, indicating that the selected music was successful in eliciting the target emotional responses as measured both through continuous and summative ratings. Analyzing the physiological data, Krumhansl found that all of the physiological measures but one (respiratory sinus arrhythmia) differed reliably by excerpt type. When correlations between dynamic emotion ratings and physiological measures were taken, patterns of correlation between different physiological systems were significant. In addition, a number of physiological measures showed increased effects over time. A factor analysis of all physiological measures resulted in six factors related to each of blood pressure, electrodermal activity (EDA), skin temperature, respiration, heart rate, pulse transmission time, and pulse volume. Importantly, in a final analysis, summative self-report measures were shown to not correlate well with summative physiological measures.

While evidence continues to mount that music does elicit real emotions that are differentiable by psychophysiological response, the question of which psychophysiological measures are required to accurately differentiate emotions at a given level of granularity remains. For example, Nyklíček et al. showed that happy, sad, serene, and agitated states can be induced with music and predicted through physiological responses [122]. However, their attempts to do so with individual physiological signals were largely unsuccessful. It was only through multivariate discriminant analysis using a wide array of physiological signals that their efforts to differentiate these emotional states using physiology alone yielded significant accuracy. These results have two other important implications for the present study. First, analyses of both self-report and physiological data provide empirical support for the claim that emotions exist in two primary dimensions: valence and arousal. Second, though two of the physiological measures they recorded (inter-beat intervals and left ventricular ejection times) did correlate with the valence dimension, other measures are important in measuring valence by physiology alone.

Blood et al. presented a study in 1999 that underscored the importance of additional psychophysiological measures in predicting valence [14]. They presented subjects with excerpts of music while systematically increasing or decreasing the level of dissonance in the excerpt, which has been shown to correspond to different levels of pleasantness versus unpleasantness. By analyzing positron emission tomography (PET) scans during different
levels of dissonance, they showed that not only are different regions of the brain activated in emotional response to music than in the perceptual analysis of music, but also that different regions of the brain are involved in processing unpleasant versus pleasant stimuli. In particular, these regions involved in processing unpleasant versus pleasant stimuli are regions that have been found to be important in processing emotional response. Two years later, Blood and Zatorre reported the results of a similar study [13], but now designed to measure physiological response to intensely pleasurable musical stimuli. Subjects self-selected musical stimuli they knew to elicit chills. PET scans were taken and heart rate, electromyogram (EMG), respiration depth, EDA, and skin temperature were recorded while subjects passively listened to these selections, as well as control music. Blood and Zatorre further found that distinct brain regions—specifically, those associated with reward and known to respond to such biologically-motivated stimuli as food and sex—are associated also with such intensely pleasurable responses to music. They also observed significant changes in the other physiological measures in response to experimental stimuli. These (and numerous other studies since then [140], [182]) argue not only for patterned, differentiated psychophysiological responses to affective stimuli, but also for the current need for varied measures to distinguish between these patterns.

2.2.3 Relationships Between Musical Features and Affect

It is becoming ever clearer both that music does indeed *evoke* genuine affective responses in listeners, and that emotional responses, in particular, can be differentiated from one another based on the patterns of physiological changes associated with them. What is known, though, about the specifics of the relationships between the dynamic features of music and the affective responses that music elicits?

In one of the first surveys of its kind, Rigg compiled the results of thirteen previous studies (including six of his own) of the relationships between musical features and affective response [141]. Most of these studies loosely divided affective space according to Hevner’s circumplex model (Figure 2.3) of musical emotions [62].

Each of the surveyed studies used a subset of the emotion terms around Hevner’s circumplex. For the selected terms, researchers asked music listeners to rate the amount of each of a number of emotions they felt that various musical excerpts may have elicited. Combining the data from all of these studies, Rigg found a number of important agreements. These agreements (as well as points of divergence) are summarized in Table 2.2. Of particular note are the importance of tempo and register in determining the nature of emotions aroused by a particular piece of music. Beyond
Table 2.2: Summary of survey from Rigg [141, p. 430–434]. Emphasized characteristics of musical features indicate common findings across surveyed studies.

<table>
<thead>
<tr>
<th>Rhythm</th>
<th>Dignified; sacred; solemn; serious</th>
<th>Sad; tragic; melancholy; mournful; lamentation</th>
<th>Dreamy; longing; sentimental; pleading; sorrowful longing</th>
<th>Serene; tranquil; peaceful; dreamy; calm</th>
<th>Humorous; playful; whimsical; delicate; graceful; mischievous; amusing; funny; flip-pant; delicate</th>
<th>Happy; gay; very happy; exuberant; joy</th>
<th>Triumphant; agitated; exciting; restless; uneasy</th>
<th>Vigorous; majestic; exalting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Firm; regular; trochaic</td>
<td>Firm; moderately regular; moderately irregular</td>
<td>Flowing; regular</td>
<td>Flowing; regular; smooth</td>
<td>Flowing; regular; smooth; trochaic; iambic</td>
<td>Firm; rough</td>
<td>Firm; regular</td>
<td></td>
</tr>
<tr>
<td>Phrasing</td>
<td>Legato</td>
<td>Legato</td>
<td>Legato</td>
<td>Staccato</td>
<td>Staccato</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tempo</td>
<td>Slow; moderately slow</td>
<td>Slow; moderately slow</td>
<td>Slow; moderately slow</td>
<td>Fast; moderately fast; moderately slow</td>
<td>Fast</td>
<td>Fast; moderately fast</td>
<td>Fast; slow</td>
<td></td>
</tr>
<tr>
<td>Register</td>
<td>Low; moderately low</td>
<td>Low; moderately low</td>
<td>High</td>
<td>High; moderately high</td>
<td>High; moderately high</td>
<td>Low; high</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>Major</td>
<td>Minor</td>
<td>Major</td>
<td>Major</td>
<td>Major</td>
<td>Major</td>
<td>Minor</td>
<td></td>
</tr>
<tr>
<td>Melody</td>
<td>Ascending; narrow range</td>
<td>Narrow range</td>
<td>Ascending; narrow range</td>
<td>Descending; wide range; narrow range</td>
<td>Wide range</td>
<td>Descending; wide range; narrow range</td>
<td>Descending</td>
<td></td>
</tr>
<tr>
<td>Harmony</td>
<td>Simple (primarily consonant)</td>
<td>Complex (with dissonance)</td>
<td>Simple</td>
<td>Simple; moderately simple</td>
<td>Simple; moderately simple</td>
<td>Simple</td>
<td>Complex</td>
<td>Complex; simple</td>
</tr>
<tr>
<td>Dynamics</td>
<td>Few changes; moderately soft</td>
<td>Soft; moderately soft; few changes; moderately loud</td>
<td>Quick changes; moderately loud</td>
<td>No changes; soft; moderately soft</td>
<td>Quick changes; soft</td>
<td>Moderately few changes; moderately soft</td>
<td>Loud; moderately loud; some changes</td>
<td>Moderately few changes; moderately loud</td>
</tr>
</tbody>
</table>
the listener checks all the adjectives which he finds appropriate to the music. This makes the business of choosing responses easy and convenient for the listener, and allows an objective and quantitative treatment by the experimenter. From the original response sheets of the listeners, each of whom had a fresh copy of the adjective list for every different musical composition, it was a simple matter to tabulate all the votes for any one adjective from all the listeners for each composition alone, or for all the compositions which are alike in certain respects. Then, by comparing the numbers of votes for different adjectives, the meanings or affective characteristics of the compositions could be ascertained.

Figure 2.3: Hevner’s circumplex model of musical affect [62, p. 249]. (From American Journal of Psychology. Copyright ©1936 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.)

this, as Rigg notes, there are far more points where the various studies agree than those where they disagree. Since this survey, numerous other studies have presented additional evidence in agreement with these earlier results [36], [52], [89]. In a later survey of the effects of time-, pitch-, and texture-related musical elements on affect, Bruner came to many of the same conclusions as Rigg [19]. His survey included many of the same studies that Rigg surveyed, but also included a great many other surveys from researchers in the marketing domain. Unsurprisingly, Bruner noted all of the same agreements in the literature as Rigg.

Several years later, Gabriellsson and Juslin added to this body of knowledge by examining different musical performances of the same melodic segments played by different performers and with different expressive intentions. While the focus of their study was on performative intent, some of their findings have important implications for this work. In particular, their analyses focused not only on higher level features of music (e.g.,
Table 2.3: Lower-level musical features as they relate to basic emotions [51].

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>Moderate timing variations, sharpened contrasts in dotted rhythms, “airy” articulation, quick onsets, and bright timbres</td>
</tr>
<tr>
<td>Sadness</td>
<td>Large timing variations, softened contrasts in dotted rhythms, legato, and softened onsets</td>
</tr>
<tr>
<td>Anger</td>
<td>Sharpened contrasts in dotted rhythms, no ritards, avoidance of legato, very sharp onsets, and harsh timbres</td>
</tr>
<tr>
<td>Fear</td>
<td>Very large timing variations and staccato articulation</td>
</tr>
<tr>
<td>Tenderness</td>
<td>Reasonably large timing variations, softened dotted contrasts, legato, soft and slow onsets, and soft timbres</td>
</tr>
<tr>
<td>Solemnity</td>
<td>Small timing variations and mostly sharp onsets</td>
</tr>
<tr>
<td>“No expression”</td>
<td>Very little or no timing variations, no ritards and neutral note onsets</td>
</tr>
</tbody>
</table>

tempo and sound level), but also on lower-level features of the music (e.g., deviations in timing between the score and performance, use of different timbres, and articulation). A summary of their findings related to these lower-level features is given in Table 2.3.

Bruner noted that “music is not simply a generic sonic mass, but rather a complex chemistry of controllable elements” [19, p. 94]. This observation led Bruner and other researchers to consider the roles that individual subcomponents play in shaping the affective responses that music evokes (e.g., tempo, rhythm, melodic shape, and harmonic structure). Ironically, though, while many studies have considered more granular dimensions of music, few (if any) have considered what is arguably the most important dimension of music: time. A piece of music—and its rhythm, dynamics, and other features—unfolds over time, and psychophysiological response to music similarly unfolds over time. There seem to be no existing studies of music and psychophysiology that consider specifically the relationships between changes in musical features and changes in psychophysiological response.

It is probable, though, that because of the complexity of the relationship between music and affect and the various interindvidual differences in psychophysiological response that have already been discussed, it is impossible to predict affective change in a person, given a particular piece of music, based solely on the characteristics of the music. Context, motivations, characteristics of the person, and other factors must be considered carefully [90], [156]. Being careful to keep this in mind, if the research community can begin to better to understand how human psychophysiological response behaves in time in response to changes in richly affective stimuli, it will be better positioned to develop and deploy more effective affective systems.
2.3 SUMMARY

The aim of this chapter has been to summarize recent work in the areas of affect, music, and psychophysiology, with a specific focus on their confluence. This background material is the foundation of the following chapters. Chapter 3 introduces EiM, which itself is a study of music, affect, and psychophysiology. Informed by previous work with EiM, Chapter 4 lays out a new, flexible strategy for outlining similar studies. Chapter 5 then specifically deals with analysis of affective and psychophysiological data gathered through EiM. As such, a grasp of these fundamentals is important in not only understanding, but effectively critiquing the chapters to follow. It should be noted that a great amount of related work applicable directly to databases of affective response and strong responses to music has been placed directly in Chapters 4 and 5 directly, in order to make this background material immediately available while reading those chapters.
EMOTION IN MOTION

The last two chapters have motivated the need for large-scale studies of human affect in naturalistic environments, and argued for the use of music as a stimulus in prompting characteristic affective responses. The Emotion in Motion (EiM) study was designed as exactly this type of large-scale study. Reflection upon early experiences with EiM and its extended data collection schedule provided the opportunity to consider the research problems of how to optimally plan, design, and execute these kinds of studies. This chapter provides an overview of the EiM study itself, detailing the experimental design, apparatus, methods, and procedures. Then, with an understanding of the design and trajectory of EiM, Chapter 4 will present the knowledge that this study has produced with respect to these larger research questions (Research Question 1 (RQ1) & Research Question 2 (RQ2)).

3.1 STUDY OVERVIEW

EiM is a study of human affective response to musical stimuli that began in 2010, and has run almost daily since then [81]. Emotion in Motion explores the links between affect and music through the lens of psychophysiology. As of the publication of this work, the study has recorded 65,883 pairs of psychophysiological signals from 23,273 participants. EiM was originally launched by the Music, Sensors, and Emotion (MuSE) research group in the Sonic Arts Research Centre at Queen’s University Belfast, under the direction of Benjamin Knapp, of which I was a member. In 2012, I assumed the responsibilities of continued improvements to (and a complete overhaul of) the apparatus, iterations on the study design and methods, and staging of the study in a number of countries around the world. I and other members of the MuSE research group have published results drawn from the EiM database [15], [29], [67], [76]–[78], [80], [81]. Context for this chapter is provided first by examining the goals of the study itself. As EiM has developed iteratively over time, this chapter provides an overview of these various iterations—where they have occurred, their durations, and the number of participants that have participated. In the context of these iterations, the study apparatus and methods are discussed.

1 The processing and cleaning of incoming data into the database is an ongoing process, and these figures represent only the data currently available in the database—a great deal of other data have been recorded, as well, and await addition to the database.
It is important to note that, at times, this dissertation refers to the EiM study, database, and system. This chapter lays out the study and the data gathered through it. Chapter 4 deals with the EiM system that was developed to execute later iterations of the EiM study, and the resulting EiM database that is used as part of this system for dissemination of and collaboration on the EiM data itself.

### 3.2 Study Goals

EiM was originally designed with three specific goals. First, data should be collected outside of a laboratory, but still in a semi-controlled setting, in order to address the concern of the generalizability of results when physiological responses are measured in a laboratory environment [22], [110]. Second, a large sample size is imperative, precisely because of issues that may arise in such a semi-controlled environment, and to address the many potential inter- and intraindividual differences in physiological responses (see Section 1.4). Not only is a large sample size imperative, but it was important that the overall sample draw from numerous demographically diverse populations [150]. Finally, it was important to design EiM in such a manner as to allow for the exploration of a number of different research questions.

To meet these goals, the experimental design is structured around three stages (Figure 3.1). First, general data about each participant are collected, including demographics, data about their musical experience and preferences, and in later iterations, data about participant personality. Next, participants are presented with a series of musical stimuli. During the presentation of these stimuli, psychophysiological responses are recorded from the participant. Following each stimulus, the participants provide self-reported measures of affective response for a variety of constructs. Finally, after the series of musical stimuli have been presented, participants are asked a series of final questions (e.g., which stimulus they preferred most and with which stimulus they were the most engaged).
3.3 Iterative Design

The *EiM* study was originally designed to run for several months as an individual exhibit within the BIORHYTHM\(^2\) exhibition, one of the headline exhibitions of Science Gallery Dublin.\(^3\) BIORHYTHM ran from July 2, 2010 to October 10, 2010. This period provided the *MuSE* research group with the opportunity to explore several study designs (five, in total). Following its initial launch at Science Gallery Dublin, the BIORHYTHM exhibition traveled to and was installed in a number of different venues around the world. These installations have given the *MuSE* group further opportunities to construct a far larger and diverse sample, as well as continue to refine and modify the study design. Thus far, *EiM* has seen those iterations (an iteration indicates a new location and/or modifications to the study design) shown in Table 3.1. The number of participants that have participated and physiological recordings that have been collected are given Table 3.2. My own involvement with the *EiM* project began during the preparation for the seventh iteration, installed in Bergen, Norway.

3.4 Apparatus

The original study was run on an individual computer terminal with which the participant interacted. For recording physiological data, the study used the Mediaid POX020-105 sensor coupled with the M15HP

---

Table 3.1: *EiM* iterations, locations, and start and end dates.

<table>
<thead>
<tr>
<th>Iter.</th>
<th>Location</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Science Gallery Dublin, Ireland</td>
<td>July 1, 2010</td>
<td>July 21, 2010</td>
</tr>
<tr>
<td>3</td>
<td>Science Gallery Dublin, Ireland</td>
<td>Aug. 17, 2010</td>
<td>Aug. 29, 2010</td>
</tr>
<tr>
<td>6</td>
<td>Eyebeam, New York City, United States</td>
<td>June 14, 2011</td>
<td>Aug. 5, 2011</td>
</tr>
<tr>
<td>7</td>
<td>Vilvite, Bergen, Norway</td>
<td>Dec. 1, 2011</td>
<td>Ongoing</td>
</tr>
<tr>
<td>8</td>
<td>Science Centre Singapore, Singapore</td>
<td>June 14, 2012</td>
<td>Aug. 13, 2012</td>
</tr>
<tr>
<td>10</td>
<td>National Taiwan Science Education Centre, Taipei City, Taiwan</td>
<td>Dec. 7, 2014</td>
<td>June 10, 2015</td>
</tr>
<tr>
<td>11</td>
<td>National Museum of Natural Science, Taichung City, Taiwan</td>
<td>July 1, 2015</td>
<td>Nov. 29, 2015</td>
</tr>
<tr>
<td>12</td>
<td>The Health Museum, Houston, Texas</td>
<td>January 13, 2018</td>
<td>July, 31, 2018</td>
</tr>
</tbody>
</table>

\(^2\)https://dublin.sciencegallery.com/biorhythm

\(^3\)https://dublin.sciencegallery.com/
Table 3.2: EiM sample sizes by location.

<table>
<thead>
<tr>
<th>Location</th>
<th>Participants</th>
<th>Excerpt Auditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taichung City</td>
<td>4855</td>
<td>2517</td>
</tr>
<tr>
<td>Taipei City</td>
<td>8595</td>
<td>30657</td>
</tr>
<tr>
<td>Singapore</td>
<td>1730</td>
<td>5459</td>
</tr>
<tr>
<td>New York City</td>
<td>879</td>
<td>2790</td>
</tr>
<tr>
<td>Manila</td>
<td>2212</td>
<td>7201</td>
</tr>
<tr>
<td>Dublin</td>
<td>3677</td>
<td>12675</td>
</tr>
<tr>
<td>Bergen</td>
<td>1325</td>
<td>4584</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>23273</strong></td>
<td><strong>65883</strong></td>
</tr>
</tbody>
</table>

pulse oximetry module⁴ for pulse oximetry and the Infusion Systems BioEmo⁵ for electrodermal activity (EDA). Analog readings from each of these sensors were collected at a sampling rate of 250Hz [16] by dedicated Arduino Uno microcontroller boards.⁶ Each Arduino was connected via USB to the study terminal. The participants wore the sensors on three fingertips of one hand, and interacted with the study interface with a mouse alone using their free hand. The study interface and data recording software were originally implemented in the Max/MSP visual programming environment.⁷ To listen to the stimuli, participants wore a pair of Sennheiser HD 280 Pro headphones.⁸ These closed, around-the-ear headphones incorporate up to 32dB of ambient noise cancellation and have a frequency response of 8Hz–20kHz.

In each study location, four terminals for participation in the study were provided, both to meet visitor demand, as well as to mitigate issues with problems such as sensor breakage. The study terminals were located in a quiet area of the larger exhibition and lit at comfortable levels. In addition, a short wall was set at the back of each terminal in order to limit distractions to the participant (Figure 3.2).

Up until the iteration deployed to Taipei City, Taiwan, EiM used the apparatus originally designed and constructed for the initial 2010 iterations. The underlying architecture of the apparatus, as well as components of the study design, changed dramatically for the iterations deployed to Taipei

---

⁴ http://www.mediaidinc.com/OemSolutions.html
⁵ http://infusionsystems.com/catalog/product_info.php/products_id/203
⁶ http://www.arduino.cc/en/Main/ArduinoBoardUno
⁷ https://cycling74.com/
⁸ http://www.sennheiser.com/hd-280-pro
When a participant chose to participate with the EiM study and sat at one of the terminals, they were initially presented with a brief description of the study, and were provided with the opportunity to provide their consent to participate or cease their participation. Following this, instructions for placing the physiological sensors on their fingers and headphones over their ears were provided. Indications of whether or not the physiological signals were of acceptable quality were given, and the participant was guided to reseat the sensors until such a signal was attained. After this was completed, a series of screens gathered a number of responses to questions about the participant’s demographics and musical experience and expertise. (For screen shots of all of the screens seen by the participant, see Appendix A.)

3.5 Data Collected

The data collected during the various iterations of EiM included demographic data, measures of musical experience, data related to participant personality, self-reported measures of affective response, and measures
of psychophysiological response. Summative self-reported measures of affective response and continuous psychophysiological measures were recorded for each musical selection, while demographic, musical experience, and personality data were recorded once for each participant. This section describes each of these measures in turn. A number of these measures were common across all iterations of the study, while several of them were unique to a single iteration or subset of iterations of the study. To make these distinctions clear, those iterations for which data were collected for each variable are delineated, and those measures that were common across all iterations are later highlighted.

3.5.1.1 Demographic Variables

Typical measures of participant demographics were taken in all iterations. The details of these measures are given in Table 3.3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Text/Choices</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male, Female</td>
<td>1–9</td>
</tr>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>10–12</td>
</tr>
<tr>
<td>Age</td>
<td>Year of birth</td>
<td>1–9</td>
</tr>
<tr>
<td></td>
<td>Age in years</td>
<td>10–12</td>
</tr>
<tr>
<td>Nationality</td>
<td>Irish, British, Rest of the World</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Irish, Rest of the World</td>
<td>2–5</td>
</tr>
<tr>
<td></td>
<td>U.S.A., Rest of the World</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Norwegian, Rest of the World</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Singaporean, Rest of the World</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Filipino, Rest of the World</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Complete list of nationalities</td>
<td>10–12</td>
</tr>
<tr>
<td>Hearing Impairments</td>
<td>“Do you have any hearing impairments?”: Yes, No</td>
<td>1–12</td>
</tr>
<tr>
<td>Visual Impairments</td>
<td>“Do you have any visual impairments?”: Yes, No</td>
<td>6–9</td>
</tr>
</tbody>
</table>

3.5.1.2 Variables Related to Musical Training and Experience

One’s levels of musical training and experience have been shown to have an effect on musical preference and affective response to music [100], as well as on psychophysiological response to music [11]. As such, it is important to take this into account when designing studies of affective response to musical stimuli. Thus, information about participants’ musical
training and expertise was also gathered. These measures are given in Table 3.4.

Table 3.4: Measures of musical training and experience.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Text/Choices</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musical Knowledge</td>
<td>“Do you consider yourself a musician or to have specialist musical knowledge?”: Yes, No</td>
<td>1–9</td>
</tr>
<tr>
<td>Musical Expertise</td>
<td>“From 1 to 5, how would you rate your musical expertise?”: Five-point scale from “No musical expertise” to “Professional musician”</td>
<td>1–6</td>
</tr>
<tr>
<td></td>
<td>“How would you rate your musical expertise?”: Five-point scale from “No expertise whatsoever” to “An expert”</td>
<td>7–12</td>
</tr>
<tr>
<td>Musical Preferences</td>
<td>“Select which of the following styles you regularly listen to (check all that apply)”: Rock, Pop, Classical, Jazz, Dance, Hip-Hop, Traditional Irish, World, None</td>
<td>1–6</td>
</tr>
<tr>
<td></td>
<td>“Select which of the following styles you regularly listen to (check all that apply)”: Rock, Pop, Classical, Jazz, Dance, Hip-Hop, World, None</td>
<td>7–12</td>
</tr>
</tbody>
</table>

3.5.1.3 Variables Related to Personality

Because of the role that personality plays both in psychophysiological response as well as music perception [44], [165], later iterations (beginning with iteration 10) included measures of personality. The research community around personality has come to the general consensus that personality traits (defined by Allport and Odbert as “generalized and personalized determining tendencies—consistent and stable modes of an individual’s adjusting to his environment” [2, p. 26]) can be categorized into five broad dimensions: extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience [86]. A number of inventories have been built around this model, but many of these inventories, including the widely used Big Five Inventory-44 (44-item) (BFI-44) [10], [85], [86] take considerable time to complete. In an effort to balance time expectations of participants with the use of an inventory with proven reliability and validity, the shorter Big Five Inventory-10 (10-item) (BFI-10) was used [135]. The BFI-10 includes two items for each of the five constructs. Participants were asked to rate their level of agreement with each of the prompts in Table
using a five-point rating scale (“Disagree strongly”, “Disagree a little”, “Neither agree nor disagree”, “Agree a little”, and “Agree strongly”).

<table>
<thead>
<tr>
<th>Trait Prompt</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extroversion (Reversed)</td>
<td>“I see myself as someone who is reserved.”</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>“I see myself as someone who is generally trusting.”</td>
</tr>
<tr>
<td>Conscientiousness (Reversed)</td>
<td>“I see myself as someone who tends to be lazy.”</td>
</tr>
<tr>
<td>Neuroticism (Reversed)</td>
<td>“I see myself as someone who is relaxed, handles stress well.”</td>
</tr>
<tr>
<td>Openness to experience (Reversed)</td>
<td>“I see myself as someone who has few artistic interests.”</td>
</tr>
<tr>
<td>Extroversion</td>
<td>“I see myself as someone who is outgoing, sociable.”</td>
</tr>
<tr>
<td>Agreeableness (Reversed)</td>
<td>“I see myself as someone who tends to find fault with others.”</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>“I see myself as someone who does a thorough job.”</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>“I see myself as someone who gets nervous easily.”</td>
</tr>
<tr>
<td>Openness to experience</td>
<td>“I see myself as someone who has an active imagination.”</td>
</tr>
</tbody>
</table>

### Variables Related to Affective Response

For each participant, both explicit and implicit measures of affective response to each musical excerpt were taken. The explicit measures were in the form of ratings along several different rating scales. Implicit measures of affective response were the recorded psychophysiological signals: EDA and pulse oximetry (POX).

Self-reported measures of affective response Over all of the iterations of EiM, a number of different self-reported measures of affective response were taken. More often than not, these included typical measures of affect, such as ratings taken using the Self-Assessment Manikin (SAM) [17] for measurement along Russell’s dimensions of affect [144]. Other measures, such as those scales developed to measure music-specific emotions, were used in some iterations, also. All of these measures are detailed in Table 3.6.
## Table 3.6: Self-reported measures of affective response.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Text/Choices</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity</strong></td>
<td>“How active or passive did the music make you feel?”: Five-point scale from “Very lively” to “Very drowsy”</td>
<td>1–3</td>
</tr>
<tr>
<td></td>
<td>“How active or passive did the music make you feel?”: Five-point scale from “Very drowsy” to “Very lively”</td>
<td>6–9</td>
</tr>
<tr>
<td></td>
<td>“How active or passive did what you have just heard make you feel?”: Five-point scale from “Very drowsy” to “Very lively”</td>
<td>10–12</td>
</tr>
<tr>
<td><strong>Intense Reactions</strong></td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Chills”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Shivers”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Thrills”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Goosebumps”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“Did you experience any of these physical reactions while you were listening?: Chills, Shivers, Thrills, Goosebumps”: Five-point scale from “Not at all” to “Intensely”</td>
<td>2–3, 6–9</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>“How in control did you feel?”: Five-point scale from “Weak (without control, submissive)” to “Empowered (in control of everything, dominant)”</td>
<td>6–12</td>
</tr>
<tr>
<td><strong>Distraction</strong></td>
<td>“How concentrated were you during this experiment?”: Five-point scale from “Very distracted” to “Very concentrated”</td>
<td>10–12</td>
</tr>
<tr>
<td><strong>Engagement</strong></td>
<td>“How involved and engaged were you with the music you have just heard?: Five-point scale from “I was engaged with the music and responding to it emotionally” to “Not at all engaged, my mind was elsewhere”</td>
<td>1–3</td>
</tr>
<tr>
<td></td>
<td>“How involved and engaged were you with the music you have just heard?: Five-point scale from “Not at all engaged, my mind was elsewhere” to “I was engaged with the music and responding to it emotionally”</td>
<td>6–9</td>
</tr>
<tr>
<td></td>
<td>“How involved and engaged were you with what you have just heard?: Five-point scale from “Not at all engaged, my mind was elsewhere” to “I was engaged with the music and responding to it emotionally”</td>
<td>10–12</td>
</tr>
<tr>
<td></td>
<td>“Which piece of music were you most engaged with?”</td>
<td>1–9</td>
</tr>
<tr>
<td><strong>Enjoyment</strong></td>
<td>“Which piece of music did you enjoy most?”</td>
<td>1–9</td>
</tr>
<tr>
<td><strong>Familiarity</strong></td>
<td>“How familiar are you with this music?”: Five-point scale from “Never heard it before” to “I listen to it regularly”</td>
<td>1–3, 6–9, 9</td>
</tr>
</tbody>
</table>
Table 3.6 – Continued from previous page

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Text/Choices</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inspiration⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Inspired”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td>Joyful Activation⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Joyful Activation (Joyful, Amused, Bouncy)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Like/Dislike</td>
<td>“How much did you like/dislike the song?”: Five-point scale from “I loved it” to “I hated it”</td>
<td>1–3</td>
</tr>
<tr>
<td></td>
<td>“How much did you like/dislike the song?”: Five-point scale from “I hated it” to “I loved it”</td>
<td>6–9</td>
</tr>
<tr>
<td></td>
<td>“How much did you like/dislike what you have just heard?”: Five-point scale from “I hated it” to “I loved it”</td>
<td>10–12</td>
</tr>
<tr>
<td>Nostalgia⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Nostalgia”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Nostalgia (Nostalgic, Dreamy, Melancholic)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Overwhelmedness⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Overwhelmed”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td>Peacefulness⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Peacefulness (Serene, Calm, Soothed)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Power⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Power (Fascinated, Overwhelmed, Feelings of transcendence and spirituality)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Sadness⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Sadness (Sad, Sorrowful)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Spirituality⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Spirituality”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td>Tenderness⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Tenderness”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Tenderness (Tender, Affectionate, In love)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
</tbody>
</table>

⁹ These measures were taken from the Geneva Emotional Music Scale (GEMS) [182]
Table 3.6 – Continued from previous page

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Text/Choices</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tension</td>
<td>“How tense or relaxed did you feel while you were listening?”: Five-point scale from “Very tense” to “Very relaxed”</td>
<td>1–3, 10–12</td>
</tr>
<tr>
<td>Tension⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Tension (Tense, Agitated, Nervous)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Transcendence⁹</td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Transcendence (Fascinated, Overwhelmed, Feelings of transcendence and spirituality)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
<tr>
<td>Valence</td>
<td>“How positive or negative did the music make you feel?”: Five-point scale from “Very positive” to “Very negative”</td>
<td>1–3</td>
</tr>
<tr>
<td></td>
<td>“How positive or negative did the music make you feel?”: Five-point scale from “Very negative” to “Very positive”</td>
<td>6–9</td>
</tr>
<tr>
<td></td>
<td>“How positive or negative did what you have just heard make you feel?”: Five-point scale from “Very negative” to “Very positive”</td>
<td>10–12</td>
</tr>
<tr>
<td>Wonder⁹</td>
<td>“How strongly did you experience any of these physical reactions while you were listening?: Wonder”: Five-point scale from “Not at all” to “Intensely”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“How strongly did you experience any of the following feelings while you were listening?: Wonder (Filled with wonder, Dazzled, Moved)”: Five-point scale from “Not at all” to “Intensely”</td>
<td>4</td>
</tr>
</tbody>
</table>

**Psychophysiological Measures** In every iteration, POX and EDA were recorded during the presentation of each musical excerpt. The Medi-aid POX020-105 sensor with the Mediaid M15HP pulse oximetry module was used for recording pulse oximetry, and the Infusion Systems BioEmo was used for recording EDA. Each signal was recorded through the analog-to-digital converter, providing 10-bit precision, at a sampling rate of 250Hz. For each musical excerpt, the two signals were stored together along with several features derived using EDATool [⁷⁷]¹⁰.

3.5.1.5 Common Data

While the experimental design did vary somewhat from location to location, a large subset of the design remained the same in each location. This permitted both the exploration of new research questions in new locations and refinements to the study design with respect to existing research questions, but also the comparison of data between locations. The

¹⁰ These raw and derived signals are stored in the database, but resampled signals and features extracted using modified feature extraction procedures are what have been used in the analysis for this dissertation. (See Section 4.5.1.2.)
Table 3.7: Common data from all EiM locations.

<table>
<thead>
<tr>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of birth</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Hearing impairments</td>
</tr>
<tr>
<td>Most engaging musical excerpt</td>
</tr>
<tr>
<td>Most enjoyable musical excerpt</td>
</tr>
<tr>
<td>Musical background</td>
</tr>
<tr>
<td>Musical expertise</td>
</tr>
<tr>
<td>Nationality</td>
</tr>
<tr>
<td>Visual impairments</td>
</tr>
<tr>
<td>EDA</td>
</tr>
<tr>
<td>POX</td>
</tr>
</tbody>
</table>

measures that were taken for every participant at every location are given in Table 3.7.

3.5.2 Musical Stimuli

In every iteration, the participants listened to at least one musical excerpt. Table 3.8 gives a list of all musical stimuli used in EiM, identifying those iterations in which each stimulus was used. The number of stimuli presented in each iteration varied, as well. Beginning with the iterations in Taiwan, participants first listened to a recording of pink noise as a control stimulus, followed by a single randomly selected musical excerpt; each of these auditions were followed by collection of the same measures of affective response, with one exception: a measure of familiarity was not recorded for the pink noise excerpt. Table 3.9 gives the size of the media pool and stimuli count for each iteration.

Table 3.8: Musical stimuli used in EiM iterations.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
<th>Iterations Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adele</td>
<td>Someone Like You</td>
<td>10</td>
</tr>
<tr>
<td>Alicia Keys</td>
<td>Empire State of Mind (Part II) Broken Down</td>
<td>6–9</td>
</tr>
<tr>
<td>Anton Bruckner</td>
<td>Te Deum in C major</td>
<td>2–9</td>
</tr>
<tr>
<td>Aphex Twin</td>
<td>Digeridoo</td>
<td>1–5</td>
</tr>
</tbody>
</table>
Apocalyptica Coma 6–9
Arvo Pärt Spiegel im Spiegel for Violin and Piano 2–5
Beastie Boys Intergalactic 6–9
Beck Everybody’s Got to Learn Sometime 6–9
Bing Crosby White Christmas 1–5, 10
Bruce Springsteen Born in the U.S.A. 6–9
Buena Vista Social Club Candela 6–9
Camille Saint-Saëns The Carnival of the Animals: XIV. Finale 4–5
Cheer Chen Groupies 10
Ceoltóirí Chualann Marcshlua Uí Néill 2–5
Ceoltóirí Chualann and Seán Ó Sé Marbhna Luimnigh 2–5
Clint Mansell Lux Aeterna (Theme from Requiem for a Dream) 2–5
Commodores Easy 1–5
Damien Rice The Blower’s Daughter 10
Edward Elgar Variations on an Original Theme for Orchestra (“Enigma”): Variation IX. (“Nimrod”) 2–9
Elvis Presley In the Ghetto 6–9
George Frideric Handel Solomon: Act Three, Sinfonia (“Arrival of the Queen of Sheba”) 2–5
Gioachino Rossini William Tell Overture 1–5
Gráinne Hambly Eleanor Plunkett 2–5
Johann Sebastian Bach Cello Suite No. 1 in G major: I. Prelude 2–5
John McSherry An Bhean Chaoiante 2–5
John Williams Theme from Schindler’s List 2–5
Jolin Tsai Play 10
Journey Don’t Stop Believin’ 1–5, 10
Juan Luis Guerra A Pedir Su Mano 2–10
Louis Armstrong What a Wonderful World 1–9
Max Bruch Kol Nidrei, Op. 47 2–5
Mazzy Star Into Dust 2–5, 10
Minnie Riperton Reasons 1–5
Modest Mussorgsky Night on Bald Mountain 2–10
Neil Young Only Love Can Break Your Heart 1–5
Nina Simone I Get Along Without You Very Well (Except Sometimes) 1–9
Nirvana Smells Like Teen Spirit 1–10
Orestis Karamanlis Atención Gringo - Remix 6–9
In addition to a number of common measures across experiments, a number of the musical excerpts were common to the excerpt pools used in all iterations. These common excerpts are given in Table 3.10.

3.5.3 Stimuli Selection

The stimuli (Table 3.8) used in EiM were selected with the specific goal of eliciting emotional responses within participants. In order to achieve this goal, discussions were held with a variety of musical experts (e.g., per-
Table 3.9: Media pool size and experimental stimuli count per iteration.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Size of Media Pool</th>
<th>Stimuli Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>46</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.10: Common excerpts from all EiM locations.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Black Eyed Peas</td>
<td>I Gotta Feeling</td>
</tr>
<tr>
<td>Louis Armstrong</td>
<td>What a Wonderful World</td>
</tr>
<tr>
<td>Slayer</td>
<td>Raining Blood</td>
</tr>
<tr>
<td>U2</td>
<td>One</td>
</tr>
<tr>
<td>The Beach Boys</td>
<td>Good Vibrations</td>
</tr>
<tr>
<td>Nina Simone</td>
<td>I Get Along Without You Very Well (Except Sometimes)</td>
</tr>
<tr>
<td>Nirvana</td>
<td>Smells Like Teen Spirit</td>
</tr>
<tr>
<td>Sinéad O'Connor</td>
<td>Nothing Compares 2 U</td>
</tr>
<tr>
<td>The Verve</td>
<td>Bittersweet Symphony</td>
</tr>
</tbody>
</table>

forming musicians, composers, and musicologists) and their feedback was incorporated into the selection process. Musical stimuli used in previous studies were also included. In selecting musical stimuli for the elicitation of emotional responses, target emotion areas were guided by Russell’s circumplex model of affect \[144\], rather than basic or discrete emotions, and participant questionnaires were designed accordingly (Section 3.5.1.4).
In particular, selections about which experts agreed were representative of the four quadrants of the circumplex model were selected for inclusion, with the aim of creating equal distributions of selections across these quadrants. Preliminary affect labels of happy (high-valence/high-arousal), sad (low-valence/low-arousal), tense (low-valence/high-arousal), and relaxed (high-valence/low-arousal) were assigned to each stimulus.

In addition to a variety of affective trajectories across the musical selections, two other properties of music played an important role in the selection process: musical genre and lyrical content. First, it was important to select stimuli across a variety of musical genres (e.g., classical, pop, rock, folk). Second, to address potential differences in affective response between stimuli that contain lyrical content and those that do not, stimuli in both categories were included in the stimuli pools (and in some cases, pairs of the same musical excerpt—one with lyrical content and one without).

For each musical selection, an approximately ninety-second excerpt most representative of the target affect quadrant was selected. As necessary, half-second fade-in or fade-out transitions were added to the beginning or end of each excerpt, respectively. Finally, each excerpt was preceded by a period of two seconds of silence. All stimuli audio files were normalized to the same level of perceptual loudness. The stimuli selection and preparation process is detailed in [76], and the same methods have been employed in the iterations of EiM used for this dissertation work (Iterations 7–12).

3.6 SUMMARY

This chapter has provided an overview of the EiM study itself across the various iterations it has seen. The EiM study has been developed iteratively over the last nine years. In most of these iterations, changes were made to the study design and apparatus, the selection of affective stimuli, and to the data collection methodology. This chapter has provided the details of these changes throughout these iterations, as well an account of the data collected during each iteration. As stated previously, the EiM study began before the work described in this dissertation. The work for this dissertation began with the seventh iteration (Bergen, Norway), and continued through the twelfth iteration. During the work performed for this dissertation, specific changes to the study itself have been identified and implemented, and the experience with EiM gathered from the commencement of the study until now has been incorporated in developing and refining a set of methodologies for large-scale studies of affect that are useful for anyone beginning similar efforts. The next chapter will elaborate these changes to EiM and present the flexible methodologies that have been developed.
The original Emotion in Motion (EiM) study provided a wealth of data from which to build new knowledge about human affective response to musical stimuli. Chapter 3 provided an overview of the study design, apparatus, methods, procedures, and staging of EiM. During the initial implementations of the study, it became clear that the original study apparatus did not lend itself well to the rapid design and deployment of smaller “substudies” aimed at exploring more focused areas of inquiry with smaller groups of participants. In addition, it was inflexible in terms of the types of media stimuli that were presented, how data were persisted (structured and stored), the types of settings in which the experiment could be conducted with relative ease, the ease of remote management of the study, the types of measures that were taken, and the different physiological signals that could be recorded. These inflexibilities prompted reconsideration of the overall EiM study as a system for large-scale, distributed studies of psychophysiological response. This redesign process resulted in not only a tool for designing and executing studies of music and emotion, but also in an adaptable system suitable for a number of different research applications. In fact, what has emerged is an entire system for the rapid development of a range of not only psychophysiological studies but also for human computer interactions capable of capturing data from a range of sensors (physiological or otherwise); recording and responding to participants’ active input through a web interface, media presentation, or other real-time feedback; and more.

While there are no other systems available for the design, execution, and dissemination of data from these particular kinds of large-scale studies, the first section of this chapter begins by considering approaches taken by other researchers for similar efforts. Various approaches are considered with respect to their experimental design; uses of psychophysiological signals; study procedures; and analysis, storage, and dissemination of data. The chapter then moves to the enumeration of the specific issues with the original EiM system that prompted the exploration of a more flexible system for designing and executing studies like EiM. Informed by these issues, the following sections discuss the strategies that have been developed for hardware agnostic sensor input; physiological signal processing; and, synthesizing, representing, and disseminating multimodal data collected through the system. In addition, there are a number of
auxiliary tools and resources that this process has produced that will be presented. Finally, the chapter concludes with a discussion of the larger contribution that this system and the EiM database represent.

4.1 SIMILAR SYSTEMS

A number of study approaches similar to EiM have been explored by the research community for several years. While many of these deal more generally with biomedical physiological signals, as differentiated from psychophysiological signals, these resources have been both instructive in informing the choices made during the development of EiM as well as useful in highlighting the value of EiM as a contribution to the research community. As this dissertation argues for the importance of EiM as a holistic approach to study design and procedures; the use of psychophysiological signals in such a study; and the analysis, storage, and dissemination of these data; this section considers how these other approaches have dealt with these issues.

4.1.1 Research Resource for Complex Physiologic Signals

The Research Resource for Complex Physiologic Signals is a well-known resource that is similar to the EiM system in a number of ways. It provides several different resources: Physiobank, a collection of databases of physiology recorded from a number of different studies, along with other associated data from those studies; PhysioToolkit, a repository of various software tools for physiological signal processing and analysis; and PhysioNet, a site facilitating the communication and collaboration of researchers working with data like these [54].

While PhysioBank provides a large (and growing) resource of physiological signals, the majority of these signal databases do not speak directly to experimental design or study procedures, except through publications attached to the database archive. Often, study design and procedures are documented in one or more publications, but these details are not integrated directly within the database structure, and the depth with which study designs and procedures are compared and contrasted across various databases is cursory, at best.

In addition to this, most PhysioBank databases do not include psychophysiological signals, per se. While some of these databases do include signals useful in psychophysiological applications, circumstances limit their usefulness in these settings. For instance, while the MGH/MF Waveform Database includes recordings of a number of cardiovascular and

1 https://physionet.org/
respiratory signals useful in affective science, all of the signals it contains were recorded from hospital patients and include only relevant clinical data and event annotations [176].² PhysioBank databases are useful to the affective scientist, however, in that they provide a wealth of high-quality prototypical signals that might be used, for example, in the development of signal processing algorithms.

4.1.2 MAHNOB-HCI Database

The MAHNOB-HCI Database draws closer to the database provided through the EiM system, but is still “only” a database [162].³ This database was constructed as the result of two separate studies. In the first, participants watched “emotional” videos and provided ratings of their emotional response via self-report and emotional keywords. In the second, videos and images were shown with and without emotion tags (which may or may not have been correct). When a tag was shown, participants responded with their agreement or disagreement with the displayed tag. In both studies, audio and video of the participants were recorded, as well as eye gaze information and a number of physiological signals.

As noted above, while the database is a valuable resource to the affective sciences community, what is provided is little more than the database itself. In other words, the MAHNOB-HCI database does not provide a platform upon which to extend the study in further work, to design and execute other similar studies, or to correlate iterations on study procedures with resulting data. In general, this lack of extensibility is a common theme amongst all of the other approaches considered here.

Where the MAHNOB-HCI database does shine is through the physiological and other signals it makes available to the research community. These tightly synchronized signals include audio, video (from six angles), 32-channel electroencephalogram (EEG), electrocardiogram (ECG), electrodermal activity (EDA), respiration amplitude, and skin temperature. In addition, a large amount of metadata (e.g., whether or not the participant wore glasses, felt emotion ratings) are included for each participant. Notably, facial feature data are missing from the database. Nevertheless, these data represent a rather thorough “picture” of participants in different emotional states. Given the complexity of the experimental apparatus, it is understandable that the sample size is quite small (n = 30).

This small sample size and number of studies did not present the researchers constructing the MAHNOB-HCI database with the same obstacles present in disseminating a database of the size and scope of EiM.

² https://physionet.org/physiobank/database/mghdb/
³ https://mahnob-db.eu/hci-tagging/
The two studies from which the MAHNOB-HCI data were drawn were largely disconnected—in other words, the problem of representing and communicating ongoing iterative study design was not a consideration in the dissemination of this database. In particular, relationships between participants, study designs, and results are not modeled—and are altogether unnecessary—in this database; participant data are largely presented as individual, unconnected entities via collections of flat files. Irrespective of this, the means by which interested researchers can access these data are quite intuitive and usable. Facilities for nuanced or complex searches of the database are not made available through the web interface, however.

4.1.3 **MMI Facial Expression Database**

The MMI Facial Expression Database was built in an effort to address the need for a comprehensive set of facial imagery covering a range of affective and expressive states that could be used for benchmarking systems for machine analysis of facial expressions [130].

Nineteen subjects displayed 79 series of facial expressions. Each of these series of expressions included one or more action units. These displays were captured as still images or on video, and for many of the images or videos, these are provided as a dual-view that combines both a frontal and profile view of the participant’s face.

The MMI database is unique among other databases here in that the authors make clear that considerations of the kinds of samples and metadata to include in order to meet the needs of other scientists, of the methods to employ to elicit such samples, and of facilitating efficient and secure retrieval of data by other scientists were explicitly stated before mounting the effort to build the database. However, while these considerations did directly inform their own experimental design and procedures, the database in and of itself is not built to aid in other scientists’ designs and procedures, per se. It is nevertheless important to note the difference in the resulting database when these concerns were considered in the design of the initial study itself, as opposed to when databases are made available to others after a study is completed, and such concerns are merely an afterthought. In the case of the MMI database, the result is in fact a database that is easily accessible and searchable, and that represents a comprehensive sample.

Autonomic psychophysiological data are not included in the MMI database. However, affective data in the form of either audio or video of the participant’s face during the session, Facial Action Coding System (FACS) annotations, Onset-Apex-Offset FACS annotations, and emotion tags for

4 https://mmifacedb.eu/
the session are included. These metadata and annotations are provided alongside other files made available for the session.

Taking note of the fact that in other databases data are often provided to users in the form of large, flat, unstructured files, a stated specific aim of the the MMI database is to provide data that are easily searchable and accessible. To address this issue, the authors provide a web interface to their database that permits quite flexible and nuanced search queries and provides a means of downloading subsets of the database and individual data files. Most data files are Extensible Markup Language (XML) documents, a format that is both machine-readable and human-readable, and allows for the representation of relationships between data files through the file structure and contents.\(^5\) While the authors did describe extensibility of the database by other scientists to be a specific goal of the MMI database, this seems to only be possible for administrators of the database, however.

4.1.4 DEAP Database

One of the first databases to provide an extensive collection of video and psychophysiological signals was the DEAP database [97].\(^6\) To create this database, the authors staged an experiment in which participants watched a selection of forty one-minute segments of music videos. Participants rated each segment in terms of Russell’s circumplex model of emotion (employing valence, arousal, and dominance) using the self-assessment manikin[17]. Throughout the experiment, an EEG, EDA, blood volume pressure, respiration pattern, skin temperature, an electromyogram (EMG), and an electrooculogram (EOG) were recorded. In addition, frontal video of twenty-two participants was captured. These data were made publicly available following the study.

As with most other databases described here, the DEAP Database does not itself provide a tool for extension of the study or for the creation of similar study designs and procedures. The DEAP Database does stand out as a valuable contribution to the research community as one of the first databases to provide an extensive suite of psychophysiological stimuli. At the time it was published, no other publicly-available databases had provided a collection of psychophysiological data of its breadth, albeit one taken from a rather small sample size. In addition, the DEAP Database is particularly useful in that not only are raw data made available to the research community, but processed, cleaned, filtered, and segmented psychophysiological data are provided, as well.

---

\(^5\) https://www.w3.org/TR/REC-xml/#sec-origin-goals
\(^6\) https://www.eecs.qmul.ac.uk/mmv/datasets/deap/
One respect in which the DEAP Database falls short is in its structure and storage of data, producing a means of data dissemination and use in ongoing work that is unnecessarily burdensome. Here, data files are provided as a fragmented collection of flat files, although different file formats are provided. There is no database structure built around these data to impose a meaningful set of relationships among various entities in the database (e.g., participants, stimuli). In addition, the data are not queryable through any sort of interface, but are only made available through monolithic chunks of data (e.g., all participant psychophysiological data or all participant video data).

4.1.5 Summary of Similar Systems

While a number of contributions similar to EiM have been made to the affective sciences research community in recent years, EiM represents a wholly different sort of contribution in a number of important respects, which this chapter describes in detail. Most importantly, EiM is a comprehensive system for the design, procedures, and methods of studies of affect; the execution of those studies; the integration of psychophysiology into such studies; and the analysis, storage, and dissemination of resulting data. In other words, while extremely valuable as a database alone, EiM goes beyond providing a mere database to the community, and provides a means for other researchers to not only replicate and extend the results from this dissertation, but also to quickly design and deploy other kinds of studies using a system built through iteration and on the lessons learned while envisioning, gathering the data for, constructing, and disseminating the large-scale EiM database.

4.2 Issues with the Original EiM System

The original EiM apparatus and study design was only intended to be used for the iterations staged in Dublin and New York City. The possibility of a long-running study that would go on to collect data from tens of thousands of participants in numerous locations was never considered. Changes to study design in preparation for the first iterations after New York City, as well as the new demands of staging EiM in locations where study personnel would not be present for the entire duration, revealed a number of existing issues with the original apparatus and system for the synthesis and dissemination of the data collected through EiM. (In fact, wide dissemination of this database was never a goal of the original effort, and thus ease of dissemination was not considered in the design.) In addition, analysis of the data collected during the first six iterations
also revealed a number of ways that the specific study designs could be improved.

4.2.1 Issues with the Study Apparatus

In the first nine iterations of EiM, the study design was loosely structured around the following stages:

- Collection of participant demographic and musical background information
- Presentation of musical excerpts, each followed by a series of questions designed to gauge the participant’s affective response to the excerpts, as well as liking and familiarity, among other questions
- Collection of responses to questions about the overall experience (e.g., “Which song did you enjoy the most?” and “With which song were you most engaged?”)

Originally, the study was designed as a Max/MSP patch that coordinated the playback of musical excerpts, questionnaire presentation and response processing, signal acquisition, and data persistence. The system was packaged with a companion Max/MSP patch that allowed for the reordering of segments in the study design, as well as insertion of new segments. There were a number of issues with this initial apparatus.

First, the original study software architecture was not particularly well-suited for this use case. For EiM, Max/MSP was effective for the organization and presentation of audio stimuli, which was one of the tasks of the EiM apparatus. However, the primary issue in the original apparatus design was a rigidity in the hardware and software configuration, study design itself, and study data products.

While the same hardware (including everything from the study terminals and peripherals to microcontroller boards and sensors) are still used in current iterations of the study, the original configuration of hardware was rigidly bound to the Max/MSP implementation. As a result of this, altering any component of the physiological signal chain, from sensor to storage, necessitated reconfigurations throughout the software. As a concrete example, a change as minor as increasing the sampling rate of a given sensor would require modifications to the Max/MSP objects for polling the serial port, the microcontroller configuration and software, and possibly even the data persistence mechanisms.

The software that presented the study to the user was equally rigidly designed, and wholly dependent on Max/MSP. The primary effect of  

7 https://www.cycling74.com
this was to tie a given iteration to an inflexible study design (e.g., what questions would be asked and the order in which they would be asked, what excerpts might be selected for presentation to the participant, and what languages were available in which text would be presented). In order to make changes to the study design, new Max/MSP components needed to be written. As a trivial but demonstrative example, in order to make the study available in Norwegian, each “screen” of the study had to be individually duplicated in Norwegian. Then, the routing logic within the study application had to be amended to allow the participant to select Norwegian as their language of choice. In effect, the apparatus had to be duplicated in its entirety—once in English and again in Norwegian—with entirely separate paths through the study in each language. Little behavior in the software was abstracted in such a way as to make it possible to easily modify participant language, study design, or even aspects as trivial as visual layout.

The data that was produced as a result of one’s participation in the study presented problems as well. First, all data produced by the Max/MSP application were saved in comma-separated values (CSV) flat files or files in a proprietary format. The format of these files was problematic, as from one study iteration to the next, it did not always remain consistent (e.g., different columns and column orderings). Furthermore, because study designs differed from one iteration to the next, a participant in one iteration may have had a completely different number and formatting of data files than those from a participant in another iteration, even though a large subset of the study design between two iterations remained identical. With data files generated in this way, comparing data across iterations required a great deal of work (see Section 4.2.2).

**Impacts of these issues** These issues negatively impacted the original iterations of the study in several costly ways. To make changes, it was necessary for one of the experimenters to manually update each study terminal in order to mitigate any issues that may have arisen during the update process. In addition, the study software was originally designed such that only one study design at a time could be run on a given terminal—simply changing the wording of a questionnaire question was a non-trivial process. Furthermore, the lack of modularity in both the hardware and software meant that the initial installation and any repairs had to be performed directly by one of the experimenters. As the study was installed in additional and remote venues, these issues became very costly, requiring (often) intercontinental travel in order to setup or tear down a new installation.
4.2.2 Issues with Data Synthesis, Management, and Dissemination

As EiM grew in size and scope, performing analyses that included data drawn from different iterations of the study became a more pressing and tedious task. In addition, as the database itself grew to be a more valuable resource to the affective sciences community, providing access to the data to other researchers presented additional obstacles.

To begin with, one of the primary disadvantages of storing study data (especially the amounts of data produced by EiM) in flat files is the difficulty of overlaying structure on and communicating relationships between these files. Crude strategies for organizing flat files can be devised—for instance, in the example of EiM, by embedding dates and iterations into the flat file directory structure—but these make other, even simple, operations difficult. For example, the process of extracting data for all female participants in the original database, even only those that participated in a single iteration, was onerous. Furthermore, individual participant data communicated no information about the study design itself, which is desirable, for example, in integrating the data from participants from iterations with different study designs into the same analysis. As another example, two iterations may have followed identical study designs with the exception of using different pools of musical excerpts for stimuli, though these pools may have had some excerpts in common (Table 3.10). Identifying those musical excerpts that were or were not common between the two iterations was cumbersome without any higher-level structure.

Inconsistencies in data formatting, as mentioned in Section 4.2.1, further complicated matters. Consider the two file fragments shown in Listings 4.1 and 4.2. These fragments came from iterations that used the same two study designs, yet they are formatted entirely differently. This creates considerable problems in synthesizing and representing EiM data in a consistent and reliable manner.

Listing 4.1: Excerpt of example participant answers file from iteration 5.

```
01_Sex Male
01_DOB 1960
...
Song1_R006
Song1_Scale1_Engagement 5
Song1_Scale2_Positivity 3
Song1_Scale3_Activity 4
...
FinalQ_Most_Engaged 2
FinalQ_Most_Ejoyed 2
```

Listing 4.2: Excerpt of example participant answers file from iteration 9.

```
"01b_DOB" , 1970 ;
"02_Nationality" , symbol Filipino ;
...
"Song1_Scale1_Engagement" , 3 ;
"Song1_Scale2_Positivity" , 4 ;
"Song1_Scale3_Activity" , 4 ;
...
FinalQ_Most_Engaged , 1 ;
FinalQ_Most_Ejoyed , 1 ;
```
Even if data file formatting had been consistent across iterations, the heterogeneous nature of data within the files introduced additional difficulties. Specifically, for a given participant, there were numerous questionnaire responses as well as physiological signal recordings that needed to be persisted and shared amongst researchers. Often in a study like EiM, when a single type of physiological signal is recorded and saved to a database, either a special-purpose software is used for persistence and analysis, or individual recordings are saved as flat files with other data (e.g., annotations, questionnaire responses, and schemas) saved alongside these as separate flat files. Not only does this additional horizontal structure aggravate difficulties in overlaying higher-order structure on the database (e.g., data pertinent to entire iterations or subsets of iterations), it has non-trivial implications for data analysis itself. For example, if all EiM questionnaire data from participants had been coalesced into a single file (in order to make analysis more efficient), it would have been necessary to create pointers to all external files associated with an individual participant (and, such an approach also introduces the problem of maintaining the integrity of these pointers). Furthermore, while a single file containing all participant questionnaire responses may make the analysis of questionnaire data more efficient, access to and analysis of individual signal files as necessary would remain inefficient.

The size of the database itself and, in particular, the ongoing nature of the study also dramatically problematized the process of disseminating these data. At present, the EiM database is on the order of hundreds of gigabytes. Though this may not qualify as “big data”, sharing a database of this size with other researchers is difficult, particularly as the study is still in progress. Furthermore, it is possible and desirable for partner researchers to contribute their own products back to the database (these products may include derived features from signals, data annotations, and the like—Requirement 11, Section 4.3.2.3). Thus, periodically transferring hundreds of gigabytes of data to partner researchers was a wholly ineffective system for dissemination of this large and growing database.

4.2.3 Specific Issues with and Changes to the EiM Study Design

Specific changes in the study design between the iterations that were staged in Manila and Taipei City were driven by reflection on and analysis of the data collected up through Manila.

To begin with, in all iterations prior to the iterations staged in Taiwan, the local population spoke (or at least could speak) English. A poll taken in Taiwan in 2002, however, showed that only about 1% of respondents

---

8 For example, this software for use with ECG data: http://www.ecg-soft.com/.
considered themselves to be fluent English speakers [134]. This result, coupled with experiences working with Taiwanese venue administrators ahead of staging EiM there and experiences staging EiM in Taipei City and Taichung City, made it clear that the media pool that had been in use since the sixth iteration needed modification. These modifications included the addition of excerpts of a number of songs that were currently popular in Taiwan, as well as modifications to existing excerpts to explore research questions that previous iterations had delineated. A number of excerpts from previous iterations were retained, as well. Details of the media pool used in Taiwan are provided in Section 3.8.

Next, in 2012, Boucsein [16] surveyed a number of studies of the possible dependence of treatment recordings on baseline recordings of EDA. Although he found sufficient evidence to question the validity of the law of initial value [178] as it applies to EDA, he did also find that possible baseline dependencies cannot be ignored when “response scores must be calculated in the form of differences to baseline scores” [16, p. 239]. In the iterations prior to those staged in Taiwan, no baseline physiology was measured from participants. Thus, beginning in Taiwan, a “baseline excerpt” (an excerpt of pink noise—noise with equal amounts of energy per octave) was presented to each participant prior to the presentation of experimental stimuli.

Also problematic in previous iterations of the study were the number of experimental stimuli that were presented. In particular, ordering effects between stimuli have been observed, as well as habituation to the stimuli [29], [80]. In other words, the order in which stimuli are presented and the number of stimuli may have an effect on the participant’s affective response. To account for this, the number of experimental stimuli in the iterations in Taiwan was reduced to one, preceded only by the baseline excerpt.

Finally, numerous researchers have found links between personality and physiological response. Of particular interest was the evidence suggesting a relationship between one’s personality and EDA lability [32], [40]. Crider noted that “[g]reater EDR lability is associated with a relatively undemonstrative and agreeable disposition, whereas greater EDR stability is associated with a relatively expressive and antagonistic disposition” [32, p. 141]. Numerous other studies also suggest a relationship between personality and psychophysiological response [121], [159], [179]. Thus, it was important to at least have some measure of personality traits available for integration into the data analyses. To this end, the use of a ten-item short version of the Big Five inventory, a commonly used inventory for the measurement of personality traits [10], [85], [86], was introduced.
4.3 REQUIREMENTS FOR A FLEXIBLE SYSTEM

Not only did these issues need to be addressed, but what does it mean for a system to be “flexible”? Looking back to Schuller’s call [150], there is a pressing need in the affective sciences community for large samples of multimodal data from numerous populations that are collected in as naturalistic environments as possible. It is thus immediately obvious that, as mentioned in Section 1.4, the means by which data are collected should be constructed in such a way as to permit collection in multiple locations, rapid on-the-fly iteration of study design, and the investigation of simultaneous (possibly divergent) lines of research inquiry. Concretely, and driven specifically by the issues identified in Sections 4.2.1 and 4.2.2, a flexible system for the collection, synthesis, management, and dissemination of data of this type and magnitude should satisfy the requirements discussed in the following sections.

4.3.1 Requirements for the Data Collection System

4.3.1.1 Requirement 1: Multiple Stimuli Types

Previous iterations of EiM were restricted to only musical stimuli. A system flexible enough to provide the diversity of data suggested by Schuller should permit multiple types of stimuli. When designing psychophysiological studies that use music as a stimulus, it is important to consider the possibility that the range and quality of affective responses that music is capable of eliciting may be different than the range and quality of all possible affective responses [101], [140]. Because of this, a range of modalities of stimuli should ultimately be used, in general, in the study of human psychophysiological response. Thus, a flexible system should at least be extensible enough to permit stimuli beyond only auditory stimuli.

4.3.1.2 Requirement 2: Localization

While the initial iterations of EiM were made available, when necessary, in another language (Norwegian), the modifications required to add support for another language were extensive. In order to collect data from participants across language boundaries while maintaining validity, the study should be readily available in participants’ native languages. Adding available languages to the study—once the study materials have been translated, of course—should require only minimal effort, and the process by which a user switches between languages should be straightforward.
4.3.1.3 Requirement 3: Changes to Study Design/Procedures

In order to allow for the simultaneous investigation of multiple lines of research inquiry, the rigid study structure that existed in the initial iterations of EiM needed to change. In addition to allowing for different types of stimuli, changing the structure of the study design should be straightforward. It should be possible to reorder and modify study components (e.g., to reword a question or add a new questionnaire), to create new types of components (e.g., to add a new interface for participant interaction or response to a question), and to run multiple simultaneous study designs (e.g., for a given participant a study is randomly selected from several available studies).

4.3.1.4 Requirement 4: Remote Study Deployment

The costs involved with deploying EiM to remote locations were substantial, largely due to the need for a trained researcher to be available for initial setup and testing. To permit the long-term operation of this and other similar studies, any capable technician (e.g., an IT or audio/visual staff member at a museum) should be able to install the study apparatus and run operational tests.

4.3.1.5 Requirement 5: Virtual Study Deployment

It is unreasonable to expect any potential study participant to have all of the required sensors in their home to participate in a study of psychophysiology. However, the ability to deploy such studies virtually and allow participants to, at a minimum, participate in portions of a study at home, or in any other environment outside of the laboratory or other moderated study location, would support the collection of a set of data taken in genuinely naturalistic environments. Thus, a flexible study apparatus should at least allow for those portions of the study that a participant might engage with at home to be made available online.

4.3.1.6 Requirement 6: Remote Study Administration

It is imperative that such a flexible system allow for a study to be effectively monitored and administrated remotely. This includes the ability to monitor incoming data quality in (near) real-time, to test terminal operation, to monitor feedback provided by the study terminals (e.g., logging and reporting output), to make changes to the study, and to address problems with the apparatus, all from a remote location.
4.3.1.7 Requirement 7: Commodity Hardware in Study Apparatus

In practice, when a physical study apparatus is used by tens of thousands of participants, the normal cycle of wear and tear is accelerated. Using laboratory-grade equipment in a study in which terminals are deployed to remote locations and study hardware quickly wears can quickly become cost prohibitive. A flexible system that recognizes this reality must at least allow for the use of commodity hardware, including computer systems and peripherals, sensors, and other components of the physical study apparatus. It will be up to the individual researchers to ensure that, if selected, commodity hardware will meet their requirements for data quality. Nevertheless, an ideal system should allow for this option.

4.3.1.8 Requirement 8: Multiple Sensor Types

The original EiM study was designed and constructed to measure and record only two types of physiological signals: blood oxygen saturation (providing pulse and other derived features) and EDA. While these signals have provided a wealth of useful data, additional physiological signals would permit even more robust analyses. More importantly, additional physiological signals would make the EiM database and other similar databases a more valuable contribution to the greater affective sciences community. To this end, a flexible system should be sensor agnostic, allowing any sensor type to be used for data collection. This should be feasible in all cases where the raw sensor data is accessible.

4.3.2 Requirements for the Data Synthesis, Management, and Dissemination System

4.3.2.1 Requirement 9: The Use of Commodity Software for Dissemination

Section 4.4.2 discusses the lack of special-purpose software for synthesizing, managing, and disseminating data of the types discussed here. The broad range of disciplines and expertises represented in the affective sciences community suggests that a flexible approach to data dissemination should provide a means of accessing and using the data that does not require expertise specific to one discipline, preferably using general-purpose tools. Thus, this flexible system for the synthesis, management, and dissemination of these types of data should include a means for performing these tasks using general purpose software that not only remains flexible enough to meet new requirements that these tasks present and is usable by a broad range of users, but that also satisfies those other requirements specified in the next sections.
4.3.2.2 Requirement 10: Centralized, Online Data Repository

The exchange between interested researchers of multiple copies of folders containing data in flat files has already troubled the EiM analytical process. The most problematic part of this system was the fact that such copies of the data quickly became stale, not accurately reflecting the current state of the growing database. As multiple researchers require ongoing access to the database, the system for synthesis, management, and dissemination should satisfy three requirements. First, the canonical, authoritative database should exist in a centralized repository. Second, this repository should be accessible online to all interested parties through a selection of various tools or language libraries with which they are comfortable. Finally, while the canonical database will be maintained in a centralized repository, it should be possible and straightforward for researchers to capture all or subsets of the database for offline analysis.

4.3.2.3 Requirement 11: Contribution of Derived Data and Products of Analysis

Quite regularly in analyses of these data, derived features of the data were found that would prove useful to other researchers in their own analyses. For example, a transformation of a physiological signal that proved especially useful in predicting affect may have been discovered, or researchers may have augmented stimuli data with annotations that were useful in later analysis. Rather than restricting this derived data to exist only on the workstation of the investigator that created them or requiring other researchers to retrieve these manually, the synthesis, management, and dissemination system should allow for such data to be easily contributed back to the centralized repository itself. Furthermore, it should allow contributors of derived data to integrate such contributions meaningfully into the database itself (i.e., to specify relationships between new and existing data).

4.3.2.4 Requirement 12: Imposing Structure on Data

The lack of structure at many levels of the existing data generated by the first iterations of EiM complicated analysis tremendously. It was, for example, tedious to subset participants within a single iteration by a single demographic characteristic. As levels of desired structure became more complex—for example, the structure necessary to identify all female participants who took part in possibly different iterations; were all exposed to the same number of stimuli, and heard the same second stimulus; and participated before noon—the analytical process became even more difficult. A complex database permits and often requires complex analyses, and such complex analyses demand structure at many levels. Structure
of this complexity simply is not possible with even the most ingenious system of structuring flat files. Thus, and crucially, a flexible system should allow for complex structures to be communicated by the data itself, and thereafter leveraged in analysis. This, at a minimum, includes the ability to easily capture and express the commonalities and divergences between the studies in which individual participants participated.

### Requirement 13: Coalescing Heterogeneous Data

Beyond imposing structure on the data, special consideration must be given to the heterogeneity of the data, as well. Databases like EiM and those from similar research endeavors will likely contain questionnaire responses from participants (with possibly multiple sets of responses for multiple stimuli), time series data from physiological recordings from participants, continuously recorded feedback from participants (e.g., a continuous measure of affect recorded through some response device or interface), and the stimuli themselves (e.g., audio and video files, or pictures) to serve as references during analysis (or for direct use in the analysis as in music feature extraction, for example), among other things. Not only is imposing structure on these different forms of data important, it is equally important to represent and make available these heterogeneous forms of data in such a way that they are easily accessible and navigable. As a counterexample, providing a CSV file of participant responses embedded with references to the locations of associated stimuli and physiological signal recordings is not only a fragile strategy, but also one that makes collecting these data together during analysis quite inefficient. A flexible system for data synthesis and management should coalesce these heterogeneous data in a meaningful way. For instance, what may in actuality be a physical separation between a participant’s questionnaire responses and their physiological data recordings, should still be represented and accessible as a whole. Coupled with the necessary structure described in the previous requirement, the process of accessing all of the recordings of a certain physiological measurement that were elicited in response to a specific stimulus becomes much more efficient, for example.

### Requirement 14: Semantically Meaningful and Flexible Data Structures

At a more granular level, it is important that the structures used to store data be both semantically and structurally meaningful. At the same time, these structures must be flexible enough to allow for extension and modification required to represent data collected through further iterations of a study. For example, consider again the example file fragment given in Listing 4.1. Within this fragment, there is no structure beyond that in a single line, structured as `<Label> <Value>`. For a researcher interested in
using data in this format in their own work, barring the existence of yet another flat file describing the format of this file, the structure of this file is difficult to parse. What is the significance of the number at the beginning of each line? Should the lack of a value on the line beginning with 05_ be taken to mean that there was no response, or a negative response? Is there a way to know the number of stimuli presented other than iterating over all Song# lines? Finally, how are differences in study design communicated through the structure of data represented in this fashion? The lack of structure and semantics in this data representation inhibits the involvement and contributions of partner researchers. To this end, a flexible system for synthesis and dissemination should communicate in a meaningful way through structure and semantics, and such structure and semantics should be flexible enough to permit extension.

4.4 REALIZING THE FLEXIBLE SYSTEM

In preparation for the iteration EiM staged in Taipei City, Taiwan, a system that realizes this strategy and meets these requirements was architected and implemented. This work has taken shape as a software system that is, indeed, flexible enough to support not only EiM as it continues to be developed and staged, but also to support the research endeavors of other investigators looking to execute similar large-scale, possibly distributed, studies of affective response. This section describes this software system.

Figure 4.1: EiM system block diagram.
Responsibilities in the new system are more clearly defined and largely broken across three different components (Figure 4.1). These three primary components are signal capture and processing, a web application (the component of the software with which a study participant directly interacts), and the database, which plays a central role not only in data collection and dissemination, but also in specifying and driving individual participant interactions with the study. This section discusses each of these components, in turn, beginning at the physiological input from the participant, moving next to the database, and ending with the application that coordinates between these two components.

4.4.1 Signal Capture and Processing

Two of the above requirements apply directly to the signal capture and processing component, namely, that it be possible to use general-purpose, commodity hardware for physiological signal capture, and that it be possible to use any sensor for measurement of physiological signals.

4.4.1.1 Hardware-Agnostic Sensor Input

When working with physiological indicators of affect, the literature shows a range of signals that have been used (see [23] for a general overview, and [65], [101], [165] for music-specific overviews). These can be as varied as ECG, respiratory features, EEG, EDA, EMG, functional magnetic resonance imaging (fMRI), and pupillary response, to name but a few. Moreover, each feature can be obtained using different technologies and methodologies, which do not necessarily communicate using similar protocols and data structures. In addition, while low-level raw sensor data is at times made available by the sensor manufacturer, at other times it is not (for example, a digital communication protocol used to transmit sensor measurements over the wire may not be externally parsable).

In order to address this uncertainty, the EiM software system can receive and process sensor data from two different sources (Requirement 8, Section 4.3.1.8). First, when the analog, raw signal measurements are available through the sensor, this signal can be sent to the analog input of an Arduino microcontroller board. In the software system, software for the Arduino microcontroller board is available for capture of any analog input and labeling the captured data with the signal type. These recordings are made available to the rest of the system for association with the participant from whom they were recorded (Requirement 7, Section 4.3.1.7).

When raw analog signals are not directly available from the sensor, EiM can receive sensor data through Open Sound Control (OSC) messages.

http://opensoundcontrol.org/
Here, each received message can be time-stamped with the stimulus presentation/playback time, allowing for synchronization between the media stream and the sensor data, which is essential in the analysis of physiological signals that cannot be synchronized manually afterward recording [79]. Additionally, the experimenter can label the different channels of data used in the study, and these labels are embedded into the sensor data file. This design allows for heterogeneous signal sources to be incorporated into the system and recorded synchronously alongside the experiment stimulus.

### 4.4.1.2 Signal Processing Tools

One of the key challenges of working with physiological signals for real-time applications is being able to continuously assess the status of the signal (e.g. to check for artifacts or disconnections) and to extract significant features that can inform and drive the desired process/application.

Drawing from data available in the EiM database, two robust real-time signal processing tools for heart rate (HR) and EDA have been developed by other members of the Music, Sensors, and Emotion (MuSE) group. These tools were calibrated using the physiological signals of over 4,000 participants from the first iterations of EiM, learning from aspects such as user interaction with the sensors, variability of signals, and expected ranges for different demographics. The HR and EDA processing modules are distributed as signal processing modules with the EiM software system.

Each tool consists of a preprocessing stage, where the signal is formatted and analyzed for artifacts that can be caused by motion or connection issues. This analysis results in the production of a quality index, which represents the percentage of the signal that is free from artifacts. The signals are then processed to extract significant features, such as phasic and tonic components from EDA, and HR from ECG or pulse oximetry (POX) signals.

These tools are presented in detail in [77] and are available online.10

### 4.4.2 Database

The heterogeneity at multiple levels of the data in a database produced by a study like EiM makes finding a strategy for synthesizing and sharing these data difficult. (As an excellent example of this, one need only peruse the harmonization standards11 developed by the National Institute of Mental Health for their data archive.) Heterogeneity at multiple levels means that, at a high level, pieces of data in the database may represent

10 [http://www.musicsensorsemotion.com](http://www.musicsensorsemotion.com)

11 [https://ndar.nih.gov/standards.html](https://ndar.nih.gov/standards.html)
very different “things”—one file may gather together the responses given by a particular participant as an observation, while another file may be a sound file used as a stimulus; at a lower-level, two participant observation files, though they represent the same kind of “things”, may have divergent internal structures altogether. The difficulties that these differences in data create exist on both conceptual and practical levels, and it is because of them that many researchers choose to simplify data management and dissemination by packaging data into flat files, providing documents that describe the layouts of these files, and perhaps relying on a special-purpose tool and opaque data formats for handling those kinds of data that are especially difficult (physiological signals, for example). Representing data like these with most generally available and accessible tools is difficult, but it is not impossible. To motivate the tools that were chosen to meet the requirements of this system, this section first describes how data are usually structured using traditional general-purpose tools and the specific problems this poses for working with data that EiM and similar studies produce.

4.4.2.1 Traditional Approach

Much like in a spreadsheet application, in a traditional relational database management system (RDBMS) there are three main levels of structure with which both the designers and users of the database are concerned: tables and the rows and columns that combine to form these tables. Beginning at the outermost level, the tables in a well-designed database contain information about a collection of a specific type of data entity. For example, in a traditional e-commerce application, one table may hold information about a store’s customers while another holds information about that store’s products, and yet another the store’s orders. At the next structural level, a row in a table represents a specific instance of one of the entities that the table represents. Continuing the example begun above, a row in the customer table may hold information for one specific customer (e.g., their name and address). For the purposes of EiM, were a participant’s data to be stored in an RDBMS, a row may hold the demographic details for the subject of an experiment. The columns of a table, on the other hand, stand at the finest level of granularity in a relational database—specifically, the intersection of a row and column. In a customer table, one column may represent the street address of the customers while another contains postal codes. On the other hand, one column in a table of EiM participant observations may hold the participants’ genders, and the intersection of this column with a given row holds the gender for a specific participant.

This separation of concerns can provide the users of traditional databases with a high level of flexibility in designing a database schema that meets
with respect to the flexible data synthesis system proposed here, there are several issues with this approach, however.

**Differing structures between rows** One obvious mapping of raw qualitative experimental data to a database schema in designs such as those used in EiM is to have a table of trials. Each row in the trials table would contain, among other things, the responses to answers asked after each stimulus presentation. Across different iterations of EiM, however, the number, type, and content of these questions have all varied, as described in Section 3.5.1. A table in a traditional relational database requires that all rows in a table contain values for all columns in that table, even if some values are explicitly null. However, while it may be the case that null values are allowed for certain columns, in a well-designed, normalized database, such instances will be rare, if they are indeed present [28]. In a database the purpose of which is to amass data from studies of differing designs and various interactive scenarios, it is difficult, if not impossible, to design table structures that allow for both unified and normalized data representation, without resorting to a separate tables for the trials for each study iteration, for example.

As another example, it is meaningful in this database to store physiological signals together that were collected from participants. Grouping these signals in separate tables (e.g., by study or some other criterion) necessitates the collection and analysis of data from multiple sources (tables) when exploring broader questions of how people respond to different stimuli, for example. Furthermore, it is often the case that during the analysis of physiological data, meaningful features are extracted from physiological signals, and it is helpful to persist these features alongside the raw signals. The rigid table structure of traditional relational databases provides no clean way to do so.

**Differing structures within rows** The means by which a physiological signal is represented in a traditional database presents a separate challenge [84]. Given the above description of tables, rows, and columns, how does one represent a discrete time-sampled physiological signal (or an audio or other discrete time-sampled signal) in a traditional database? Three approaches to this problem are routinely employed. First, each column may represent a particular sample in the signal. This calls for as many columns as are necessary to represent the entirety of the signal that the table represents. While a large number of columns in a table is not a problem in and of itself, how does one store signals of differing lengths? Here, the problem of null values for columns at the end of all signals but the longest presents itself again. The second common solution is to store signals (which are, after all, nothing more than lists of numbers), as
data a single cell (i.e., the intersection of a table row and column). This approach provides for the possibility of representing, say, various signal types for a single subject in one row. Here, efficiency quickly becomes the problem. Analyzing any respectably sized set of signals, whether or not in real-time, translates to the process of agglomerating all such textual representations of signals together, converting them back to numeric representations (with the added risk of loss of precision), and only then performing the analysis—a very inefficient process. Efficiency problems, especially in large databases, plague the third approach, as well. Here, the data for each signal are stored in a file on disk, and entries in columns in a signal table simply point to those files. In a large database, the inefficiency here is clear, as well. Furthermore, new approaches for in-database analysis of large databases are continually becoming available, but most often these techniques require the data to live in the database, not in separate files referenced from within the database [41]. Taking any of these approaches may be acceptable in databases of a moderate size, but in large (and growing) databases such as the EiM database, these approaches are not feasible.

4.4.2.2 The EiM Approach

Partially in an effort to address these problems, NoSQL document-based stores have been developed (for example, Cassandra, CouchDB, MongoDB, and Redis) [107]. Instead of storing individual data entities in the rows of tables, a document-oriented database stores them within structured documents. Documents are often grouped into logical collections, and documents and arrays can be embedded inside of other documents. There are no predefined schemas for documents in a collection (though enforcement of structural “rules” can be performed at the application level, if required). This architecture allows for the straightforward creation of hierarchical relationships within and between documents, as well as the addition and removal of fields to documents with no implications for other documents within the same collection (Requirement 12, Section 4.3.2.4). Flexible schema and the ability to embed documents and arrays within documents exactly solve the problems discussed in Section 4.4.2.1.

The system described in this chapter uses the popular and freely available MongoDB database.12 In MongoDB, documents are stored in binary JSON (BSON)13 format, a binary serialization of JavaScript Object Notation (JSON)-like documents.

Like JSON, BSON supports the embedding of documents and arrays within other documents and arrays. BSON also contains

12 https://www.mongodb.com/
13 http://bsonspec.org
realizing the flexible system

| In the implementation of this flexible strategy for EiM, signal documents represent participant physiological signals, and trial documents represent all other participant data. For the data that have come from the EiM studies, a typical trial document looks something like that shown in Listing B.1. This example will also serve as a brief but useful introduction to JSON. Objects in a JSON document are surrounded by {}, and arrays (or lists) in a JSON document are surrounded by []. An object is simply an unordered list of key-value pairs. Here, for instance, the key _id is associated with the value ObjectId("570eb7bf83a73509d0e04f0f"). A key can also be associated with another object, allowing for objects to be nested inside other objects. For simplicity, when objects are nested, this discussion will use period-delimited names to explicitly refer to keys within this nesting. As an example, in this document, answers.personality.thorough is associated with NumberInt(3). Arrays/lists are simply ordered lists of values, and these values can themselves be objects (which themselves can contain further nested objects and arrays, and so on). Individual values in an array are referenced in this document by their index surrounded by square brackets, and these indices begin at 0 (the second value in an array lies at index 1). Concretely, in this document, answers.ratings.activity[1] maps to NumberInt(4).

Every document in the database has a unique identifier, given by its _id field. Wherever a field references an ObjectId("..."), this acts as a reference to another document in the database. Thus, using a document-oriented database does not preclude one from establishing relationships between documents in different collections. For example, the first media item (a stimulus) in this trial corresponds to the document in the media collection given in Listing 4.3. In this way, participant trial documents are easily associated with their related physiological signal recordings and stimuli, for instance.

Listing 4.3: Example media document.

{  
  "_id" : ObjectId("56ccbb59d4c6ac453fb18f02"),  
  "type" : "audio",  
  "artist" : "Radiohead",  
  "title" : "Paranoid Android",  
  "label" : "T019",  
  "has_lyrics" : true,

14 For a more thorough treatment of JSON, see https://www.json.org

4.4 REALIZING THE FLEXIBLE SYSTEM 71
Everything in an interaction driven by the system is stored as an object in the database. These may be signal and trial documents, stored media, or documents that describe the structure of a study itself. These structure documents are used by the web component of the system (Section 4.4.3) in order to dynamically construct interactions on demand (Section 4.4.3.1).

The combination of these relationships with the flexible structure of documents in the database allows the use of MongoDB to meet a number of the requirements outlined above. First, MongoDB is itself general-purpose, freely available software that allows access to the database through a number of tools and languages, the choice of which is left to the individual researcher (Requirement 9, Section 4.3.2.1). Second, MongoDB allows the entire database to be made continuously available online to all interested researchers—this database is now the canonical source for all data gathered through the various EiM studies (Requirement 10, Section 4.3.2.2). Finally, because of the relationships specified within the structure of various items in the database (as above where a trial references its associated physiological signals, the stimuli used in the study, and the study design itself, described in the object referenced by experiment), meaningful structure is built into both the design of the database itself as well as the individual documents it houses.

4.4.2.3 Remaining Requirements

To describe the ways in which this approach satisfies the remaining requirements with respect to data synthesis, management, and dissemination, one must first examine several interrelated database documents more closely. Listing B.1 shows an excerpted trial document (in these listings, “...” indicates locations where information has been removed for conciseness). In this study session, three stimuli were presented. This is made evident through the structure of the document; the values associated with both the media and signals keys are arrays that each contain three values. The first, second, and third values in each array correspond to the first, second, and third associated media documents (the stimuli) and signal documents (the physiological signal recordings), respectively. This is one example of many throughout this implementation of the strategy that

---

See https://eim-data.musicsensorsemotion.com/ for information on how to access the database.
demonstrates the cooperation of and communication of meaning through structure and semantics in these data documents, while still maintaining flexibility (Requirement 14, Section 4.3.2.6).

The signal document to which the first value in the signals array refers is given in Listing B.2. In this document, the value with which the trial key is associated refers back to the document in Listing B.1 (note the matching ObjectId("...") values). This document also contains the data_file, derived_eda_data_file, and derived_pox_data_file keys. The ObjectId values of each of these keys refer directly to documents in the database that hold the raw signal data, derived EDA data, and derived POX data, respectively.

Listings B.3 and B.4 show the metadata attached to the objects in the database that store the raw signal data and derived EDA data, respectively. Each of these metadata objects are also JSON documents (stored as BSON in the database) that describe the actual raw and derived sensor data. Those raw sensor data are stored as CSV files directly in the database, and thus each metadata object describes in detail the structure of these files. In this way, while CSV files are still used to store sensor data, these files are accessible directly through the queryable relationships described here. In addition, the metadata describing these physiological data files include high-level queryable information (e.g., the EDA decomposition parameters available in the decomposition_tau and decomposition_error names in the derived EDA metadata [9]).

This structure of relationships is what permits the coalescence of heterogeneous data in the database (Requirement 13, Section 4.3.2.5). All data that can be queried in the database are available as first-class documents in the database. For instance, relationships between a signal document are related back to even the investigator responsible for the study through the signal document’s related trial, its related experiment, and this experiment’s related researcher(s). Where it is more difficult to query data such as that stored in individual files (e.g., specific values within raw sensor data), these are nevertheless still stored in the database, and thus enjoy these rich and queryable relationships with other data in the database through their attached metadata documents. Additionally, if desired, signal values could be stored directly in an array on a signal document (for instance). This would introduce an obvious performance penalty when retrieving these documents, however.

Furthermore, this flexibility at the schema level allows for easy contribution of derived data generated either through ongoing analyses like those for this dissertation, or through the analyses of other researchers. It is rather trivial, for instance, for a researcher to append a new column of data to a derived signal data file, or to create a new derived data file object altogether, and associate it back to the data in the database from
which it was generated. In this way, the researchers working concurrently with the database can easily and quickly enjoy the contributions of other researchers (Requirement 11, Section 4.3.2.3).

4.4.3 Web Application

The web application with which study participants interact makes its own unique contributions in this implementation of the system toward meeting the requirements the strategy defines. In particular, the web application specifically enables localization, flexible study designs, remote and virtual deployment, remote administration, and data collection to meet the data synthesis and management requirements.

The application dynamically structures interactions, coordinates the presentation of media and acquisition of sensor signals, persists collected data, and provides a graphical display to those interacting with the application. This component is implemented in JavaScript, using AngularJS for the frontend, and Node.js for the backend. Structuring the web application itself around JSON documents, like those stored in the database, is of central importance in maintaining the flexibility specified by the requirements.

4.4.3.1 AngularJS Frontend

The frontend of EiM is a single-page application (SPA) written in AngularJS. While up to now the study has been hosted on terminals specifically for this purpose (Requirement 4, Section 4.3.1.4, and Requirement 5), this approach allows other studies built with it to be hosted on a remote server (Section 4.3.1.5). AngularJS is a JavaScript Model-View-Controller (MVC) framework that is used for web development and allows one to extend the syntax of Hypertext Markup Language (HTML); provides real-time, two-way data binding between data models and rendered views; and, handles complex dependency injection scenarios. In practice, this allows for any portion of a study built with the EiM system to be written as an AngularJS module and dropped into the system. Modules for such tasks as media playback, physiological data collection, questionnaires, and real-time continuous response inputs are included in the system, and it is trivial to write new modules to be included in the system (Requirement 1, Section 4.3.1.1). External modules are leveraged by the system, as well. Of note, the angular-translate module is incorporated into the EiM system in order to easily dump the entire textual contents of a study for translation, and load these translations back into the system for use in allowing users to select this new language for their trial (Requirement 2, Section 4.3.1.2).

16 https://angularjs.org/
A study design itself is described by a JSON document, as well. The main implication of specifying a study design in JSON is that this is the only new technology an interested researcher must know in order to build a new study using this system, refine an existing design used in a previous EiM iteration to meet their own purposes, or simply to work with the existing EIM database (Requirement 3, Section 4.3.1.3). To better understand how the design of a study is specified in a JSON document, consider Listing B.6. This listing is an excerpt of the study design document for EiM iteration 11 (deployed to Taipei City, Taiwan).

The first two names in the main object are the mediaPool and sensors arrays. The mediaPool array refers to media documents (like that shown in Listing 4.3). Including a reference to a media document here makes the media available to the system for use in a study. Next, the sensors array specifies that EDA and POX are to be recorded for this study. Other than the configuration described in Section 4.4.1.1, this is all that is required in the study design document to make media available for use in the study and to specify the physiological data that are to be recorded.

Next, the structure key in this study design document maps to an array of nested objects. Each object in this array represents a view (or screen) with which the participant will interact during the study. Each object in the structure array must provide a name, the value of which is a string (text label) that maps to a module available in the EiM system. Here, after the participant sees several introductory slides (indicated by the first …), they are presented with the sound-test and eda-instructions views. The modules for these two views contain the necessary functionality for playing test sounds and providing visual instructions for the participant in placing the EDA electrodes on their fingertips. It should be noted that numerous other modules such as these can be derived from the modules made available with the system. In particular, the process of customizing these modules that are already available in the system is trivial: a welcome screen, a consent form, general informational screens, questionnaires soliciting different types of responses, and media playback, among others.

Some objects in the structure array take additional values for configuration. The next object shown in the excerpt references the media-playback module. This module takes an additional mediaType parameter. When the value of the mediaType parameter is random, the module will playback a media excerpt for the participant drawn at random from the study’s media pool (that defined by the top-level mediaPool array). This module can also be configured to present a predetermined media stimulus (e.g., in EiM iterations 10–12, neutral pink noise is presented as a neutral, first stimulus). Most interesting here is the questionnaire module, which allows for the flexible creation of questionnaires of a number of different formats. Here
(though only one is shown in the excerpt), a range of rating scales are presented for the participant to provide their self-report ratings for the media that was just played.

Listing 4.4: Example questionnaire view.

```json
{
  "name" : "questionnaire",
  "data" : {
    "title" : "Musical Background",
    "structure" : [
      {
        "questionType" : "likert",
        "questionId" : "musicalExpertise",
        "questionLabel" : "How would you rate your musical expertise?",
        "questionLabelType" : "labelLeft",
        "questionLikertMinimumDescription" : "No expertise whatsoever",
        "questionLikertMaximumDescription" : "An expert",
        "questionStoragePath" : "data.answers.musical_expertise"
      },
      {
        "questionType" : "radio",
        "questionId" : "hearingImpairments",
        "questionLabel" : "Do you have any hearing impairments? (If so, you may still participate in the experiment!)",
        "questionOptions" : {
          "choices" : [
            {
              "label" : "Yes",
              "value" : true
            },
            {
              "label" : "No",
              "value" : false
            }
          ]
        },
        "questionStoragePath" : "data.answers.hearing_impairments"
      }
    ]
  }
}
```
While the server that hosts this system was placed on the EiM terminals in the study locations, as noted above, this server could be placed in a remote location, or a centralized location in the study location with which the study terminals communicate. Whether or not the terminals host their own server, remote access software allows for straightforward administration of the study (Requirement 6, Section 4.3.1.6). In particular, EiM used TeamViewer\textsuperscript{17} for this purpose. Other administrative actions, such as data backups, are scripted to occur automatically on the study terminals, and push their output to a centralized server.

### 4.4.3.2 Node.js Backend

Node.js is a platform built on Chrome’s V8 JavaScript engine for building fast, scalable network applications.\textsuperscript{18} Node.js uses an event-driven, non-blocking I/O model that makes it lightweight and efficient, making it perfect for data-intensive, real-time applications that run across distributed devices.

The primary roles of the EiM Node.js backend are the coordination of OSC communication for sensor control, recording of physiology, and media presentation; and communication with the MongoDB database. With the help of the Socket.IO library\textsuperscript{19}, any custom OSC message can be passed from the AngularJS frontend, through the Node.js backend, and on to a running instance of Max/MSP (as mentioned, Max/MSP is an excellent tool for tasks such as coordinating media playback, and thus is still used for this purpose in the system). In the same way, any custom OSC message can be sent from Max/MSP and arrive at the AngularJS frontend as a JSON object. Similarly, data of nearly any structure can be passed to and retrieved from the MongoDB database.

### 4.5 WIDER CONTRIBUTIONS OF THIS WORK

The value of the contributions of this particular area of this research are threefold. First, it has resulted in the revised EiM study itself, which continues to run daily and, in turn, has generated a valuable resource for the affective computing and larger affective sciences communities in the EiM database. Second, the realization of this strategy has taken shape as the new EiM software system. Not only is this system what now drives the ongoing EiM study, but it is also a tool that can be quickly and easily adapted to meet the needs of other researchers looking to mount similar studies, irrespective of scale. Third, this work has generated not only the

\textsuperscript{17} https://www.teamviewer.com/
\textsuperscript{18} https://nodejs.org/
\textsuperscript{19} https://socket.io/
requirements for a flexible strategy for collecting and disseminating the data that Schüller called for and that the affective sciences community so desperately needs [150], but also the knowledge that has been gathered in realizing such a strategy, and recommendations for researchers that will undertake similar ventures in the future (see Chapter 6).

4.5.1 Open Database and Supporting Software Libraries

As previous sections have mentioned, the EiM database is made freely available to researchers in the wider affective sciences community for use in their own research. The database continues to grow daily, and is already in use by a number of other researchers around the world. 20 Throughout the course of work for this dissertation, additional software libraries have been developed to ease the task of working with this very large database. In particular, there are libraries for psychophysiological signal analysis and for working directly with the EiM database itself. These libraries are now regularly utilized by others in research and commercial applications.

4.5.1.1 eim Python Package

The eim 21 Python library provides functionality specifically for working with the EiM database. The core functionality includes:

- Accessing and authenticating against the EiM database
- Querying and retrieving documents from the EiM database
- Preprocessing and analysis of psychophysiological signals (through the Pypsy Python package; Section 4.5.1.2)

The eim library is fully-documented. 22 For a thorough treatment of all of the functionality provided by the library, the reader is directed to this documentation.

4.5.1.2 Pypsy Python Package

The Pypsy 23 (Python psychophysiology) Python library houses the non-EiM-specific logic necessary for preprocessing and analysis of psychophysiological signals. At present, modules for the processing of POX and EDA signals (the signals recorded during EiM sessions) are provided. Notably,
the module for EDA signal processing provides a refined implementation of the primary algorithm used in Ledalab\textsuperscript{24}—perhaps the standard tool for EDA processing in MATLAB/academic work—correcting several fundamental errors in their currently-available release.

4.6 SUMMARY

This chapter began by identifying a range of issues with the original EiM study. In particular, a number of problems were identified around the study apparatus; the approach to data synthesis, management, and dissemination; and the study design itself. In response to these problems, fourteen requirements for a were derived, each aimed at addressing a specific issue identified in the beginning of the chapter. Taken together, these requirements represent a recommendation not only for the successful design and execution of large-scale, distributed studies of affect, but also for data synthesis and dissemination. Next, this chapter demonstrated a concrete implementation of a system that meets each of these requirements and has proven successful in a demanding, real-world study scenario. The data from this study, EiM, as well as several tools built to aid in the process of accessing and analyzing them, are now available to interested researchers.

\textsuperscript{24} http://www.ledalab.de/
5

STRONG PHYSIOLOGICAL RESPONSES

This dissertation has already introduced a new, flexible system for large-scale, distributed studies of affect. Furthermore, it has discussed how this system has been effectively used to establish a database of human affective response to musical stimuli that is unrivaled in its size and scope. This chapter serves to demonstrate the type of analysis that is only possible with systems and databases like this. Specifically, it describes the work of modeling strong physiological responses to musical stimuli in terms of the characteristics of a given participant. Both statistical and machine learning-based approaches are leveraged in this effort. This chapter will demonstrate that sufficient information is present in data about the characteristics of a participant in order to determine whether or not they will demonstrate the notable responses discussed here.

5.1 BACKGROUND

Several places where large numbers of participants had very strong psychophysiological reactions at very specific moments within some of the Emotion in Motion (EiM) musical stimuli have been interesting [76]. However, the precise causes of these reactions or what differentiates the participants that present such reactions from those that do not (not all participants present such reactions, though large numbers do) had not yet been identified. Knowing what produces such exceptional responses would be informative both from a musical and an interaction design perspective. This chapter explores such reactions first by considering what other researchers have learned about them. Some specific stimuli found in the EiM study that elicited these responses are then characterized. After this, opportunities for modeling these cases in the EiM data are explored, investigating possibilities for predicting such reactions.

5.1.1 What are Strong Reactions to Music?

Previous studies have explored "strong" reactions by considering them through a purely experiential lens, by considering physical/psychophysiological responses in isolation, and by studying feelings and emotions as they relate to these strong experiences. Gabrielsson provides an excellent

Considering multiple stimuli, about half of participants manifest the notable responses described in this chapter, similar to what other studies have observed [55, 58].
example and overview of the experiential side of strong reactions to music [50].

Gabrielssön describes the Strong Experiences with Music (SEM) project in [50], the purpose of which was to collect recounts of such (often very) strong experiences with music. Through qualitative analysis of these descriptions, Gabrielssön built an ontology of these types of experiences, breaking them down componentially. Specifically, he provides descriptions of not only these components, but also of how these components illuminate the causes of strong experiences with music. Importantly, among these components, participants showed that a physical reaction was often a very central element of strong experiences with music. Participants reported strong experiences across sixteen different categories of music (most categories corresponded with genres of music), including classical, religious, scenic, folk, jazz, rock, pop, dance, improvisation, and music from other cultures. The various studies under Gabrielssön’s SEM umbrella underscore two critical points. First, very strong experiences with music—at times, of the magnitude of out of body experiences—occur very infrequently. Second, the ways in which people experience music (specifically, experiences like those studied in SEM) vary widely. These experiences are a result of not only musical, but also personal and situational factors. Indeed, because of these factors the same person who has a strong experience with music may have heard the same music many times before or after their strong experience without such another such strong experience. Even when the experience is not specifically a strong experience, in Gabrielssön’s terms, these pronounced reactions to hearing certain music are determined largely by factors unique to the individual [149].

Far more common than explorations of such life-changing experiences with music are studies that examine music-listening experiences that evoke a strong physical reaction in the listener. These reactions may range from a racing heartbeat to tears, from shivers down the spine to goosebumps (piloerection). These last sorts of reactions, referred to by various researchers alternately as “chills”, “shivers”, “thrills”, and “frisson”, have been explored by a number of authors. As in [69], this section uses “frisson” to refer collectively to chills, shivers, and thrills.

In 1980, Goldstein documented a study, the hypothesis of which was that frisson in response to music may be mediated in some way by the release of endorphins [55]. To test this, he presented subjects with self-selected musical excerpts before and after naloxone injections (naloxone blocks the effects of opiates). During each listening, subjects indicated the locations in the music where they experienced frisson, the intensity of these reactions, and their physical localizations. Goldstein found that these sensations occurred most commonly along the upper spine and
the back of the neck, and found significant agreement among subjects that these sensations were pleasurable. Questionnaire responses from the study indicated that only about half of all respondents experienced frisson commonly, but for those that did, music was cited as the most reliable stimulus for eliciting frisson (according to 96% of respondents that experienced frisson commonly). Goldstein also noted that as the systems involved with electrodermal activity (EDA) are entirely innervated by the sympathetic nervous system, EDA—and by extension, thrills, which are marked by a strong EDA response—have a linear relationship with physiological arousal, if not also emotional arousal.

To test the significance of self-selected music for eliciting frisson, Rickard exposed subjects to four treatments—self-selected music that the subject indicated was effective in eliciting frisson (EPM treatment), relaxing music, arousing music, and an emotional film scene—with the hypothesis that higher incidences of frisson and increased physiological arousal would accompany the EPM treatment as compared to the other treatments. The study found some evidence in support of this hypothesis, noting however that while the mean number of occurrences of frisson was significantly different between EPM and both the relaxing music and emotional film scene, this difference was observed (but not significant) between EPM and arousing music. (Differences in all measures of emotional impact between EPM and all other treatments were significant, however.) EDA and occurrences of frisson were the only measures that demonstrated this difference, in contrast to measures of cortisol levels, skin temperature, electromyogram (EMG), and heart rate. This suggests that EDA and frisson are likely good physiological indicators of intense affective responses to music (others also argue that frisson is associated with prominent electrodermal responses (EDRs)—see [31], [60], for example). Additionally, Rickard noted that it was quite possible that other unmeasured variables introduced unwanted variance into their measures, identifying differences in participant personalities as a likely confound.

Drawing from Rickard [140], Mori and Iwanaga studied tears in addition to frisson as responses to musical stimuli [118]. They proposed not only that frisson would induce subjective pleasure and subjective/physiological arousal (as shown by Gabriëlsönn and Rickard), but that tears would induce subjective pleasure, relaxation, and physiological calming. Again, this study demonstrated a larger number of intense physical responses when self-selected musical stimuli were used, and that these responses accompanied higher heart rates, respiration depths, and EDA levels. They also found significant differences between frisson-inducing and tear-inducing music in musical features extracted through music information retrieval (MIR) techniques.
A study by Grewe et al. [58] also found further evidence of inter- and intrapersonal differences in frisson responses to music in exploring two questions: are listener characteristics important in determining whether or not frisson is experienced, and what musical/acoustical features can elicit strong emotional responses? Considering the latter first, the authors found that incidences of frisson are irregularly distributed across both participants and pieces of music. (Furthermore, the authors also found that one participant’s physiological responses to a given self-selected piece of music can vary from day-to-day—the same participant would identify frisson in sometimes consistent and at other times different locations in the music upon repeated auditions.) There was also evidence that participants who experienced frisson were less prone to seek adventure (this contrasts sharply with other literature [112], [153], both of which found that people who were more open to experience—one of the five personality traits measured in EIM—were more likely to experience frisson) and more reward dependent, and that music played a larger role in the lives of participants who experienced frisson. Sloboda also found that musicians are more prone to experiencing frisson [155]. Additionally, similar to [63], the authors found that only 55% of participants demonstrated frisson. Concerning the musical/acoustical features that can elicit strong emotional responses, they found the following to be especially effective: the beginning of a piece of music, the entry of a human or instrumental voice, changes in dynamics, important melodies/themes/motives, the contrast of two voices, and harmonies.

Grewe et al. later explored strong physiological responses to aural, visual, tactile, and gustatory stimuli, in addition to music [57]. Not only were they able to demonstrate similar physiological responses for all types of stimuli, but for musical stimuli they found that positively valenced music was more effective at inducing frisson, independent of the level of arousal participants perceived in the stimulus. Auditory and musical stimuli were both equally effective in inducing frisson, though there was less variance in response to musical stimuli. Again, a pronounced EDR accompanied all instances of frisson.

Many researchers other than Grewe et al. have considered what musical features are particularly effective in inducing frisson, though some authors consider this question in simpler terms of a strong emotional or psychophysiological response. In addition to those results from [58], researchers have found that frisson can be induced by changes in dynamics [129], [155], changes in texture and sustained high pitches [60], a solo instrumental voice emerging from an orchestral background [129], or averted or delayed fulfillment of musical expectations [69], [115], and that slower music is more effective in inducing frisson [60], though this list is not exhaustive. Clearly, the means by which music may elicit frisson are
as varied as the situational and personal factors that may have influence over whether or not a person experiences frisson in response to music.

In strong psychophysiological responses, an increase in physiological arousal is expected as the sympathetic nervous system is activated. Increases in heart rate, EDA, blood pressure, respiration rate, and muscle tension; decreased temperature; and increases in arousal hormones all provide evidence of activation of the sympathetic nervous system. However, while personal, situational, and musical factors may vary widely, one thing is consistent. In most of the cited and in other studies not mentioned here, these strong responses to musical stimuli are almost always accompanied by a sharp increase in EDA. This is the hallmark of the strong psychophysiological responses considered in the remainder of this chapter.

5.2 STRONG PSYCHOPHYSIOLOGICAL RESPONSES IN EiM

Though many important moments are visible in responses across the EiM database, three musical stimuli in particular have been very effective in reliably eliciting strong psychophysiological reactions. Jeff Buckley’s rendition of *Hallelujah* by Leonard Cohen, *Raining Blood* by Slayer, and *Into Dust* by Mazzy Star elicited the strongest physiological responses found in the database. The indications of these strong responses are prominent peaks in the aggregate EDA signals for these stimuli (the mean of all EDA signals across all participants for a given stimulus). For most other stimuli, the aggregate signals tend to regress to their central tendency, making prominent peaks in the aggregate signals fairly rare. The aggregate EDA signals for each of these songs, however, each display a prominent peak at a specific moment in time.

5.2.1 *Hallelujah*

One experimental stimulus that has been included in the battery of EiM stimuli since the beginning of the study is Jeff Buckley’s *Hallelujah*. Specifically, an excerpt from 5′00″.27 through 6′50″.43 has always been included in the available stimuli. In early visual inspection of EDA signals recorded in response to this experimental stimulus, a very regular and prominent EDR was evident toward the end of the excerpt (at 5′57″.5). The excerpt begins with the final verse of the song where the melody, unlike in previous verses, climbs into a higher register. Under the vocals a clean electric guitar repeats an arpeggiated line; though not a pure ostinato, its lulling effect is similar. Following the verse, two repetitions of the chorus follow, where both the vocals and the guitar settle to oscillate between the minor I and the major VI on the simple lyric *Hallelujah / Hallelujah*. These choruses
gradually decrescendo toward 5′57″.5. At this point, the vocals leap up a major ninth to the octave above that has yet to be used to this point in the excerpt; this leap is accompanied by a change to a forte dynamic.

The prominent EDR is shown in the top plot of Figure 5.1. The salient moment in the musical excerpt is marked with a vertical line labeled “Strong stimulus”. The plot shows the average EDA signal of 239 subjects that listened to this excerpt in the Dublin and Taiwan implementations. This, by itself, is an interesting finding as one can observe a similar response in the average phasic EDA signals from participants from two different geographical locations and cultural backgrounds. *Hallelujah* was the excerpt with the highest mean familiarity in the Dublin implementation (a mean of 4.1 on a five-point scale), whereas in Taiwan it was rated on average on the bottom half of the familiarity scale (a mean of 2.4 on a five-point scale). In order to further investigate this phenomenon, a different section of the same musical piece was added to the pool of excerpts for the Taiwan implementation. Specifically, the excerpt was taken from 0′45″.16 through 2′38″.25 and presents a musically similar structure as that of the original excerpt, with two verses followed by the chorus, but this excerpt included no leap in the vocal. In the signals gathered from this excerpt, there was no indication of a prominent EDR besides the initial reaction that was present in most of the averaged responses. This observation highlights the uniqueness of the response in the original excerpt. The bottom plot of Figure 5.1 separates the mean signal of the participants that listened to the original excerpt in Dublin and Taiwan that presented an EDR at this moment in the excerpt (*n* = 91) from those subjects who did not (*n* = 148). (The specific criteria by which a particular signal was labeled as a reaction or no-reaction signal are described in Section 5.3.1.) It is interesting to note that the group that did not exhibit a response has a similar mean EDA signal over most of the remainder of the excerpt to the mean EDA signal of the group that did exhibit a response.

5.2.2 *Raining Blood*

Slayer’s *Raining Blood* is remarkably different than *Hallelujah*. The excerpt used in *EiM* was taken from 3′01″.23 to 4′24″.69. The excerpt begins with the drums and rhythm guitar—the drums sound a standard driving rhythm on the hi-hat, bass, and snare, while the heavily distorted rhythm guitar plays sharp, repeated power chords on every beat. After a bar and a half, an overdriven lead guitar joins the fray with inverted sixteenth note arpeggios, beginning each bar on the tonic, stepping down the scale each beat, and returning to the tonic on each downbeat. This pulsing forward motion continues for six bars until the vocal enters with the lyric *Raining blood / From a lacerated sky / Bleeding its horror / Creating my structure now*
**Strong Physiological Responses**

Figure 5.1: Mean EDA signals from participants who listened to Jeff Buckley’s *Hallelujah* in Dublin, Taipei City, and Taichung City.

*I shall reign in blood.* As the vocal enters, the rhythm guitar continues its trudge forward, but the drums begin repeated sextuplets on the double pedal, adding cymbal crashes to the rhythm. Also increasing anticipation, the lead guitar sounds a similar figure, only now moving to power chords on the fifth and sixth strings. Up to this point (and throughout the rest of the excerpt), the bass matches the bass drum in rhythm, hammering the tonic all along the way. When this lyric finishes, the rhythm guitar is left alone sounding the tonic on the sixth string alone for two bars, relaxing tension only slightly. The hi-hat sounds out in the second half of the second bar, accelerating the tempo. On the next downbeat, an orgasm of sound explodes as the lead guitar sends notes whining up and down the fretboard. The rhythm guitar remains on the sixth string repeating a Locrian figure, only momentarily peeling away to the sixth to outline a major seventh chord with sharp fifth—a rather unexpected shift—and then returning to the previous figure. The drummer focuses his attention on the tom-toms, cymbals, and double bass drums, and as noted above, the bass hammers the tonic to the end of the excerpt. This continues for twelve bars, with an *accelerando* in the last two. As the band only briefly begins to slow, the music is interrupted by a load thunder clap. The echo of the thunder subsides to reveal the sound of heavy rainfall, which slowly fades to silence. The salient EDR occurs after this thunder clap, and is shown in Figure 5.2, broken out in the same way as Figure 5.1.
5.2 Strong Psychophysiological Responses in EIM

Figure 5.2: Mean EDA signals from participants who listened to Slayer’s Raining Blood in Dublin, Taipei City, and Taichung City.

5.2.3 Into Dust

Another example is shown in the top plot of Figure 5.3 from the excerpt of Into Dust by Mazzy Star (the excerpt is taken from 2’08” 25 through 3’40” 50). The vertical line in each plot again indicates the moment in the stimulus that is believed to trigger the EDR. In the music, this corresponds to the first appearance of the performer’s voice (at 0’25” 26 in the excerpt). Physiological arousal elicited from the first entrance of a singing voice supports previous findings in [59]. It is interesting to note that in Dublin and Taiwan Into Dust ranked as one of the least familiar musical excerpts among participants. This could be an indication that these events are independent, to a certain degree, from the listeners’ existing knowledge of the music.

The Into Dust excerpt begins (similarly to the corresponding section of the Hallelujah excerpt) with a simple ostinato played by an acoustic guitar. The ostinato outlines a Gm7-Em7-Dm progression. This eight bar sequence is repeated a second time, now joined by slowly descending lines from a flute and cello. At the end of the repeated sequence, the vocal enters on the lyric I could possibly be fading / Or have something more to gain. It is this entrance of the voice that seems to elicit a strong psychophysiological response.
88 Strong Physiological Responses

Figure 5.3: Mean EDA signals from participants who listened to Mazzy Star’s Into Dust in Dublin, Taipei City, and Taichung City.

5.3 Modeling Strong Psychophysiological Reactions

5.3.1 Identifying Reactions Deterministically

Since previous observations of these strong psychophysiological reactions had been primarily visual, it was important to establish a means for deterministically categorizing a signal as a reaction signal or a no-reaction signal. A peak finding algorithm was used to identify peaks in a candidate signal. This algorithm was parameterized by the minimum peak distance, the minimum peak prominence, the minimum peak width, and the minimum peak height. The algorithm operated by first identifying all peaks in the signal. Beginning with the tallest peak, any other peak within a given distance was excluded from further consideration; the minimum peak distance governed this distance. After this, the algorithm examined the next tallest peak and continued excluding candidate peaks until reaching the end of the list of peaks. Peak prominence is a measure of how a given peak stands out relative to other nearby peaks. Peak width is the measure of time between the two valleys surrounding a given peak, and peak height is the measure of the absolute value of the peak. Specifically, this algorithm was parameterized to find all peaks with a minimum peak distance of 1 second, a minimum peak prominence of 1µS, a minimum peak height of 0µS, and a minimum peak width of 0.9 seconds. After peaks that failed to reach these criteria were removed, any signal with a peak within a time window around the moment of interest in the signal was labeled as a reaction signal, and all others were labeled as no-reaction signals. These time windows were set to begin 1.5 seconds before the moments of interest identified in each song in Sections 5.2.1, 5.2.2, and 5.2.3. The windows
each ended 8.5 seconds after the moment of interest, for a total window length of 10 seconds. The windows were permitted to begin before the moments of interest in order to account for participants who may have anticipated it due to familiarity with the music. Furthermore, the windows were permitted to extend 8.5 seconds after the moment of interest due to not only the natural latency in EDRs, but also individual differences in EDR latency [9], [16].

After segmentation, all results were visually inspected to confirm classification of each signal. This inspection confirmed that in almost all cases, this algorithm performed well. For some stimuli, there were rare cases when a strong EDR appeared to be associated with the moment of interest in the music, but the signal was classified as no-reaction. In most of these cases, an initial EDR was identified as beginning within the window, but an additional EDR began before the fall of the original EDR, causing the ultimate maximum to fall outside of the window. In other cases, an initial EDR was identified as beginning within the window, but the peak of this EDR fell outside of the window. In these few cases, such false negatives were manually reclassified as reaction examples. Improvements to this algorithm, especially for practical purposes, would be useful.

5.3.2 Exploring Reaction Groups

The analysis discussed here was performed using Jeff Buckley’s *Hallelujah*. This includes the work described in this section up through the beginning of Section 5.3.5. Later analysis, beginning in Section 5.3.5.2, incorporated data from all of the stimuli described in Sections 5.2.1 through 5.2.3.

As a first step, the data collected during a trial of EiM from participants who listened to *Hallelujah* are considered, in order to identify what relationships may exist between, on the one hand, characteristics of the participants and the way they rated their affective and other responses during the audition, and on the other hand, whether or not they demonstrated the strong psychophysiological reaction.

The data considered here was gathered from all participants who listened to *Hallelujah* during the EiM iterations in Dublin, Taipei City, and Taichung City. From this group, only those participants who had EDA signals 90% of which were free of artifacts were included (n = 303). The mean age of these participants was 24.7 years (Figure 5.4a). Age was not normally distributed, included a handful of outliers (e.g., several 121-year-olds), and was skewed older. This was likely due to the fact that EiM is often staged in science museums, which are frequented more heavily by younger audiences (e.g., large groups of students in field trips). Included in the sample were 153 males and 150 females (Figure 5.4b). The nationalities represented by participants in the sample were overwhelmingly Taiwanese,
Irish, and British (Figure 5.4c; in the iterations in Dublin, “Other” was provided as a selection for Nationality, but beginning in Taiwan, a full list of nationalities was provided from which participants could make a selection). Nevertheless, irrespective of location, the majority of participants interacted with the study in English, as opposed to Taiwanese.

Figure 5.4: Demographics for participants who heard *Hallelujah* in Dublin, Taiwan City, and Taichung City: (a) distributions of ages by reaction group; (b) counts of sexes by reaction group; (c) counts of nationalities; (d) counts of languages.

5.3.2.1 Univariate Distributions

Considering relationships between univariate distributions and reaction groups reveals some important distinctions between the two groups—the group that did manifest a strong physiological response (the reaction group), and the group that did not (the no-reaction group). Moreover, these comparisons highlight the striking similarities between the two groups.

PARTICIPANT CHARACTERISTICS Inspection of the distributions of various participant characteristics underscores this similarity. Plots of example distributions broken out by whether or not the participant demonstrated the reaction are shown in Figure 5.5.
With respect to age, the ages of those participants that did demonstrate the reaction appear to be more narrowly concentrated around the peak of the distribution, which is greater than that of the group of participants that did not demonstrate the reaction. This difference was insignificant, however ($\mu_{\text{Reaction}} = 22.4$, $\mu_{\text{No-reaction}} = 23.4$, Mann-Whitney $U = 8079.5$, $p = 0.9835$). Also insignificant were differences in the distribution of how well participants rated their concentration during the study (most participants saw themselves as very focused; $\mu_{\text{Reaction}} = 3.92$, $\mu_{\text{No-reaction}} = 4.01$, Mann-Whitney $U = 3464$, $p = 0.5489$). Most other distributions were quite similar. The combined distribution of musical expertise across the groups appears to be bimodal, with the group that demonstrated the reaction appearing to contribute a larger portion of lower ratings in musical expertise. Nevertheless, the difference between these two distributions was not significant in this analysis ($\mu_{\text{Reaction}} = 2.35$, $\mu_{\text{No-reaction}} = 2.65$, Mann-Whitney $U = 3501$, $p = 0.1155$). There does seem to be a relationship worthy of further exploration here—musical expertise may play some role in whether or not one experiences a strong psychophysiological reaction such as this, and these results seem to indicate that the less musically experienced one is, the more likely they may be to experience such a reaction. This may mean that previous claims of more musical experience predisposing one to stronger affective responses to music [155] are not as straightforward as researchers previously believed.

AFFECTIVE AND MUSICAL RESPONSES With respect to participants’ affective and musical responses to the musical stimulus, results were largely similar to those from participant characteristics (Table 5.1). While there are no appreciable differences between the reaction groups in terms of engagement, song like/dislike, felt positivity, and felt tension, felt activity and familiarity are more interesting (Figure 5.6). First of all, activity is a measure of the participant’s rating of the intensity of their affective response (felt arousal). The difference in these ratings between the reaction groups is worthy of future exploration. It is possible that those participants who did present the reaction were more likely to rate their felt arousal as less intense—this is rather surprising.

Perhaps counterintuitive are the participants’ ratings of their familiarity with the musical stimulus, in which there may have been a difference between reaction groups. Specifically, participants who were more familiar with the stimulus may have been less likely to demonstrate the reaction. One possible explanation for this is the likelihood that the salient moment in the stimulus was less predictable—and thus more novel—for those participants who were less familiar with the music. On the other hand, people who are more familiar with a particular piece of music arguably anticipate such a moment in the music, thus potentially amplifying their
Figure 5.5: Univariate distributions of participant characteristics by reaction group for participants who heard *Hallelujah* in Dublin, Taiwan City, and Taichung City: (a) age; (b) concentration; (c) musical expertise.

affective response [68], [69], [111], [115], [119]. In either case, this should be considered in future research.
Table 5.1: Comparisons of affective/musical responses between reaction groups from participants who heard *Hallelujah* in Dublin.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\mu_{\text{Reaction}}$</th>
<th>$\mu_{\text{No-reaction}}$</th>
<th>Mann-Whitney $U$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>3.19</td>
<td>3.45</td>
<td>6944.5</td>
<td>.063</td>
</tr>
<tr>
<td>Engagement</td>
<td>3.59</td>
<td>3.43</td>
<td>8710.0</td>
<td>.291</td>
</tr>
<tr>
<td>Familiarity</td>
<td>2.64</td>
<td>3.03</td>
<td>6931.0</td>
<td>.062</td>
</tr>
<tr>
<td>Like/dislike</td>
<td>3.57</td>
<td>3.33</td>
<td>8773.0</td>
<td>.248</td>
</tr>
<tr>
<td>Positivity</td>
<td>3.59</td>
<td>3.50</td>
<td>8498.0</td>
<td>.478</td>
</tr>
<tr>
<td>Tension</td>
<td>4.20</td>
<td>4.06</td>
<td>8706.5</td>
<td>.276</td>
</tr>
</tbody>
</table>

Figure 5.6: Distributions of activity and familiarity by reaction group for participants who heard *Hallelujah* in Dublin: (a) Activity (felt arousal); (b) Familiarity.

**Participant Personality** The results of tests between distributions of scores of participant personality items are given in Table 5.2. As with
most of the variables measured for participant characteristics and affective/musical responses, most of these comparisons revealed no significant differences between groups. However, several deserve closer inspection.

Table 5.2: Comparisons of personality items between reaction groups for participants who heard *Hallelujah* in Taiwan (*n* = 188).

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\mu_{\text{Reaction}}$</th>
<th>$\mu_{\text{No-reaction}}$</th>
<th>Mann-Whitney <em>U</em></th>
<th><em>p</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Artistic</td>
<td>2.38</td>
<td>2.33</td>
<td>3797.5</td>
<td>.670</td>
</tr>
<tr>
<td>Fault</td>
<td>3.44</td>
<td>3.10</td>
<td>4140.0</td>
<td>.087</td>
</tr>
<tr>
<td>Imagination</td>
<td>3.81</td>
<td>3.87</td>
<td>3392.0</td>
<td>.532</td>
</tr>
<tr>
<td>Lazy</td>
<td>3.85</td>
<td>3.63</td>
<td>3796.0</td>
<td>.668</td>
</tr>
<tr>
<td>Nervous</td>
<td>3.83</td>
<td>3.57</td>
<td>3941.5</td>
<td>.273</td>
</tr>
<tr>
<td>Outgoing</td>
<td>2.96</td>
<td>3.31</td>
<td>2975.0</td>
<td>.059</td>
</tr>
<tr>
<td>Reserved</td>
<td>3.33</td>
<td>3.08</td>
<td>4054.5</td>
<td>.224</td>
</tr>
<tr>
<td>Stress</td>
<td>2.80</td>
<td>3.18</td>
<td>2995.0</td>
<td>.045</td>
</tr>
<tr>
<td>Thorough</td>
<td>3.26</td>
<td>3.03</td>
<td>4015.0</td>
<td>.190</td>
</tr>
<tr>
<td>Trusting</td>
<td>3.53</td>
<td>3.77</td>
<td>3293.0</td>
<td>.251</td>
</tr>
</tbody>
</table>

The comparisons of fault (“I see myself as someone who tends to find fault with others”) and outgoing (“I see myself as someone who is outgoing, sociable”) indicate that there may be a relationship between these variables and whether or not the participant demonstrated a reaction. It does appear that participants who presented the reaction generally scored themselves as less outgoing and more likely to find fault in others than those who did not present the reaction. The test for differences in stress scores (“I see myself as someone who is relaxed, handles stress well”), however, provides sufficient evidence of a significant difference between the two groups. In general, participants who presented the reaction scored themselves as less relaxed than those participants who did not present the reaction.

On the Big Five Inventory-10 (10-item) (*BFI-10*), two items are presented corresponding to each of the five personality traits the inventory is designed to measure: extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Fault, outgoing, and stress correspond to three different personality traits: agreeableness, extroversion, and neuroticism, respectively. When considering comparisons of the overall traits (Table 5.3), there appears to be a relationship between agreeableness and reaction group (participants who presented a reaction may have scored themselves as less agreeable than those who did not). The significant difference found in comparing the stress item does carry over to the trait it
measures (neuroticism); there is sufficient evidence also at the trait level to conclude that in these data there is a relationship between neuroticism and reaction group (in general, those who rated themselves as more neurotic were more likely to present the reaction).

Table 5.3: Comparisons of personality traits between reaction groups for participants who heard *Hallelujah* in Taiwan ($n = 188$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\mu_{\text{Reaction}}$</th>
<th>$\mu_{\text{No-reaction}}$</th>
<th>Mann-Whitney $U$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td>3.06</td>
<td>3.34</td>
<td>2982.0</td>
<td>.062</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>2.71</td>
<td>2.70</td>
<td>3637.5</td>
<td>.888</td>
</tr>
<tr>
<td>Extroversion</td>
<td>2.53</td>
<td>2.62</td>
<td>3412.0</td>
<td>.459</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>3.50</td>
<td>3.20</td>
<td>4285.0</td>
<td>.034</td>
</tr>
<tr>
<td>Openness to experience</td>
<td>3.74</td>
<td>3.77</td>
<td>3493.5</td>
<td>.768</td>
</tr>
</tbody>
</table>

Something of a picture, though conflicting, begins to develop here. The relationships that these data suggest are that those participants who did experience the strong psychophysiological reaction experienced lower felt arousal, were less familiar with the music, less musically experienced, more prone to find fault in others, less outgoing, and less agreeable than those participants who did not present the reaction. Furthermore, the results did reach significance in supporting the idea that the participants who did experience the reaction were less relaxed and more neurotic, as well.

5.3.2.2 *Bivariate Distributions*

Considering participant characteristics, affective and musical responses, and personality scores produced several paths for future exploration. Comparisons of bivariate distributions of independent variables (taking reaction group as the dependent variable), or the relationships between two variables, proved to be interesting, as well.

The impetus for this analysis grew out of a visual exploration of these data. As the data are largely ordinal, scatter plots reveal little in terms of relationships between the variables. However, bivariate kernel density estimate plots by group show that while there may be little difference between the reaction groups in terms of individual predictors, this may not be true for the joint distributions of pairs of predictors. For this analysis, the Likert-type items used for measurement of various factors (e.g., personality and affective response factors) were treated as continuous measures [87].

Hotelling’s $T^2$ is a useful extension of the univariate $t$ test to a multivariate setting. Hotelling’s $T^2$ values were calculated for every pair of variables
in the dataset in order to compare these bivariate distributions between reaction groups. As expected, the majority of the results of these 192 tests were inconsequential. Over ten percent of them, however, revealed significant differences between the joint distributions between groups (Table 5.4). In these 22 relationships, only 8 different variables are involved: felt arousal (activity), age, artistic (“I see myself as someone who has few artistic interests”), concentration, engagement, imagination (“I see myself as someone who has an active imagination”), song like/dislike, positivity (felt valence), and felt tension. In fact, most (91%) of these results deal in one way or another with only three variables: concentration, artistic, and imagination—artistic and imagination are both items that measure the personality trait of openness to experience.

Table 5.4: Test results from comparisons of bivariate joint distributions between groups for participants who heard *Hallelujah* in Dublin and Taiwan (where Bonferroni-corrected \(p\)-values are < .05).

<table>
<thead>
<tr>
<th>Variables</th>
<th>(T^2)</th>
<th>(F)</th>
<th>(D^2)</th>
<th>Corrected (p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artistic</td>
<td>Activity</td>
<td>0.7894</td>
<td>0.3926</td>
<td>0.0203</td>
</tr>
<tr>
<td>Artistic</td>
<td>Age</td>
<td>0.1316</td>
<td>0.0654</td>
<td>0.0034</td>
</tr>
<tr>
<td>Artistic</td>
<td>Concentration</td>
<td>0.3430</td>
<td>0.1706</td>
<td>0.0088</td>
</tr>
<tr>
<td>Artistic</td>
<td>Engagement</td>
<td>0.1661</td>
<td>0.0826</td>
<td>0.0043</td>
</tr>
<tr>
<td>Artistic</td>
<td>Imagination</td>
<td>0.1165</td>
<td>0.0580</td>
<td>0.0030</td>
</tr>
<tr>
<td>Artistic</td>
<td>Like/Dislike</td>
<td>0.1112</td>
<td>0.0553</td>
<td>0.0029</td>
</tr>
<tr>
<td>Artistic</td>
<td>Positivity</td>
<td>0.0876</td>
<td>0.0436</td>
<td>0.0023</td>
</tr>
<tr>
<td>Artistic</td>
<td>Tension</td>
<td>0.1442</td>
<td>0.0717</td>
<td>0.0037</td>
</tr>
<tr>
<td>Concentration</td>
<td>Activity</td>
<td>0.8330</td>
<td>0.4143</td>
<td>0.0214</td>
</tr>
<tr>
<td>Concentration</td>
<td>Age</td>
<td>0.2948</td>
<td>0.1466</td>
<td>0.0076</td>
</tr>
<tr>
<td>Concentration</td>
<td>Engagement</td>
<td>0.6317</td>
<td>0.3142</td>
<td>0.0162</td>
</tr>
<tr>
<td>Concentration</td>
<td>Imagination</td>
<td>0.4222</td>
<td>0.2100</td>
<td>0.0110</td>
</tr>
<tr>
<td>Concentration</td>
<td>Like/Dislike</td>
<td>0.2494</td>
<td>0.1240</td>
<td>0.0064</td>
</tr>
<tr>
<td>Concentration</td>
<td>Positivity</td>
<td>0.2483</td>
<td>0.1235</td>
<td>0.0064</td>
</tr>
<tr>
<td>Concentration</td>
<td>Tension</td>
<td>0.4948</td>
<td>0.2461</td>
<td>0.0127</td>
</tr>
<tr>
<td>Imagination</td>
<td>Age</td>
<td>0.2270</td>
<td>0.1129</td>
<td>0.0059</td>
</tr>
<tr>
<td>Imagination</td>
<td>Engagement</td>
<td>0.1841</td>
<td>0.0915</td>
<td>0.0048</td>
</tr>
<tr>
<td>Imagination</td>
<td>Like/Dislike</td>
<td>0.1917</td>
<td>0.0953</td>
<td>0.0050</td>
</tr>
<tr>
<td>Imagination</td>
<td>Positivity</td>
<td>0.1623</td>
<td>0.0807</td>
<td>0.0042</td>
</tr>
<tr>
<td>Imagination</td>
<td>Tension</td>
<td>0.1940</td>
<td>0.0965</td>
<td>0.0051</td>
</tr>
<tr>
<td>Positivity</td>
<td>Age</td>
<td>0.6132</td>
<td>0.3056</td>
<td>0.0111</td>
</tr>
<tr>
<td>Positivity</td>
<td>Engagement</td>
<td>0.8339</td>
<td>0.4155</td>
<td>0.0151</td>
</tr>
</tbody>
</table>
For those relationships different between groups involving age (Figure 5.7 gives a visualization of a small sample of these comparisons), it appears that in general there is stronger agreement/lower variance within the group of participants that did demonstrate the reaction on values of each variable. The same is true for those relationships involving imagination and concentration, though figures for the remaining relationships have been excluded here. In general, it is difficult to characterize precisely how these variables vary differently with respect to the two groups. The most that can be said is that for each of these pairs of variables, the distance between the joint distributions is significantly different between groups. It is also important to note $D^2$ in Table 5.4. Much like Hotelling’s $T^2$ is the multivariate extension of Student’s $t$, $D^2$ is an extension of Cohen’s $d$ that leverages the Mahalanobis distance to measure the effect size associated with a given $T^2$ [146], [151]. Sapp et al. recommend the following thresholds for qualitative assignments of effect size to $D^2$ values: $D^2 = 0.25$ (small effect), $D^2 = 0.5$ (medium effect), $D^2 > 1$ (large effect). It may prove fruitful for future studies to focus closely on the interactions of attention/concentration/engagement, musical preference, age, felt affect, and (particularly) openness to experience in relation to strong psychophysiological reactions to music. Given the small effect sizes observed here, careful consideration should be given to test power. Specifically, while these tests do show statistically significant differences, the magnitude of these differences are quite small.

5.3.3 Ordinal Logistic Regression

As part of the work in attempting to develop a means of classifying participants into the reaction or no-reaction group based purely on non-physiological data, both ordinal and neural network-based logistic regression models were developed. Section 5.3.5 discusses the neural network approach. This section presents the results of a standard ordinal logistic regression.

The model was fitted using the factors and covariates discussed in Section 5.3.2. Additionally, music preference, study location, and whether or not the participant had any hearing impairments were included in the model. Without imputing missing values (personality variables, for instance, were not measure in iterations of the study in Dublin), this left a model with $n = 191$ participants, all of whom participated in the study in Taipei City or Taichung City. The goodness of fit measures for the fitted model are given in Table 5.5.

The overall model fit test measures the degree to which the inclusion of the predictors improve the model fit over the base intercept-only model (Table 5.6). There is evidence ($p = .014$) for rejection of the null hypothesis.
Figure 5.7: Example bivariate distributions involving artistic between groups for participants who heard *Hallelujah* in Taiwan.

Table 5.5: Ordinal logistic regression goodness of fit for *Hallelujah* reaction in Taiwan.

<table>
<thead>
<tr>
<th></th>
<th>Chi-square</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>270.41</td>
<td>98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Deviance</td>
<td>106.743</td>
<td>98</td>
<td>.2570</td>
</tr>
</tbody>
</table>

that the model without the predictors is as good as the model with the predictors included.

Parameter estimates for those predictors of which their inclusion in the model provides a significant improvement over the intercept-only model.
are shown in Table 5.7. First, note that for factors with multiple levels, one should expect to find that successive changes in level correspond with comparable differences in the magnitude of their respective parameter estimates. For example, for familiarity, the magnitude of the parameter estimate decreases as the score for familiarity increases from 2 to 3. This is in line with what was noted above—in general, participants who were more familiar with the stimulus appeared to be less prone to demonstrate the reaction. However, no significant parameter estimates were found for levels 1, 4, or 5. This does not imply that the estimates found for levels 2 and 3 are not significant, but this observation only holds for these two levels of the factor. In general, this suggests that the assumption of proportional odds does not necessarily hold for this model, and more generalizable results will only be found or confirmed through either a more sophisticated approach to this model (for instance, stratification on those factors for which this assumption does not hold), or through a model that treats these factors drawn from Likert-type items as continuous measures (an approach argued for by Johnson and Creech [87]). With the same caveat, at isolated factor levels, as participant scores for lazy and thorough increased, the probability for presenting a reaction decreased and increased, respectively.

More straightforward results are found in the language, location, nationality, and age factors. Participants who participated in English (as opposed to Taiwanese) were less likely to present a reaction. Participants who participated in Taichung City (as opposed to Taipei City), and participants who did not include rock in their selection of music preferences were also less likely to present a reaction. Chinese participants were less likely to present a reaction, and finally, as participant age increased, their probability of presenting a reaction decreases slightly. All of these observations assume that for a change in one factor, all other factors are held constant. For instance, this does not necessarily mean that older Chinese participants are even more unlikely to present a reaction than younger Irish participants.

These results seem to present competing priorities, but in general, may lend support for the case for novelty playing a key role in determining

Table 5.6: Ordinal logistic regression overall model fit for Hallelujah reaction in Taiwan.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 Log Likelihood</th>
<th>Chi-square</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept Only</td>
<td>231.105</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final</td>
<td>106.743</td>
<td>124.362</td>
<td>92</td>
<td>.014</td>
</tr>
</tbody>
</table>
whether or not a given musical stimulus provokes a strong physiological reaction in a listener. Being less familiar with a given piece of music increases its novelty by definition. Among the music preferences available for selection in the study, rock is the listed genre most characteristic of *Hallelujah*. Arguably then, participants who did not prefer rock music were less likely to be familiar with the stimulus. There is a similar though weaker argument to be made for a relationship between an increase in age and an increase in novelty of the stimulus. Given the distribution of ages among participants however, one would have expected the opposite to be the case (with a mean participant age of only 25 years, it is possible that the older participants in this sample would be the ones more familiar with *Hallelujah*).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>95% CI ↓ Bound</th>
<th>95% CI ↑ Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction = No</td>
<td>11.240</td>
<td>226.179</td>
<td>0.000</td>
<td>1</td>
<td>.995</td>
<td>-44.17805</td>
<td>44.148284</td>
</tr>
<tr>
<td>Activity = 1</td>
<td>5.201</td>
<td>2.413</td>
<td>4.645</td>
<td>1</td>
<td>.031</td>
<td>0.4711</td>
<td>9.931</td>
</tr>
<tr>
<td>Engagement = 2</td>
<td>-7.945</td>
<td>3.639</td>
<td>7.666</td>
<td>1</td>
<td>.029</td>
<td>-15.077</td>
<td>-0.812</td>
</tr>
<tr>
<td>Familiarity = 2</td>
<td>6.044</td>
<td>2.914</td>
<td>3.030</td>
<td>1</td>
<td>.038</td>
<td>0.3341</td>
<td>11.755</td>
</tr>
<tr>
<td>Familiarity = 3</td>
<td>4.358</td>
<td>2.181</td>
<td>3.092</td>
<td>1</td>
<td>.046</td>
<td>0.0831</td>
<td>8.634</td>
</tr>
<tr>
<td>Language = English</td>
<td>-2.687</td>
<td>1.280</td>
<td>4.403</td>
<td>1</td>
<td>.036</td>
<td>-5.197</td>
<td>-0.177</td>
</tr>
<tr>
<td>Lazy = 3</td>
<td>5.929</td>
<td>2.118</td>
<td>7.834</td>
<td>1</td>
<td>.005</td>
<td>1.7771</td>
<td>10.080</td>
</tr>
<tr>
<td>Lazy = 4</td>
<td>3.755</td>
<td>1.777</td>
<td>4.463</td>
<td>1</td>
<td>.035</td>
<td>0.2711</td>
<td>7.238</td>
</tr>
<tr>
<td>Location = Taichung</td>
<td>-3.596</td>
<td>1.553</td>
<td>5.360</td>
<td>1</td>
<td>.021</td>
<td>-6.641</td>
<td>-0.552</td>
</tr>
<tr>
<td>Music Pref (Rock) = 0</td>
<td>-2.988</td>
<td>1.211</td>
<td>6.093</td>
<td>1</td>
<td>.014</td>
<td>-5.361</td>
<td>-0.615</td>
</tr>
<tr>
<td>Nationality = Chinese</td>
<td>-8.026</td>
<td>3.717</td>
<td>4.662</td>
<td>1</td>
<td>.031</td>
<td>-15.312</td>
<td>-0.740</td>
</tr>
<tr>
<td>Stress = 2</td>
<td>5.064</td>
<td>2.497</td>
<td>4.113</td>
<td>1</td>
<td>.043</td>
<td>0.1701</td>
<td>9.959</td>
</tr>
<tr>
<td>Thorough = 1</td>
<td>-8.917</td>
<td>4.389</td>
<td>4.128</td>
<td>1</td>
<td>.042</td>
<td>-17.319</td>
<td>-0.315</td>
</tr>
<tr>
<td>Thorough = 2</td>
<td>-10.196</td>
<td>3.814</td>
<td>7.146</td>
<td>1</td>
<td>.008</td>
<td>-17.672</td>
<td>-2.721</td>
</tr>
<tr>
<td>Thorough = 3</td>
<td>-8.030</td>
<td>3.304</td>
<td>5.906</td>
<td>1</td>
<td>.015</td>
<td>-14.506</td>
<td>-1.554</td>
</tr>
<tr>
<td>Thorough = 4</td>
<td>-7.659</td>
<td>3.445</td>
<td>4.943</td>
<td>1</td>
<td>.026</td>
<td>-14.411</td>
<td>-0.907</td>
</tr>
<tr>
<td>Trusting = 3</td>
<td>-6.431</td>
<td>2.252</td>
<td>8.156</td>
<td>1</td>
<td>.004</td>
<td>-10.844</td>
<td>-2.017</td>
</tr>
<tr>
<td>Age</td>
<td>-0.123</td>
<td>0.058</td>
<td>4.569</td>
<td>1</td>
<td>.033</td>
<td>-0.236</td>
<td>-0.010</td>
</tr>
</tbody>
</table>

5.3.4 Principal Component Analysis

A principal component analysis (PCA) of these data sheds some light on the difficulty in finding more meaningful relationships not only between predictors, but also between predictors and reaction groups. Table 5.8 shows the variance captured by the first ten components in this analysis. In total, these ten components explain 72.26% of the total variance in the dataset.
The relative magnitudes of the loadings of each of these components are plotted in Figure 5.8. Two things are important to note with respect to this figure. First, each successive component captures a smaller proportion of variance, and the magnitudes of loadings cannot be compared directly between two components. The line weights of each component’s plot vary proportionally with the amount of variance the component captures in order to call out this difference visually. Second, a number of the more substantial magnitudes are assigned to indicator variables. Interpretation of these magnitudes should also be approached carefully. For instance, note that in the second component the two indicator variables assigned to sex have nearly equivalent loadings, but with different signs—this is to be expected and doesn’t provide any real insight into the data. (PCA including these indicator variables is not inappropriate here, as these results will only later be used for dimensionality reduction.)

Bearing all of this in mind, the most interesting variables in this analysis are those that fall primarily to the right side of Figure 5.8. Specifically, participant scores for engagement, familiarity, song like/dislike, and positivity (felt valence) explain a great deal more variance in these data than, for instance, any of the variables that deal with participant personality or other affective response measures. Importantly, note that this has nothing to do with variance that exists between reaction groups; this deals with variance between all participants irrespective of whether or not they presented a reaction. Nevertheless, there are a few important trends to observe here, primarily those in the first component to the right of the figure. First, in this component, engagement, song like/dislike, and positivity vary in the same direction. Intuitively, the more engaged with a given selection of music, and the more positive one’s felt valence, the stronger they tend to rate their liking of the music (at least for Hallelujah). On the other hand, familiarity varies inversely with these three variables. Interestingly, although there is no measure of significance here, the data do seem to indicate that participants who were less familiar with Hallelujah tended to rate their engagement, liking, and felt positive valence of their affective response more highly.

With respect to the dimensionality reduction mentioned above, example scatter plots of pairs of component values from these data after having been transformed using the results of the PCA are given in Figure 5.9. Clearly, the PCA has correctly identified the components that capture the most variance (as demonstrated by the clear separation of groups). However, at least in two dimensions, none of these separations coincide with a separation between the two reaction groups. Due to the nature of PCA, the pairwise plots not shown here still show segmentation between the components, but this segmentation worsens as components that explain less variance are visualized. Like the plots shown here, plots of less important components

Overall, engagement, familiarity, song like/dislike, and positivity (felt valence) explain the most important variance in the first component.
Table 5.8: Variance explained as a result of Dublin Hallelujah PCA.

<table>
<thead>
<tr>
<th>Component</th>
<th>Variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.04%</td>
</tr>
<tr>
<td>2</td>
<td>10.54%</td>
</tr>
<tr>
<td>3</td>
<td>6.08%</td>
</tr>
<tr>
<td>4</td>
<td>5.02%</td>
</tr>
<tr>
<td>5</td>
<td>4.55%</td>
</tr>
<tr>
<td>6</td>
<td>4.14%</td>
</tr>
<tr>
<td>7</td>
<td>3.70%</td>
</tr>
<tr>
<td>8</td>
<td>3.33%</td>
</tr>
<tr>
<td>9</td>
<td>3.18%</td>
</tr>
<tr>
<td>10</td>
<td>2.67%</td>
</tr>
</tbody>
</table>

**Total** 72.26%

![Component loadings graph](image)

Figure 5.8: Component loadings for first five components from Dublin and Taiwan Hallelujah PCA.
also fail to segment reaction groups cleanly. For these reasons, plots of only the top four pairs of components are included here.

Figure 5.9: Example pairwise plots of component values from PCA-transformed Dublin and Taiwan Hallelujah data.

5.3.5 Neural Networks

If meaningful relationships between the predictors selected for analysis here and reaction groups do exist, the previous results (especially those from the PCA in Section 5.3.4) suggest that the decision boundary of a classifier capable of separating these two groups would be reasonably complex. Neural networks can be used as classifiers capable of learning such complex decision boundaries.
5.3.5.1 Initial Exploratory Modeling

The primary question in developing these models was, devoid of any knowledge of the psychophysiological response of a given participant, would it be possible to determine whether or not a given participant would demonstrate a strong psychophysiological response to a given stimulus? In other words, is sufficient information available in the demographic, personality, and self-reported affective/musical response variables that were measured to indicate that an individual was more or less prone to strong psychophysiological reactions in response to a given stimulus? To explore this an initial prototype neural network was built that included all of the features listed in Table 5.9.

Table 5.9: Available features for prototype neural network models of reaction vs. no-reaction groups described in Section 5.3.5.1.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Personality</th>
<th>Affective/Musical Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Artistic</td>
<td>Activity</td>
</tr>
<tr>
<td>Concentration</td>
<td>Fault</td>
<td>Engagement</td>
</tr>
<tr>
<td>Hearing impairments</td>
<td>Imagination</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Music preferences</td>
<td>Lazy</td>
<td>Like/Dislike</td>
</tr>
<tr>
<td>Musical expertise</td>
<td>Nervous</td>
<td>Positivity</td>
</tr>
<tr>
<td>Language</td>
<td>Outgoing</td>
<td>Tension</td>
</tr>
<tr>
<td>Location</td>
<td>Reserved</td>
<td>Strong response</td>
</tr>
<tr>
<td>Nationality</td>
<td>Stress</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Thorough</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trusting</td>
<td></td>
</tr>
</tbody>
</table>

The same data segmentation and preprocessing were performed for all data models presented in Section 5.3.5.1. A 70%/30% split was used to create training and validation datasets respectively by drawing samples at random from 364 observations. The same 303-observation training set and 61-observation validation set were used to train and evaluate each model discussed in this section (including the baseline classifier against which these models were compared—Section 5.3.5.1). All data were preprocessed identically to generate identical datasets from which the models could draw. Where data were missing (for example, personality variables were not measured in Dublin), the mean value of the variable was computed and used for such an observation. In general, all models presented in Section 5.3.5.1 were trained and evaluated on identical training and validation sets, respectively.
**Baseline Classifier**  As the primary purpose of these models was to determine if, given these data, it would be possible to determine automatically whether or not a given participant did or did not experience a reaction, it was necessary to construct a baseline classifier against which the other models presented here could be evaluated. There are several options for constructing such a baseline classifier. Often, a classifier that chooses a class at random for a given example is used. When working with a class imbalance, however, this approach is not appropriate. In this case, it is more appropriate that the baseline classifier always label a candidate example with the majority class—in this case no-reaction. The validation dataset used for this analysis contained 41 no-reaction examples and 20 reaction examples. Thus, the accuracy of this baseline classifier is $41/61 = .672$. The confusion matrix for this baseline classifier is given in Table 5.10.

Table 5.10: Confusion matrix for baseline classifier that predicts no-reaction for all examples for Dublin and Taiwan Hallelujah validation dataset.

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Actual Class</th>
<th>Reaction</th>
<th>No Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No Reaction</td>
<td>20</td>
<td>41</td>
<td>61</td>
</tr>
</tbody>
</table>

In this section, $F_\beta$ scores will be used to compare performance between various classifiers. Understanding $F_\beta$ scores presupposes an understanding of precision and recall, however. The *precision* of a classifier measures the ability of a classifier to correctly identify members of a relevant class. As a question, precision asks, for all the examples of a relevant class a classifier marked as members of that class, how were correctly classified? For example, for a classifier that classifies images as containing dogs or cats, the precision of the classifier on the dogs class would be measured as the ratio of the number images of dogs it correctly classified to the total number of images it classified as containing a dog. Defined mathematically, precision (in general) is

$$\frac{TP}{TP + FP}$$

where $TP$ denotes true positives (correct classifications of the relevant class), and $FP$ denotes false positives (incorrect classifications of the relevant class as the relevant class).

On the other hand, *recall* is a measure of how many examples of a relevant class a classifier correctly classified. As a question, recall asks,
among all of the provided examples of a relevant class shown to a classifier, how many examples did it classify with the relevant class? Returning to the dogs and cats example, the recall of the classifier on the dogs class would be measured as the ratio of the number of images of all dogs it correctly classified to the total number of images of dogs it was shown. Mathematically, this is

\[
\frac{TP}{TP + FN}
\]

where \(TP\) still denotes true positives, and \(FN\) denotes false negatives (incorrect classifications the relevant class as something other than the relevant class).

\(F_\beta\) scores provide a weighted combination of precision and recall. The \(F_1\) score—the most commonly used \(F_\beta\) score—is the harmonic mean of precision and recall, defined as

\[
2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

The \(F_1\) is a specific case of an \(F_\beta\) score. The general \(F_\beta\) score is parameterized by \(\beta\), which provides a means of weighting the importance of precision and recall differently. The \(F_\beta\) score is defined as

\[
(1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}
\]

Thus, an \(F_{1.0}\) score places more importance on recall, and an \(F_{0.1}\) score places more importance on precision.

Returning to Table 5.10, the \(F_1\) score for this confusion matrix (taking the reaction class as the relevant class) is ill-defined, as precision is undefined for a classifier that only predicts the irrelevant class. The weighted \(F_1\) score (not to be confused with the weighting done by choosing a different \(F_\beta\) score) is calculated as the mean of the class-wise \(F_1\) scores each weighted by the ratio of observations in each class. For the no-reaction class the classifier gives an \(F_1\) score of .80 (precision = .67; recall = .8), and for the reaction class the \(F_1\) score is 0 (precision = undefined; recall = 0). This gives a weighted \(F_1\) score of .5377. As identifying reaction examples is the more interesting (and more difficult) task here, \(F_{0.5}\) scores will be compared between classifiers, which weighs precision more highly than recall. In other words, for a given input, these comparisons place more importance on a classifier correctly classifying a reaction example or incorrectly classifying it as a no-reaction example over incorrectly classifying such an example as a reaction example. For this validation dataset, the baseline classifier gives a weighted \(F_{0.5}\) score of .4834.
ARCHITECTURE/HYPERPARAMETER TUNING A range of possible architectures and input feature subsets were explored in developing these models. As a start, a model was developed that took all available input features. A wide-and-deep classifier [27] was used consisting of 5 layers with 17 nodes, all using an exponential linear unit for their nonlinearity. Dropout regularization was not employed. Training of the model used batch Adagrad optimization (batch size = 109, learning rate = 0.35) and was run for 15,000 epochs. Initially, all available input features were provided to the deep portion of the model, and no features were provided to the wide (linear) portion of the model. Testing on the validation dataset yielded a poor AUC-PR of .395 and the confusion matrix given in Table 5.11. These results give a precision of .467, a recall of .35, and an F0.5 score of .4375 for the reaction class. Conversely, for the no-reaction class, there is a precision of .717, a recall of .805, and an F0.5 score of .733. Taken together, these give a weighted F0.5 score of .636. This does appear to be a slight improvement over the baseline classifier.

Table 5.11: Confusion matrix for wide-and-deep classifier using all basic features for Dublin and Taiwan Hallelujah data.

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Reaction</th>
<th>No Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>No Reaction</td>
<td>13</td>
<td>33</td>
</tr>
</tbody>
</table>

The model was then shifted from Tensorflow’s wide-and-deep model to a simpler network with an identical deep portion, but no wide portion. The model hyperparameters between this and the previous model, and this and other models in Section 5.3.5.1 differed as a result of Bayesian hyperparameter optimization ([160]). These parameters did vary widely, and as such, will not be described in detail.

Testing on the validation dataset yielded an improved AUC-PR of .468 and the confusion matrix given in Table 5.12. These results give a precision of .5, a recall of .45, and an F0.5 score of .438 for the reaction class. Conversely, for the no-reaction class, there is a precision of .744, a recall of .78, and an F0.5 score of .751. Improving upon the previous model, this model has a weighted F0.5 score of .665.

FEATURE SELECTION Following these prototype models, a simple single-layer network was constructed using L1 regression (least absolute shrinkage and selection operator (LASSO) regression) in order to identify those features that were most important. L1 regression is a simple linear...
Table 5.12: Confusion matrix for basic neural network classifier using all basic features for Dublin and Taiwan Hallelujah data.

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Reaction</th>
<th>No Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>No Reaction</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>41</td>
</tr>
</tbody>
</table>

regression with the L1 norm of each feature’s weight added to the ordinary least squares error term. Because of this, the weights of features that contribute most to the error are pushed toward zero. The resulting weights are shown in Figure 5.10. Excluded from this figure are those variables for which one-hot encoded vectors were used (language, location, nationality, and sex). Clearly, the weight for the “none” music preference variable dominates all other weights. Several personality variables are more important than others (thorough and stress outweigh artistic and reserved, for example). Beyond no music preference, however, there is no subset of those five or ten features that may alone be most useful in this classification problem. It is important to note, however, that these magnitudes are most relevant to a linear regression. In other words, in more complex models (such as more complex neural networks that introduce nonlinearities and compound effects between multiple variables), ignoring smaller weights may mean ignoring important information.

As such, these LASSO weights were ordered from largest to smallest and used to build successively larger models, with each next larger model incorporating four additional features added to the input space. For example, the first model took as inputs the four features with the largest weights from the LASSO regression. The second model then took as inputs the eight features with the largest weights from the LASSO regression. This process of building up models was continued until a model with the 28 features with the largest weights from the LASSO regression were included. The following hyperparameters were made available for optimization: training batch size, number of training epochs, number of hidden layers, number of nodes in each hidden layer (the same number of nodes was used for each hidden layer), dropout probability (taken as the same probability for each hidden layer), learning rate, and activation function/nonlinearity (chosen from a rectified linear unit, a...
leaky rectified linear unit, or an exponential linear unit). A summary of the confusion matrices and weighted $F_{0.5}$ scores for each of these models is given in Figure 5.13. Figure 5.11 shows the weighted $F_{0.5}$ scores plotted against the number of features in each model. Though not extreme, there is a clear positive relationship between the number of features in the model and the model’s corresponding weighted $F_{0.5}$ score. Indeed, as this is a difficult classification task, additional features do not appear to introduce enough noise to warrant their exclusion.

5.3.5.2 Refined Models

While introducing additional features into the models does tend to improve their performance in terms of weighted $F_{0.5}$ scores, it seems that the signal-to-noise ratio between most features and their predictive ability in terms of determining a reaction is low. This, and the performance of the initial prototype models, presented two additional questions. First, are there additional features that can be added to these models? Second, how generalizable is this approach? With respect to the first question, for some participants there was additional data (again, outside of data directly derived from their psychophysiological response to the experimental stimuli) for some participants that could be added to these models. Specifically, beginning in Taiwan the presentation of a control stimulus was added to the experimental procedure (Section 3.5.2). Following this stimulus, the
Table 5.13: Results for classifiers with successively larger numbers of input features as determined by LASSO regression on Dublin and Taiwan Hallelujah data.

<table>
<thead>
<tr>
<th>Features</th>
<th>True +</th>
<th>True -</th>
<th>False +</th>
<th>False -</th>
<th>Weighted $F_{0.5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>37</td>
<td>4</td>
<td>15</td>
<td>.6457</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>33</td>
<td>8</td>
<td>13</td>
<td>.6363</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>32</td>
<td>9</td>
<td>9</td>
<td>.7049</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>32</td>
<td>9</td>
<td>11</td>
<td>.6653</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>31</td>
<td>10</td>
<td>10</td>
<td>.6721</td>
</tr>
<tr>
<td>24</td>
<td>9</td>
<td>35</td>
<td>6</td>
<td>11</td>
<td>.7072</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>25</td>
<td>6</td>
<td>12</td>
<td>.6867</td>
</tr>
</tbody>
</table>

Figure 5.11: Changes in weighted $F_{0.5}$ scores in response to number of input features in the models of Dublin and Taiwan Hallelujah data.

Many of the additional affective/musical response variables in Table 5.14 come from the Geneva Emotional Music Scale (GEMS) scale [182].

same self-reported response variables were measured from the participant (with the exception of asking how familiar the participant was with the stimulus). Because of this, for all of the models discussed in this section, the set of features in Table 5.14 were used. Additionally, these refined models introduced the additional stimuli discussed in Section 5.2, Raining Blood and Into Dust.

Five strategies were employed in segmenting the data for these refined models. Five questions corresponded with these five strategies:
Table 5.14: Available features for neural network models of reaction vs. no-reaction groups discussed in Section 5.3.5.2.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Personality</th>
<th>Affective/Musical Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Artistic</td>
<td>Activity (Control, Experimental)</td>
</tr>
<tr>
<td>Concentration</td>
<td>Fault</td>
<td>Chills (Experimental)</td>
</tr>
<tr>
<td>Hearing impairments</td>
<td>Imagination</td>
<td>Chills/Shivers/Thrills (Experimental)</td>
</tr>
<tr>
<td>Music preferences</td>
<td>Lazy</td>
<td>Engagement (Control, Experimental)</td>
</tr>
<tr>
<td>Musical expertise</td>
<td>Nervous</td>
<td>Goosebumps (Experimental)</td>
</tr>
<tr>
<td>Language</td>
<td>Outgoing</td>
<td>Familiarity (Experimental)</td>
</tr>
<tr>
<td>Location</td>
<td>Reserved</td>
<td>Joyful activation (Experimental)</td>
</tr>
<tr>
<td>Nationality</td>
<td>Stress</td>
<td>Like/Dislike (Control, Experimental)</td>
</tr>
<tr>
<td>Sex</td>
<td>Thorough</td>
<td>Nostalgia (Experimental)</td>
</tr>
<tr>
<td>Visual Impairments</td>
<td>Trusting</td>
<td>Overwhelmed (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peacefulness (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positivity (Control, Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power (Control, Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sadness (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spirituality (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong response (Control, Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tenderness (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tension (Control, Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thrills (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcendence (Experimental)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wonder (Experimental)</td>
</tr>
</tbody>
</table>

1. If a model were trained and validated against data from one stimulus, how well would it classify completely unseen examples from the same stimulus? (This was repeated for each stimulus.)

2. If a model were trained and validated against data from one stimulus, how well would it classify completely unseen examples from the other stimuli? (This was repeated for each such possible grouping of stimuli.)

3. If a model were trained and validated against data from two stimuli, how well would it classify completely unseen examples from the one remaining stimulus? (This was repeated for each such possible grouping of stimuli.)
4. If a model were trained and validated against data from all stimuli, how well would it classify completely unseen examples from all stimuli?

For all of the tests presented in this section, the training, validation, and test datasets were completely segregated. This differs from those models presented in Section 5.3.5.1 in that there, judgments were made about the classifiers’ performance based on their application to the validation dataset. This validation dataset informed the training of the model (for instance, in determining optimal hyperparameters), though not nearly to the extent that the training dataset did. Importantly, the test sets used with the models in this section were completely unseen by the models until they were compared against a baseline classifier.

Additionally, the models developed to answer the above questions each required different baseline classifiers. All of these baseline classifiers were developed using the same approach as the baseline classifier discussed in Section 5.3.5.1. In all cases here, as in the classifier from Section 5.3.5.1, the classifiers were set to predict the no-reaction class, as this was always the majority class.

Once the baseline classifier was constructed for each of the tests outlined above, Dietterich’s 5x2cv paired t test [43] was used to compare the performance of the baseline classifier, and a classifier trained using the features outlined in Table 5.14. Dietterich’s test was modified to not only compare weighted $F_{0.5}$ scores, but also to use a one-sided t test, as identifying an improvement in the performance of the custom model over the baseline model was the purpose for these tests.

The results of these tests are shown in Table 5.15. The majority of tests here demonstrate sufficient evidence that the performances of nearly all of the custom models in terms of weighted $F_{0.5}$ scores are better than performances from baseline classifiers for the same datasets. The two models for which tests did not reach significance were a model trained and tested on data from *Hallelujah* and a model trained and tested on data from *Into Dust*. Even though results for these models did not reach significance, the results do still show a marked difference between the performance of the custom classifiers and the baseline classifiers. As expected, when the amount of data available for training increases, the performance of the custom model over the baseline model tends to improve. These results demonstrate that continued work in this area may result in the ability to characterize a given participant’s (or user’s) likelihood to respond in reliable ways to a given stimulus.
Table 5.15: Tests of significance for improvement in refined custom model classification performance over baseline classifiers for *Hallelujah*, *Raining Blood*, and *Into Dust*.

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Test Data</th>
<th>Custom</th>
<th>Baseline</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hallelujah</em></td>
<td><em>Hallelujah</em></td>
<td>.652</td>
<td>.581</td>
<td>1.853</td>
<td>.062</td>
</tr>
<tr>
<td><em>Into Dust</em></td>
<td><em>Into Dust</em></td>
<td>.681</td>
<td>.550</td>
<td>1.741</td>
<td>.071</td>
</tr>
<tr>
<td><em>Raining Blood</em></td>
<td><em>Raining Blood</em></td>
<td>.596</td>
<td>.474</td>
<td>2.596</td>
<td>.024</td>
</tr>
<tr>
<td><em>Hallelujah</em></td>
<td>Others</td>
<td>.631</td>
<td>.542</td>
<td>2.466</td>
<td>.028</td>
</tr>
<tr>
<td><em>Into Dust</em></td>
<td>Others</td>
<td>.625</td>
<td>.540</td>
<td>2.174</td>
<td>.041</td>
</tr>
<tr>
<td><em>Raining Blood</em></td>
<td>Others</td>
<td>.602</td>
<td>.522</td>
<td>2.517</td>
<td>.027</td>
</tr>
<tr>
<td>Others</td>
<td><em>Hallelujah</em></td>
<td>.626</td>
<td>.539</td>
<td>7.027</td>
<td>.001</td>
</tr>
<tr>
<td>Others</td>
<td><em>Into Dust</em></td>
<td>.603</td>
<td>.530</td>
<td>5.588</td>
<td>.001</td>
</tr>
<tr>
<td>Others</td>
<td><em>Raining Blood</em></td>
<td>.618</td>
<td>.528</td>
<td>3.902</td>
<td>.006</td>
</tr>
<tr>
<td>All stimuli</td>
<td>All stimuli</td>
<td>.595</td>
<td>.488</td>
<td>2.866</td>
<td>.018</td>
</tr>
</tbody>
</table>

5.4 SUMMARY

This chapter introduced the idea of strong responses to music, and explained what this meant in terms of analysis of data from *EiM*: examining particular moments in the musical stimuli and collected data where exceptional EDRs were seen in aggregate signals in the database. Attention was given to previous research on strong responses to music. The specific stimuli under investigation here were then characterized from a musical perspective before discussing analysis of the data for these stimuli and the results this analysis produced. These results will be considered in detail in Chapter 6, giving particular attention to their implications for future research. Following the presentation of these results, a suite of models built to predict from participant responses the likelihood of a participant presenting a strong psychophysiological response were introduced. While these models leave room for improvement, nearly all of them demonstrate that sufficient data is available apart from participant physiological responses to experimental stimuli to make these predictions with performance significantly better than chance. Potential areas for improvement of these models will be discussed in Chapter 6, as well.
DISCUSSION

This dissertation explored three research questions, primarily focused around the problem of how to develop a strategy to design and mount large studies that deal with psychophysiology, and how to best exploit the data gathered through such studies. This problem was considered with the immediate goal of using such a strategy in order to explore pressing questions of how humans respond physiologically to affective stimuli—specifically, musical affective stimuli. This research was broken into three research questions. This closing chapter considers again each of those research questions, discussing what has been learned through their investigation and providing suggestions for future work.

6.1 RQ1 REFLECTION

As stated in Chapter 1, the first research question considered by this work was:

What is an effective strategy for psychophysiological studies that simultaneously permits:

- The assembly of very large samples that cross numerous demographic boundaries,
- Data collection in naturalistic environments,
- Distributed and/or virtual study locations,
- Rapid iterations on study designs, and,
- The simultaneous investigation of multiple research questions?

6.1.1 Documenting the Issues with EiM

The challenges and shortcomings of the original Emotion in Motion (EiM) study execution strategy were documented to begin to answer this question. The first glaring issue with the original study apparatus was the use of a general-purpose programming environment for coordination and display of the study materials to the participants. Overall, a general-purpose tool did not meet this need, and by its nature, introduced inflexibility into the apparatus. It is short-sighted to think that a long-running, distributed
study designed to gather data at the scale and quality required by the affective science community will be perfect on its first iteration. Because of this the flexibility of the study design and apparatus is necessary.

Next, results gained from using laboratory-grade equipment will always outshine those gained from using cheaper hardware. It is true that the quality of some signals in the EiM database pales in comparison to what might be recorded using the best equipment available. However, the costs of using laboratory-grade equipment in the type of study this dissertation describes would be tremendous. Furthermore, ecological environments where a risk of equipment breakage and theft is real exacerbate these costs. Frankly, these situations more closely resemble the setting in which a realistic application would be deployed. Because of this, to gather databases of physiological signals the size of the EiM database, the only viable option is to work with lower-cost signal acquisition hardware. In the end, there are still thousands of superb quality signals in the EiM database, and even when signal quality is lacking, other quantitative data are still available for analysis. In order to be successful, other similar efforts the scale of EiM must find ways to leverage cheaper hardware.

Data must be managed well. File formats and syntax must be considered before embarking on a study of this size and scope. The issues with the data originally gathered for EiM meant that time was wasted in rectifying observations between iterations, data were lost, or data were ignored—all problems that were ultimately unnecessary. In order to avoid similar issues, researchers should consider not only their immediate needs, but also the needs of others with whom they collaborate, and unexpected future data requirements. Of course, as was the case with EiM, it would be impossible to predict all possible future uses of the data or further data collection opportunities. Nevertheless, the original EiM strategy and others would benefit from using flexible, parsable, and extensible data formats and syntax whenever possible.

Altogether, these issues led to a severe “observation attrition” rate across iterations of EiM. At times, physiological signals were unavailable or of poor quality. At other times, it was not possible to rectify self-report responses from a participant with their physiological signal data. At yet other times, uncontrollable circumstances meant data was lost altogether from a participant—perhaps they ceased participation before the end of the experiment. In a semi-controlled setting using low-cost hardware, most of these issues are unavoidable, though steps can be taken to mitigate them (e.g., by always having spare equipment available). However, regularly in the affective science community, results are reported that conflict with one another, and most often these studies use samples of moderate size, at best. Given the wide variability in psychophysiological response documented in this work and elsewhere, it is unquestionable
that very large and diverse samples like the \textit{EiM} database are required by researchers in order to continue to advance the field. By recognizing and addressing the limitations discussed above, the \textit{EiM} database is able to contribute an impressive sample in spite of this observation attrition.

6.1.2 Requirements for a Better Data Collection Strategy

After documenting these issues, a set of requirements for an improved approach to physiological data collection at scale was developed. It was necessary that these requirements define a system that was flexible enough to be quickly deployed to new remote or virtual study locations, be executed outside of the laboratory, be executed independent of investigator input or control, allow flexible and/or multiple study designs, permit instantaneous updates to these designs, and be extensible to various formats and modalities of stimuli.

These requirements are documented in Section 4.3.1. Briefly, these requirements demand a system that:

- Is extensible in order to allow for the presentation of at least audio and visual stimuli,
- Is easily localizable into foreign languages,
- Permits manipulation of the study procedures,
- Can be deployed to remote/online locations and administered online,
- Leverages low-cost signal acquisition hardware, and,
- Can work with any sensing hardware in which raw sensor data are accessible.

Once these requirements had been defined, they were used to develop a concrete implementation in the form of the \textit{EiM} system. This system was used in the field to successfully stage iterations of \textit{EiM} in three additional locations (Taipei City, Taichung City, and Houston). These additional iterations nearly doubled the size of the \textit{EiM} database and proved out the success of the requirements and system developed in this work.

6.1.3 Continued Improvements to the Strategy and System

In spite of the work performed here, the specific \textit{EiM} apparatus and the larger strategy it now implements could still be improved in a number of ways. The first improvement that has begun to be integrated into the \textit{EiM
study is the gathering of maximum electrodermal response (EDR) amplitudes from participants. This is a useful step in reducing interindividual variation in EDRs in order to more effectively compare electrodermal activity (EDA) between participants. Boucsein noted a number of methods for doing so, the most common of which appears to be the exposure of participants to the sound of a bursting balloon [16]. Comparisons of EDA between individual signals of participants—especially comparisons of individual EDR amplitudes—is difficult without a measure of maximum EDR amplitudes for each participant. In a recent trial iteration of EiM, the sound of a balloon bursting was added by means of treating it as an additional media stimulus, but a more flexible approach would be to use a module that allows study designers to choose from a sample of methods for eliciting and recording this data. In any case, this should be standard practice for other practitioners.

It should also be noted that equipment breakdown—both as a result of normal wear and tear and intentional damage of the equipment—has been a constant issue for EiM since the beginning of the study, as well as the theft of equipment. This was to be expected, and was included in preparation for each iteration. This also is one of the primary reasons why lower-cost signal acquisition hardware was selected for use in the study. In any data collection effort of the scope of EiM, investigators must plan for such issues. Moreover, this underscores the reality that for an effort of this size, commodity and low-cost hardware is the only option for all but the most exorbitant of budgets. Additionally, there were frequent issues with mischievous participants breaking into the terminals and attempting to corrupt the systems. This reinforces the need for study interfaces be completely locked down and backed up regularly, and the stated requirement for straightforward remote study administration (this was extremely useful, for instance, when it was necessary to rebuild a corrupted study terminal remotely).

Next, it has become clear through many analyses performed with the EiM data that while additional measures the study has probed have proven useful in revealing relationships between participants’ psychophysiological responses to music, inter- and intraindividual differences do vary widely, as documented in Chapter 2. While gathering additional information about participants (e.g., personality measures) has been useful, this variance still troubles analysis of the EiM data. This is made more problematic by the small effect sizes observed not only in the results from Chapter 5 but in other more recent analyses. It is also possible that these variances and small effect sizes, coupled with the possible oversight of latent and confounding factors, is what leads to the publication of conflicting results in other studies (see [177] and [139], for example). It is likely that the addition of
more thorough measures of personality, as well as introduction of other measures of personal and contextual factors [157] would have improved these results.

In conjunction with this, adding capabilities to the system (as opposed to only making the system extensible) to present stimuli of different modalities and collect data of different modalities, would be very useful. The work to add modules to the system to not only present video stimuli, but also to present audio and video stimuli natively in the participant’s browser, is currently underway. Additionally, a module for flexibly specifying the characteristics of sensors sending data either over Open Sound Control (OSC) or through a microcontroller board and mapping this data to a participant’s trial has been prototyped and will later be included in the system. It would also be quite useful in analysis to have physiological signal recordings from participants over the entire duration of their trial—something that the system does not currently support. Finally, while a number of questionnaire question types are included in the system (for specifying multiple-choice, forced choice, and Likert-type items, for example), additional question types would be useful. As with the above recommendations for stimuli and signal modalities, developing these would be relatively straightforward for others to do given the flexibility of the system, yet work is underway to include these capabilities in the core system.

Not only are these improvements currently being integrated into the EiM system, but similar efforts stand to benefit from considering these suggestions, as well. In the interest of transparency, it is important to mention that the data collected through EiM is not without its own issues. For one, it is certain that there are some extreme outliers in the data (it is highly improbable that any of the participants were 121-years-old!) Any good data preparation and analysis strategy will expect and be prepared to handle these outliers, but this is even more important with large datasets where long distribution tails can mask such outliers.

Difficult physiological signal data has been an issue—though also not unexpected—as well. For example, the analysis of Hallelujah data in Chapter 5 worked with a respectable sample size. This, however, was only a subset of the participants who actually heard the song as their musical stimulus (data from a total of 949 participants who heard the song has currently been processed into the database). While it is respectable that with the EiM apparatus nearly 40% of signals for a given stimulus were 90% free of any artifacts, and 363 participants is a sample size that dwarfs that of most other similar studies, it remains unfortunate that this precluded the use of data from nearly 600 participants in the analysis. Future research efforts need to consider and plan for this loss of data, but should also continue to improve means for acquiring consistently reliable signals at affordable costs. Other parallel studies by the Music, Sensors,
and Emotion (MuSE) group have worked with sensors like the Empatica E4\(^1\). Acquisition devices like this provide great signals (yet do sacrifice configurability and precise control). Nevertheless, the Empatica was a poor choice for EiM, as there was no reliably way to protect against theft of the devices. Developing means for using similar moderately-priced hardware in an ecological environment would be a boon to other efforts in the field.

Another real issue for large studies like this is the matter of translation of study materials into local languages. EiM took a two-fold approach to this issue. On one hand, local experts were always used for the translation of the study into other languages, and on the other, it was a priority (and requirement) for the system to be as flexible as possible to localization needs. Nevertheless, in environments like those in which the study was staged (often in science museums—international tourist points of interest), it was very likely that in many cases participants were forced to interact with the study in a language that was not their first. Due to this, it is quite possible that specific constructs the study measured, for instance, were not entirely accurate. It would be prohibitive to translate the study into every possible language, but in similar situations (especially if a study is to be deployed online) translations should be made available as many likely participant first languages as is feasible. Future research efforts should take care to ensure that participants are aware that the study may be available in their native language. Nevertheless, especially in order to develop samples that cut across wide demographic and cultural boundaries, this cannot be overlooked.

Music is a tremendously complex stimulus to use in psychophysiological studies. It brings with it a wealth of personal and situational factors that even the most complex study designs would have difficulty addressing. The language in which a lyric is sung, the content of the lyric, a previous memory that one has when hearing a piece of music, and many other factors combine to influence the affective effect of hearing a certain piece of music for a given person. Many other researchers have avoided this issue by simplifying their musical stimuli ([93] and [94], for example). While these studies may tease out the role of a particular tempo in affective response, for instance, they still can only approximate what might be the effect of combining one particular tempo with an actual acoustic instrument and a lyric. Nevertheless, this complexity of music was a known issue when EiM was designed. It may have been better, however, to do more to consider more personal and contextual factors in selecting and revising the stimuli used in the study. For example, while more local music was included in the iterations in Taiwan, these choices were largely based on what music was popular in Taiwan at the time. Better informed choices could likely have been made there—for instance, by interviewing contacts

---

\(^1\) https://www.empatica.com/research/e4/
in Taiwan and exploring the existing sample to determine what age range of participants would have been likely. Additionally, having finer-grained data about participants’ experience with music and music preferences would have been useful. Additionally, qualitative data and analysis are not often used in the hard sciences, but free-form or semi-structured questions about participants experiences and affective responses to the stimuli would have provided a wealth of exploitable data that the EiM database currently lacks.

In a certain sense, some of the same mistakes were made in the design of EiM as have been made in the design of the Max/MSP program (Section 4.2.1). Instead of trying to “be all things to all people” in terms of a visual programming tool as Max/MSP does, faced with the opportunity to amass an enormous sample size, EiM may have let the value of a wealth of data outweigh the need to consider carefully what pitfalls it might encounter along the way in terms of the data quality issues outlined here. As a simple example, examining average distributions of affective response to all of the stimuli in EiM makes it clear that most stimuli are positively valenced and high in arousal. Because of this, the study would have been well served to heed Sloboda’s encouragement to tune the design to elicit more negative emotions [157].

Where does this leave EiM, then? The database is quite useful for a number of focused research questions. Its real value, however, lies in the direction it can give researchers in exploring subsequent, more narrowly focused questions. Generalizable results will come from rigorously controlled, repeatable studies. With the number of relationships interacting in affective response to music, however, designing these types of studies without a good starting point can be very difficult. Its ability to serve as this starting point is where the true value of EiM lies. It provides a vast amount of data through which researchers can begin to form hypotheses around which such narrowly focused studies can be designed. Additionally, it provides a system for executing those same studies. Because of this, EiM is a means to many ends, not an end in itself.

### 6.2 RQ2 Reflection

Following this discussion of an effective strategy for studies like EiM, this work then turned to the next research question:

*How can commodity hardware and general purpose software tools be used to record, represent, store, and disseminate very large, multimodal databases gathered from psychophysiological studies that use varying study designs?*

The process of answering this question, similar to that for Research Question 1 (RQ1), was to develop a set of requirements for a data synthesis,
management, and access strategy that both addressed the outstanding issues identified with the original EiM strategy, but also was capable of:

1. Unifying, representing, and storing heterogeneous data (e.g., high-fidelity time series and questionnaire responses),

2. Capturing and communicating variations in the circumstances of data collection (e.g., varying study designs and populations),

3. Facilitating access to commonalities in data between study designs, and,

4. Allowing access to external investigators.

These requirements were not developed in a vacuum. Rather, they were the product of several years of experience with the original EiM iterations in Dublin, New York City, Bergen, Singapore, and Manila. In summary, the requirements (Section 4.3.2) are that the system

1. Allow the use of commodity software for data dissemination,

2. Provide a centralized, online data repository,

3. Permit the contribution of derived data and products of analysis back to the database,

4. Define an approach for imposing structure on data,

5. Effectively coalesce heterogeneous data, and,

6. Use semantically meaningful and flexible data structures.

Previous iterations of EiM had attempted to meet some of these requirements (without explicitly delineating them beforehand) using a general-purpose visual programming environment. Some of the tasks this study execution system did well, while it failed to meet other requirements here, specifically in these areas of data synthesis, management, and dissemination (these issues are documented in Section 4.2.2). In order to fully meet these requirements, both the data synthesis, management, and dissemination system and the data acquisition system had to be completely rebuilt. The result of this coupled with the work discussed in Section 6.1 is a flexible system that meets these requirements and is focused specifically on data for and from large-scale distributed studies of psychophysiology and affect. This system leverages common components throughout the architecture, with a specific focus on structuring data and configuration using JavaScript Object Notation (JSON) documents. In this way, to begin
working with the system—from designing a study to analyzing data collected through a study—researchers only need to understand and work with one “technology” that may be new to them. JSON softens the blow of learning a new technology in that it is quite straightforward to begin using.

This data synthesis, management, and dissemination strategy and system has been incorporated into the work of EiM over the last four years in iterations of the study in Taipei City and Taichung City, Taiwan; Houston, Texas; Bergen, Norway; and in smaller ad hoc runs of the study. While time constraints have precluded the incorporation of all collected data into the dataset, doing so will be straightforward thanks to this strategy. Furthermore, once it is included, interested researchers will immediately be able to access it and incorporate it in to their analyses.

6.2.1 Continued Improvements to the Strategy and System

While the new system is a tremendous improvement over the original, additional improvements could still be made. The opportunity to work with the system for an extended period of time has demonstrated a number of areas in which the system could be enhanced to better meet the needs not only of EiM, but of other researchers, as well. While these improvements are ongoing, the system software itself is open-sourced, and is open to other researchers to which they are able to contribute their own modifications and enhancements. The library this work has produced for handling psychophysiological data (Section 4.5.1) is also regularly used in other research and commercial applications. This usage has prompted a number of different improvements to the library.

Improvements to the synthesis approach should begin with devising a means for other researchers to not only contribute data derived from the EiM study data back to the database (which they can do now), but to also contribute the data from other studies to the database directly by using the EiM data collection system. In other words, if a study is run using the EiM system, study data could be immediately published to the EiM database. The problem to solve here is in developing a means for representation of data from a study that may be completely different in structure than data from the EiM studies, but also provides a means for establishing relationship between them, if possible. This would address a wider need in affective science research: in the overwhelming majority of studies, sample sizes are not only quite small, but there is no good means for establishing relationships between data from different studies. This would also provide a means of exposing data and methods to other researchers for verifying and repeating results. This needs to be done, whether or not it occurs in conjunction with the EiM database.
As discussed in Section 4.4.2, storing time series data in a traditional relational database management system (RDBMS) in a way that permits flexible, in-database analysis of these data is a difficult task. For this reason, databases systems built to work specifically with time series data now exist. These databases not only permit this in-database analysis, but do so in a far more efficient manner than would be possible in an RDBMS. The EiM system opted not to use a time series database primarily because this would force other researchers interested in using the database to learn an entirely new system, set of tools, and likely a new query language. Instead, the EiM database uses a compromise afforded by the MongoDB database that allows time series to exist in their own structured files (as is often done in physiological datasets), provides descriptions of these files directly in the database, and establishes relationships between these files and their other related data in the database. This approach allows for more efficient retrieval of these data than if they were to be stored in regular files on the database system’s host operating system. What this approach does not permit, however, is a straightforward means of in-database signal analysis, especially when it involves multiple signals. An improvement on this approach that leverages the capabilities of a time series database while also affording the ease of use and relationships exposed by the EiM approach would be preferable, but time constraints have not permitted this work.

An overarching issue with the synthesis, management, and dissemination system is the difficulty in convincing other researchers to work with and contribute to the database. To be fair, there has been no distinct effort to publicize the system and database until now with the publication of this work, and a companion paper submitted to an upcoming affective computing conference. However, within the MuSE group, some perspectives have been voiced with respect to why other members have been slow to engage with the dissemination system. The first of these is that members of the MuSE group already possess their own copies of the Dublin and Taiwan datasets, and have developed custom tools for working with these datasets. It is understandable that they would not have the time or interest in retooling their workflows in order to work with the centralized database. To this end, there is also a concern that, beyond a change in tooling, insufficient resources exist to train others in using the EiM database. Specifically, as the goal is to share this data with researchers in technical and non-technical fields, materials aimed at researchers not familiar with working with a resource like this database would be very useful; these are currently being developed. A specific objective of this documentation is to enable other researchers to easily subset and download the database, should they choose to work with the data outside the centralized repository.
6.2.2 “Isn’t this all just common sense?”

![Figure 6.1: “Code Quality” Reprinted from https://xkcd.com/1695/ with permission.](https://xkcd.com/1695/)

It’s certainly fair to ask whether or not the system this work has presented is all just common sense. While common sense (or, at least, “common sense” that equates to “best practices” in research and industry) does inform a number of areas of the strategy presented here and the system that concretely realizes this strategy, this is not all just common sense for three reasons. First, no other work has assembled and presented this strategy and its requirements as a collective whole, much less one that is specific to large-scale, distributed studies of human affect. Second, no other work has implemented a strategy/framework for such research into any sort of system that is readily available to other researchers. Most importantly, the experimental data that have been collected through such a system have never been shared with the wider research community, much less data at this scale and quality.
Indeed, single-purpose data collection systems are routinely designed by researchers for their own use in data collection. Rarely, if ever, are these systems open-sourced, and freely usable and extensible by other investigators, though. The quality of software and systems developed by researchers for data collection and synthesis is often poor, as well, making the task of modifying and extending them difficult (Figure 6.1). A specific effort has been made with EiM to provide a system, that is well-organized; easy to use, change, and extend; and is exhaustively documented and tested. The success of the overall EiM study effort stands as a testament to these claims.

6.3 RQ3 REFLECTION

The final research question this work explored was:

How can the most salient psychophysiological responses present in the EiM database be characterized?

Specifically, how are the psychophysiological responses themselves, the musical stimuli that elicited such responses, and the groups of participants that do or do not exhibit such responses characterized? and,

Further, can these relationships be effectively captured in a computational model?

To address this, it was necessary to explore several approaches to modeling these relationships. This work began with several straightforward statistical approaches, including exploring differences in the univariate and bivariate distributions of measured variables between the reaction and no-reaction groups, and modeling these relationships through an ordinal logistic regression. When exploring differences in distributions, a number of relationships became apparent.

6.3.1 Statistical Analysis

All of the following variables at least demonstrated the possibility of such a relationship with whether or not participants demonstrated the strong psychophysiological response under study: musical expertise, felt arousal, familiarity with the stimulus, outgoingness, fault-finding, agreeableness, stress, and neuroticism (Section 5.3.2.1). Table 6.1 summarizes these general relationships. Generally, these results suggested that participants who were more likely to experience a strong psychophysiological reaction to the stimulus studied here may be less musically experienced, less familiar the stimulus, and more neurotic. These relationships deserve closer investigation.

Analysis of bivariate distributions produced equally interesting potential areas for future work (Section 5.3.2.2). This analysis showed that, out
of all the variables measured in EiM, a small family of variables were involved in all of the important relationships found. All of these important relationships dealt in one way or another with how one of the following variables varied with another: felt arousal, age, artistic, concentration, engagement, imagination, song like/dislike, felt valence, and felt tension. Moreover, a large majority of these relationships dealt in one way or another with concentration, artistic, and/or imagination. Given that artistic and imagination both deal with the personality trait of openness to experience, the relationship between openness to experience and concentration seems especially ripe for future exploration.

Table 6.1: Summary of differences between Hallelujah reaction and no-reaction groups.

<table>
<thead>
<tr>
<th>The reaction group…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
</tr>
<tr>
<td>…rated themselves as less agreeable.</td>
</tr>
<tr>
<td>Fault</td>
</tr>
<tr>
<td>…rated themselves as more likely to find fault in others.</td>
</tr>
<tr>
<td>Felt arousal</td>
</tr>
<tr>
<td>…experienced lower felt arousal.</td>
</tr>
<tr>
<td>Musical expertise</td>
</tr>
<tr>
<td>…were less musically experienced.</td>
</tr>
<tr>
<td>Neuroticism</td>
</tr>
<tr>
<td>…rated themselves as more neurotic.</td>
</tr>
<tr>
<td>Outgoing</td>
</tr>
<tr>
<td>…rated themselves as less outgoing.</td>
</tr>
<tr>
<td>Stimulus familiarity</td>
</tr>
<tr>
<td>…were less familiar with the musical stimulus.</td>
</tr>
<tr>
<td>Stress</td>
</tr>
<tr>
<td>…rated themselves as less relaxed.</td>
</tr>
</tbody>
</table>

As with the EiM system, these results are not an end in and of themselves. Rather, these results are merely a starting point, as is the entire EiM database. There do not seem to be any other published studies that consider these larger relationships across constructs of participant demographics, personality characteristics, self-reported affective response, and psychophysiology. In the spirit of [173] and [174], this dissertation does not argue for the “scientific significance” of these potential relationships. Rather, it suggests that exploring these relationships together may be a promising area for future research. Only repeated studies that confirm the presence of these relationships would bear out any scientific significance to which these results hint. As such, future research should develop more focused, planned studies (as opposed to exploratory studies) that consider all of these variables in tandem (self-reported affective response to music, various measures of participant demographics and music tastes/experience, and participant personality).
While these analyses did uncover potential fruitful areas for continued research, they did not reveal any wide dissimilarities between the two groups. A principal component analysis (PCA) aided in understanding the variability within the dataset under study. (Again, it is important to understand that this PCA included all Taiwan and Dublin Hallelujah participants, without any segmentation by reaction/no-reaction group.) This perspective on the data made it clear that variability within the dataset was not concentrated in any principal component. The first component explained just over 29% of the variability in this dataset, in fact. Interesting relationships were observed here, as well: in general, between variables measuring self-reported affect and personality, engagement, familiarity, song like/dislike, and felt valence explained the most variance in the largest component. Interestingly, familiarity varied inversely with the other three. These data showed that, overall, the less familiar one was with this particular stimulus, the higher they rated their engagement, liking of the song, and felt valence.

6.3.2 Modeling Approaches

The most helpful result of the PCA was to demonstrate a wide dispersion of variability in the Hallelujah data. Because of this, it was evident that an approach to modeling the reaction/no-reaction groups would need to be able to segment these groups in spite of this dispersion of variability. Thus, this work then explored a number of machine learning-based approaches, all centered around constructing and training neural networks as representations of these models. Many modeling approaches, especially those working with high-dimensional data, have turned to neural networks for similar applications, as they often demonstrate superior predictive power over simpler models. Unfortunately, the trade-off for this predictive power is a model which is difficult to explain and interpret. In grappling with this, [124] and others are actively exploring means for practitioners to better explain their models. However, an effective way of doing so is not yet known. Nevertheless, this work did demonstrate that neural network models of the relationships between participant characteristics and reaction groups are able to at least outperform baseline classifiers in determining, based solely on participant characteristics and self-report, whether or not a given participant is more or less likely to demonstrate a strong physiological response to a selected musical affective stimulus.

Prototype neural networks were first constructed using standard models available in Tensorflow\(^2\), which differed only in their architectures and hyperparameter values (not in their input features). Both of these initial

---

\(^2\) [https://www.tensorflow.org/](https://www.tensorflow.org/)
models gave a better weighted $F_{0.5}$ score than the baseline classifier that was configured to always predict the majority class. After working with these initial models, a least absolute shrinkage and selection operator (LASSO) regression was performed to understand what the important features of a simple linear classifier of these data would be. This regression again confirmed that wide dispersion of variability in the data remained when the model was specifically attempting to discriminate between reaction/no-reaction groups (as opposed to the variability observed from the PCA, which did not consider reaction group). One feature dominated all others here when used in a simple linear model—whether or not the participant did not select any particular music preference. It would be interesting to explore both whether or not selecting no musical preference or whether or not the music is in the listener’s native language has an effect on their engagement with the music.

The next attempts to model reaction groups took the feature importance from LASSO regression in determining what variables to use as input features. A model with a small number of features was constructed, trained, and tested first. This was followed by models that integrated successively more features in order of their importance as indicated by the LASSO regression. The result of this work was the understanding that, for these data, models performance tended to increase as the number of input features increased. All of the models constructed with this approach produced higher $F_{0.5}$ scores than the baseline classifier.

Finally, this work explored different ways in which information from models constructed using participant data from one song were transferable to participant data for other songs. For example, one question it explored was whether or not a model trained on Hallelujah participant data could be used to predict reaction groups for participants that listened to Into Dust. New models for this investigation were constructed using all available measured variables. Tests for significant differences between the trained models and the baseline classifier were performed, and all but two of these models demonstrated a significant performance improvement over the $F_{0.5}$ score of their associated baseline classifiers (these were the model trained and tested on Hallelujah data alone, and the model trained and tested on Into Dust data alone). In general, the conclusion here is that, as is often the case with complex models such as neural networks, the larger the magnitude of input data in terms of both observations and input features, the better the performance of the model. Interestingly, the models that performed best were those that were used data from two stimuli for their training and validation sets, and used data from an unseen stimulus for their test sets. While the actual $F_{0.5}$ scores from these models are not particularly impressive, this does demonstrate that information does exist in
data outside of participant physiology itself to indicate whether or not they will demonstrate particularly strong psychophysiological responses.

Still, there is room for improvement of these models and potential future research directions. For one, there is a wealth of data that these models did not integrate. One approach that may be particularly auspicious would be to integrate the control physiology from participant data into these models. Specifically, work is currently underway to integrate features derived from pulse oximetry (POX) and EDA signals recorded during the control stimulus into convolutional neural networks (CNNs). It is likely the performance of these models will show a substantial improvement over those presented here. It would additionally be interesting to integrate POX data recorded during the experimental stimulus in a similar fashion. Finally, it is very probable that the feature engineering performed in constructing these models failed to find predictive features in the data that do exist. Feature engineering is more art than science, and additional exploration of the infinite number of potential features in these data may lead to superior models.

It is important to note that while research activities around human affective response to music has flourished in the last several decades, the preponderance of such efforts overlook one of the most fundamental aspects of music and affect: both music and affect are fundamentally temporally dynamic in nature. Many previous studies consider summative aspects of one or both of music and affect—music is either labeled, for instance, with affective descriptors for an entire selection; affect, as another example, might be measured using ratings of a single dimension for an entire musical example. Exceedingly few, if any, studies have considered continuous measures of affect as they relate to the continuous features of music. Music should prove useful within affective computing not only to better understand and interpret the affective state of the user, but also to provide meaningful affective feedback through music itself. To provide the most meaningful and useful contributions to the affective computing literature, though, it is imperative to consider carefully the relationships between music and affect as they change in time. This work takes a step in that direction, by considering unique moments within stimuli, but these dynamic relationships still need to be considered more completely.

6.4 Notes on Affect, Arousal, Music, and Future Work

It is worthwhile to note that while the EiM database contains measurements of self-reported affective response and recorded psychophysiology, the degree to which affective response can be inferred solely through examination of one or more psychophysiological signals remains an open question not only in affective computing, but in the affective sciences at
large. Indeed, evidence is available (Chapter 2) that psychophysiology can be used to differentiate affective responses, but it is at least still unclear whether or not psychophysiology as a sole probe into affect can provide insight into the contextual, personal, and environmental factors that may influence one’s affective responses.

Additionally, the analyses in Chapter 5 dealt only with EDA signals. EDA has long been known to be an excellent indicator of autonomic nervous system (ANS) activity—specifically, of arousal. However, arousal is only one component of the dimensional model of emotion that this dissertation has employed. Recent work has shown that EDA may contain information about affective valence after all [56], but the significant physiological responses considered in Chapter 5 are primarily classic indicators of affective arousal. Being able to predict one’s propensity for sharp increases in arousal may be useful, but robust applications in affective computing would benefit from an awareness of valence in addition to arousal. When using psychophysiology as a lens into affect, multiple fused psychophysiological signals are more effective for determining valence, and this is not only why Schuller recommended the collection of multimodal measures [150], but also why the EiM database includes more than EDA signals.

Above and beyond the analyses presented in this dissertation, music remains an excellent tool for eliciting human affective responses of different qualities and intensities. Though these emotional experiences may most often be non-referential, as Meyer observed [115], the responses themselves are nevertheless not unlike responses to referential emotional experiences [101]. Because of this, datasets like the EiM dataset provide invaluable tools for continuing to explore the relationships between affect and physiology, in order to inform progress in affective computing.

The analyses presented in Chapter 5, while thorough, are exploratory in nature. Nevertheless, their nascence suggests areas ripe for future exploration. This will be true of most analyses performed with data from a dataset generated by studies like EiM where a conscious balance has been struck between study complexity and the resulting sample size. In order to generate the large and broad dataset that was sought in developing EiM, it was necessary to simplify the study design itself, in order to survey a larger and more diverse population. Specifically, the areas of future work uncovered by the analyses in Chapter 5 revolve around personality, more finely focused musical stimuli, and more robust models of participant ambulatory physiological responses.

The battery selected for probing participant personality was an abbreviated version of the standard Big Five Inventory-44 (44-item) (BFI-44), as detailed in Chapter 3. Because of the settings in which EiM was staged, it was important that the overall duration of one’s study participation be short enough to retain their interest in the study and continue participation
through to completion. Even with the shorter Big Five Inventory-10 (10-item) (BFI-10), participation in the study required around fifteen minutes. The selection of the BFI-10 necessarily means that personality measures present in the EiM dataset are less robust than they would be had the BFI-44 been used. In spite of this, the results from Chapter 5 indicate that there may be relationships between personality and human response to musical affective stimuli that deserve more focused attention—particularly the ways in which neuroticism and openness to experience influence one’s response to musical stimuli, and the ways in which these traits interact with other participant characteristics (such as attentiveness and demographic characteristics). Future work should consider these relationships, and should do so with more thorough measures of participant personality, especially.

There are many next steps that should be considered with respect to the musical stimuli themselves. This chapter has already noted the negatively skewed responses to the EiM stimuli in terms of affective valence, and that further attention should be placed on better targeting of negatively valenced responses. It is possible that the selection of stimuli that come from genres of music specifically designed to elicit target emotions would be more effective in this regard. Specifically, responses to selections drawn from genres of music such as film music, program music, and Musak should be considered. Additionally, while emotional experiences in response to music are largely non-referential, a great many potentially informative variables were deliberately excluded (for the same reasons prompting the selection of the BFI-10) from the designs of EiM iterations that would provide insight into the relationships this dissertation has considered. Context, situational factors, and participant experience with the music beyond simple familiarity and genre preferences would likely provide deeper insights into participant responses. Questions along these lines may also help to explain why responses have been negatively skewed. Many of these types of questions are better gathered through short-form free-text responses and semi-structured interviews, rather than via rating scales and the like. Generally, future extensions of this work would be well-served by a smaller, more focused selection of stimuli and much deeper probes into participant self-reported affect.

Finally, the models presented in Chapter 5 were also exploratory, though effective. These models present a number of opportunities for further extension and improvement. While the models attempt to capture one particular feature of psychophysiological response, there is a wealth of other data in the EiM dataset that could be used to improve their performance. Specifically, psychophysiological responses to the control stimuli should be incorporated. The inclusion of summative features from these responses should be considered for use as inputs to these models, and it would likely
be fruitful to consider additional model architectures. In particular, the physiological signals themselves that were recorded in response to control stimuli should be integrated into the models through a CNN section added to these models. These additions should include data from both EDA and POX signals.

6.5 Final Remarks

This work presents three unique contributions all aimed at the common goal of enhancing continued research in the field. First, it contributes the results just presented from modeling participant strong reactions in terms of their characteristics and self-reported responses (Chapter 5). This work demonstrated that relationships between the two do exist, else the models presented could not have performed better than a baseline model. Not only this, but ideas for what relationships are worthy of future study were outlined by the previous statistical analyses of the same data.

Second, this work contributes the flexible strategy and system presented in Chapter 4. Many years of experience with the EiM study allowed and informed the development of a set of requirements outlining a strategy for both successfully executing large-scale, distributed studies of psychophysiology and affect, but also synthesizing, managing, and disseminating the data collected through such efforts. Instead of merely presenting this strategy, a functional implementation of this strategy was realized. This allowed not only reflection on the strategy and suggestions for its improvement, but also provides a concrete system that other researchers can leverage immediately.

Finally, and most importantly, this work presents the EiM database itself (Chapters 3 and 4). The database was assembled as the tangible output of the flexible system, and provides the input to the analysis and modeling efforts described in Chapter 5. No other database of human psychophysiological responses to musical stimuli approaches the EiM database in terms of size and scope. Results like those presented in Chapter 5 are only possible with databases like this one, as Schuller rightly noted [150].

As this chapter has noted, none of the above are ends in and of themselves. Rather, they are means to many ends. For example, it would be wonderful if others would note the results presented with respect to strong psychophysiological responses, use the database to confirm or explore these results further, and use the system to mount focused studies. Each contribution here serves as a link in this chain, and if Schuller was correct, marked advances in the field cannot occur without the contributions of this work.


[82] W. James, “What is an emotion?” Mind, vol. os-IX, no. 34, pp. 188–205, 1884. doi: 10.1093/mind/os-IX.34.188.


Perspectives,” in The Oxford Companion to Emotion and the Affective

generated by music: An exploratory study of music experts,” Musi-
10298649020050S106.

[150] B. Schuller, “Multimodal Affect Databases: Collection, Challenges,
Library of Psychology, R. A. Calvo, S. K. D’Mello, J. Gratch,

[151] R. E. Schumacker, Using R with Multivariate Statistics. Los Angeles,

tional’ data to improve learning in pervasive learning environ-
org/563e/3f5ead81b9491c7622695611a44b8fa9870d.pdf.

Individual differences in aesthetic chills and other unusual aesthetic
experiences,” Psychology of Aesthetics, Creativity, and the Arts, vol. 5,

in Cognitive Bases of Musical Communication, M. R. Jones and S.
Holleran, Eds., Washington, DC, USA: American Psychological

[155] ——, “Music structure and emotional response: Some empirical

of Music and Emotion: Theory, Research, Applications, P. N. Juslin and
J. A. Sloboda, Eds., New York, NY, USA: Oxford University Press,
2010.


Figure A.1: Screen shot of welcome screen.

Welcome to Emotion in Motion

Start Experiment

Emotion in Motion

SENSEORS
MUSIC

Emotion in Motion

Select Language

152 APPENDIX A: EIM SCREEN SHOTS
Consent Form

This experiment looks at how music can influence your emotional state, through on-body sensors and self-report questionnaires. The experiment will last approximately 12 minutes. If you have any questions please ask a member of the staff or contact the researchers at the email address provided below.

By clicking the 'Continue' button below you agree to participate in this experiment, that this agreement is of your own free will and that you have had the opportunity to ask any questions you may have about this study.

You may also quit the experiment at any time by clicking the 'Emotion in Motion' link at the top-left corner of the screen.

Any personal information provided by you will be kept strictly confidential and you will not be personally identifiable by any information provided.

Should you require any further information please contact the researchers conducting this experiment (Benjamin Knapp, Javier Jaimovich, and Brennon Bortz) at emotion.in.motion@musicsenseemotion.com. If you have any questions regarding your rights as a human research participant, please contact Dr. David Moore at Virginia Polytechnic Institute and State University’s Institutional Review Board at moored@vt.edu.

If you consent to the above terms, click 'Continue'. Otherwise, click 'Cancel Experiment'.

Figure A.2: Screen shot of implicit consent form.
Figure A-3: Screen shot of demographics questionnaire.
Additional Questions

How well do the following statements describe your personality?

<table>
<thead>
<tr>
<th>Statement</th>
<th>Disagree strongly</th>
<th>Disagree a little</th>
<th>Neither agree nor disagree</th>
<th>Agree a little</th>
<th>Agree strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see myself as someone who is reserved.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who is generally trusting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who tends to be lazy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who is relaxed, handles stress well.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who has few artistic interests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who is outgoing, sociable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who tends to find fault with others.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who does a thorough job.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who gets nervous easily.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I see myself as someone who has an active imagination.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure A.4: Screen shot of abbreviated Big Five personality battery questionnaire.
Appendix A: EIM Screen Shots

Media Questions
This questionnaire uses some simple scales to find out how you responded to the media excerpt. We will compare your responses to the biosignals that we measured as you were listening.

How involved and engaged were you with what you have just heard?

1. Not at all engaged, my mind was elsewhere.
2. I was engaged with it and responding to it emotionally.

How positive or negative did you just hear make you feel?

1. Very negative
2. 3. 4. 5. Very positive

How active or passive did you just hear make you feel?

1. Very sleepy
2. 3. 4. 5. Very lively

How in control did you feel?

1. Weak (without control, submissive)
2. 3. 4. 5. Empowered (in control of everything, dominant)

How tense or relaxed did you feel while you were listening?

1. Very tense
2. 3. 4. Very relaxed

How much did you like/dislike what you have just heard?

1. I hated it
2. 3. 4. 5. I loved it

Figure A.5: Screen shot of questionnaire following media excerpt playback.
Musical Background

How would you rate your musical expertise?

- ○ No expertise
- ○ whatsoever
- ○ An expert

Do you have any hearing impairments? (If so, you may still participate in the experiment!)

- ○ Yes
- ○ No

Continue

Figure A.6: Screen shot of musical background questionnaire.
Figure A.7: Screen shot of final questionnaire.
APPENDIX B: EXAMPLE DATA FILES

B.1 trial document

Listing B.1: Example trial document.

```json
{
  "_id": ObjectId("570eb7bf83a73509d0e04f0f"),
  "metadata": {
    "location": "taichung_city",
    "language": "en",
    "terminal": NumberInt(4),
    "session_number": "69d67ac2-6c74-46bf-82fe-748304910aeb"
  },
  "media": [
    ObjectId("56ccaa8d4c6ac453fb18ec7"),
    ObjectId("56ccbb59d4c6ac453fb18f02")
  ],
  "date": ISODate("2015-08-13T06:09:12.425+0000"),
  "experiment": ObjectId("570eae52b74655eacecf892e"),
  "answers": {
    "ratings": {
      "power": [
        NumberInt(3),
        NumberInt(3)
      ],
      "like_dislike": [
        NumberInt(3),
        NumberInt(3)
      ],
      "tension": [
        NumberInt(3),
        NumberInt(4)
      ],
      "engagement": [
        NumberInt(2),
        NumberInt(5)
      ],
      "activity": [
        NumberInt(3),
        NumberInt(4)
      ],
      "positivity": [159
    ]
  }
}```
B.2 signal DOCUMENT

Listing B.2: Example signal document demonstrating document relationships.

```json
{
  "_id" : ObjectId("54111f5f08ad6ee3090e9c84"),
  "label" : "R015",
  "music_styles" : [ "classical" ],
  "nationality" : "taiwanese",
  "age" : NumberInt(58),
  "emotion_indices" : [ NumberInt(1), NumberInt(7) ],
  "sex" : "male",
  "hearing_impairments" : false,
  "personality" : {
    "trusting" : NumberInt(2),
    "artistic" : NumberInt(1),
    "imagination" : NumberInt(4),
    "reserved" : NumberInt(4),
    "outgoing" : NumberInt(4),
    "lazy" : NumberInt(4),
    "stress" : NumberInt(3),
    "nervous" : NumberInt(4),
    "fault" : NumberInt(4),
    "thorough" : NumberInt(3)
  },
  "musical_expertise" : NumberInt(1),
  "concentration" : NumberInt(4)
}
```

```json
{
  "_id" : ObjectId("570fbef683a7351e9c266e0d"),
  "label" : "R015",
  "signals" : [ ObjectId("570fbef683a7351e9c266e07") ]
}
```
B.3 RAW SIGNAL FILE DOCUMENT METADATA


```json
{
    "description": "<Textual description of columns in referenced CSV file containing originally-recorded sensor data.>,
    "contentType": "text/csv",
    "encoding": "utf-8"
}
```

B.4 PROCESSED SIGNAL FILE DOCUMENT METADATA

Listing B.4: Example processed EDA signal document metadata.

```json
{
    "decomposition_tau": [
        0.001,
        0.011
    ],
    "decomposition_error": {
        "discreteness": NumberInt(0),
        "negativity": 0.0000000228038530153711,
        "compound": 0.0000001140192650768555,
        "mse": 0.0,
        "rmse": 0.0
    },
    "contentType": "text/csv",
    "tonic_trend": {
        "intercept": 1.004999905470891,
        "slope": 0.0000000014091576559104416
    },
    "description": "<Textual description of columns in referenced CSV file containing derived EDA features.>",
    "encoding": "utf-8",
    "fs": 25.0,
    "quality": 0.0004212299915753448
}
```
B.5 media DOCUMENT

Listing B.5: Example media document.

```json
{
   "_id" : ObjectId("56ccbb59d4c6ac453fb18f02"),
   "type" : "audio",
   "artist" : "Radiohead",
   "title" : "Paranoid Android",
   "label" : "T019",
   "has_lyrics" : true,
   "year" : ISODate("1997-01-01T00:00:00.000+0000"),
   "emotion_tags" : [
      ObjectId("538bd9352212e1eda2ff529b")
   ],
   "file" : ObjectId("56cca6f3d4c6ac453fb18e2e")
}
```

B.6 STUDY DESIGN DOCUMENT

Listing B.6: Example study design document (excerpt).

```json
{
   "mediaPool" : [
      ObjectId("547c92686577a50a2ebde518"),
      ObjectId("547c92956577a50a2ebde519"),
      ObjectId("547c92cb6577a50a2ebde51a"),
      ...
   ],
   "sensors" : [
      "eda",
      "pox"
   ],
   "structure" : [
      ...
   ],
   {
      "name" : "sound-test"
   },
   {
      "name" : "eda-instructions"
   },
   ...
}
"name": "media-playback",
"mediaType": "random"
},
{
"name": "questionnaire",
"data": {
"title": "Musical Background",
"structure": [
{
"questionType": "likert",
"questionId": "musicalExpertise",
"questionLabel": "How would you rate your musical expertise?",
"questionLabelType": "labelLeft",
"questionLikertMinimumDescription": "No expertise whatsoever",
"questionLikertMaximumDescription": "An expert",
"questionStoragePath": "data.answers.musical_expertise"
},
{
"questionType": "radio",
"questionId": "hearingImpairments",
"questionLabel": "Do you have any hearing impairments? (If so, you may still participate in the experiment!)",
"questionOptions": {
"choices": [
{
"label": "Yes",
"value": true
},
{
"label": "No",
"value": false
}
]
},
"questionStoragePath": "data.answers.hearing_impairments"

}]}
Some portions of the work presented in this dissertation have also been published in the following:


