TOWARD AN EMOTIONALLY INTELLIGENT PIANO: REAL-TIME EMOTION DETECTION AND PERFORMER FEEDBACK VIA KINESTHETIC SENSING IN PIANO PERFORMANCE

By

Matan Ben-Asher

A THESIS PROJECT

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science in Music Engineering Technology

Coral Gables, Florida

May 2013
UNIVERSITY OF MIAMI

A Thesis Project submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Music Engineering Technology

TOWARD AN EMOTIONALLY INTELLIGENT PIANO: REAL-TIME
EMOTION DETECTION AND PERFORMER FEEDBACK VIA
KINESTHETIC SENSING IN PIANO PERFORMANCE

Matan Ben-Asher

Approved:

Dr. Colby N. Leider
Associate Professor of Music Engineering

Dr. Shannon K. de l'Etoile
Associate Dean of Graduate Studies

Dr. Christopher L. Bennett
Research Assistant Professor of Music Engineering

Dr. Daniel S. Messinger
Professor of Psychology
Toward an Emotionally Intelligent Piano: Real-Time Emotion Detection and Performer Feedback via Kinesthetic Sensing in Piano Performance

Abstract of a Master’s Research Project at the University of Miami

Research Project supervised by Dr. Colby N. Leider
Number of Pages in Text: [124]

This research employs an agglomeration of knowledge from the fields of electrical engineering, digital signal processing, physics, computer music, music performance, music psychology, machine learning, and pattern recognition. The integration of this knowledge is applied to the technology of inertial measurement sensors in an attempt to detect musical expression and emotion in intuitive gestures based solely on kinesthetic data from piano players in real-time. A system is presented for detecting common gestures, musical intentions and emotions of pianists in real time using kinesthetic data retrieved by wireless inertial measurement sensors. The algorithm is implemented using the gesture recognition toolbox in EyesWeb software and employs the lens model of communication of emotions in music. The algorithm can detect common Western musical structures such as chords, arpeggios, scales, and trills as well as musically conveyed emotions such as cheerful, mournful and vigorous, completely and solely based on motion sensor data. The algorithm can be trained per performer in real time or can work based on previous training sets. The system presents feedback to the user by mapping the emotions to a color set and projecting them as a flowing emotional spectrum on the background of a piano roll. The detected emotion is also shown as an object floating in the two-dimensional emotion space of the adjective circle. The system was tested on a study group of pianists, detected and displayed structures and emotions, and it provided some insightful results and conclusions.
In loving memory of my grandfather, emeritus supreme court justice professor Menachem Elon, who passed away while I was writing this thesis. He inspired me to strive for the very truth and towards great achievements, but to forever maintain a practice of solidarity, justice, and human dignity. He knew very little about music and even less about engineering, but he knew a great deal about emotion.
Hence, again, it becomes possible for motion in music to imitate the peculiar characteristics of motive forces in space, that is, to form an image of the various impulses and forces which lie at the root of motion. And on this, as I believe, essentially depends the power of music to picture emotion. [Helmholtz, 1863].
TABLE OF CONTENTS

DEDICATION ................................................................. iii

PREFACE ................................................................. iv

LIST OF TABLES ......................................................... ix

LIST OF FIGURES ......................................................... x

CHAPTER

1 Introduction ............................................................ 1

2 Background .............................................................. 8
  2.1 Inertial Measurement Units (IMUs) ................................. 8
    2.1.1 Overview ....................................................... 8
    2.1.2 Accelerometers ................................................. 9
    2.1.3 Gyroscopes ...................................................... 10
    2.1.4 The APDM Software Development Kit ......................... 12
  2.2 Motion Analysis and Gesture Recognition .......................... 12
    2.2.1 Motion Capture Data ........................................... 12
    2.2.2 Similarity Measures .......................................... 14
  2.3 Bayes Decision Theory ............................................... 15
  2.4 Expressive Notation ............................................... 17
    2.4.1 Dynamic Markings ............................................ 18
    2.4.2 Articulation ................................................... 19
  2.5 Expression and Communication of Emotion in Music Performance 22
    2.5.1 Philosophical Problems and Theories of Emotions ............ 22
2.5.2 Hevner’s Adjective Circle ........................................... 29
2.5.3 The GERMS Model ................................................. 31
2.5.4 The Functionalist Perspective ................................. 33
2.5.5 The Standard Paradigm ............................................ 35

3 Previous Work ............................................................... 38

3.1 Computational and Perceptual Models ....................... 38
  3.1.1 The Todd Model ...................................................... 38
  3.1.2 The KTH Rule System ............................................. 40
  3.1.3 The Lens Model ..................................................... 43
3.2 Real-Time Machine Learning Tools .......................... 47
  3.2.1 The Wekinator ....................................................... 47
  3.2.2 EyesWeb and The Gesture Recognition Toolbox ......... 49
3.3 Previous Applications .................................................. 50
  3.3.1 The Fuzzy Analyzer ................................................ 50
  3.3.2 Mapping Emotions to Colors in Musical Performance ... 53
  3.3.3 IMUs in Piano Teaching ......................................... 54

4 Proposed System .......................................................... 58

4.1 System Design .......................................................... 58
  4.1.1 System Specifications ............................................. 58
  4.1.2 Setup and Top-Level Design .................................. 60
  4.1.3 Data Collection and Storage ................................. 63
4.2 Algorithm Description .................................................. 63
  4.2.1 Feature Extraction ............................................... 64
4.2.2 Classification .................................................... 71
4.3 Post Processing and Visual Feedback ......................... 78
  4.3.1 Post Processing ............................................... 78
  4.3.2 Visual Feedback ................................... 79

5 Evaluation ................................................................. 82
  5.1 Evaluation and Research Goals ................................ 82
  5.2 Subjects .......................................................... 83
  5.3 Experiment Procedure ........................................... 84
    5.3.1 Stage 1 – The Standard Paradigm, One Piece Different
           Emotions ....................................................... 84
    5.3.2 Stage 2 – Different Pieces, Different Emotions ........ 85
    5.3.3 Stage 3 – Free Playing, Self-report, and Listener Evaluation 86
  5.4 Tools for Classifier Evaluation ................................. 86
    5.4.1 Confusion Matrix ........................................ 87
    5.4.2 Precision, Recall, Accuracy, and Specificity .......... 87
    5.4.3 RMS Error ............................................... 88
    5.4.4 $\chi^2$ Test ............................................ 89
    5.4.5 Cohen’s $\kappa$ test and weighted $\kappa$ test ........... 91
  5.5 Results ............................................................. 93
    5.5.1 Stage 1 - The Standard Paradigm, One Piece Different
           Emotions ....................................................... 93
    5.5.2 Stage 2 – Six Pieces, Different Emotions, testing on Free-
           playing and Self-report ................................... 97
    5.5.3 Stage 3 – Performer vs. Listener vs. Algorithm ....... 101
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6    Discussion</td>
<td>107</td>
</tr>
<tr>
<td>6.1  System Performance, Limitations, and Possible Improvements</td>
<td>107</td>
</tr>
<tr>
<td>6.2  Research Discoveries and Ideas for Future Research</td>
<td>112</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>118</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Test Subjects Description</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>Stage 1 – Classification results per subject over all categories in the Standard Paradigm</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>Precision, Recall, Specificity, and Accuracy per category for Stage 1</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>$\chi^2$ test results per category for Stage 1</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>$\kappa$ test results for Stage 1</td>
<td>97</td>
</tr>
<tr>
<td>6</td>
<td>Precision, Recall, Specificity, and Accuracy per category for Stage 2</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>$\chi^2$ test results per category for Stage 2</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>$\kappa$ test results for Stage 2</td>
<td>101</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diagram of accelerometer design</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Diagram of MEMS gyroscope tuning fork design</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Opal monitors on hands.</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>First measures of Beethoven Sonata No. 8 <em>Pathetique.</em></td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Excerpt from Schubert Impromptu 142 D 935.</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>Hevner’s Adjective Circle.</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>Five components of performance expression, the GERMS model.</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>Todd simulation results.</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>Summary of KTH rule system.</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>Portraying emotions with the KTH rules.</td>
<td>42</td>
</tr>
<tr>
<td>11</td>
<td>Modified Lens Model.</td>
<td>43</td>
</tr>
<tr>
<td>12</td>
<td>Naïve Bayes Classification in EyesWeb.</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>Fuzzy Analyzer for Estimating Emotions.</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>Mapping Colors to Emotions.</td>
<td>54</td>
</tr>
<tr>
<td>15</td>
<td>IMU Data in Piano Playing Patterns.</td>
<td>56</td>
</tr>
<tr>
<td>16</td>
<td>System Diagram Illustration.</td>
<td>61</td>
</tr>
<tr>
<td>17</td>
<td>Tempo from Motion Algorithm.</td>
<td>68</td>
</tr>
<tr>
<td>18</td>
<td>Dynamics from Motion Algorithm.</td>
<td>70</td>
</tr>
<tr>
<td>19</td>
<td>Articulation from Motion Algorithm.</td>
<td>72</td>
</tr>
<tr>
<td>20</td>
<td>Tempo, Dynamics, and Articulation displayed in EyesWeb.</td>
<td>72</td>
</tr>
<tr>
<td>21</td>
<td>Display of piano roll with emotion colored background.</td>
<td>80</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>22</td>
<td>Adjective Circle on HSV Color Wheel.</td>
<td>81</td>
</tr>
<tr>
<td>23</td>
<td>Overall Confusion Matrix Stage 1.</td>
<td>95</td>
</tr>
<tr>
<td>24</td>
<td>Overall Confusion Matrix for Stage 2</td>
<td>99</td>
</tr>
<tr>
<td>25</td>
<td>Three dimensional confusion cube</td>
<td>102</td>
</tr>
<tr>
<td>26</td>
<td>Emotion paths of Performer, Listeners, and Algorithm though one minute of Classical piece</td>
<td>103</td>
</tr>
<tr>
<td>27</td>
<td>Emotion paths of Performer, Listeners, and Algorithm though one minute of Jazz improvisation</td>
<td>105</td>
</tr>
<tr>
<td>28</td>
<td>Emotion paths of Performer, Listeners, and Algorithm though 20 seconds of a Contemporary improvisation</td>
<td>106</td>
</tr>
<tr>
<td>29</td>
<td>Intensity, tempo, and articulation 3D scatter plot</td>
<td>110</td>
</tr>
<tr>
<td>30</td>
<td>3D scatter of linear combination of features projected to three dimensions using MDA</td>
<td>111</td>
</tr>
</tbody>
</table>
1

Introduction

“Music exists only at the moment of its performance” [Kierkegaard, 1843]. Although nearly two centuries have passed since this statement was first made, it seems that it is still well established that live performance is in fact the paramount musical experience. It is in live performance that the dyad between performer and audience creates that intimate setting where extreme emotions manifest in music and gestural expressions. And it is the integration and interaction of senses – sound, sight, and even touch – that the performer exploits in order to convey the verbally ineffable. It seems therefore, that while investigating the emotions aroused by music it is only reasonable to explore the relationship of motion to the musical experience.

The expressions and emotions communicated and aroused by music have been addressed from a wide range of directions, some of which are quite distinct and some that overlap. These include the fields of philosophy, music psychology, musicology, computer music, artificial intelligence, machine learning, and music information retrieval.

From a philosophical and psychological perspective, emotions in music have been of core interest for centuries and have attempted to define, model, and justify the problematic, controversial but yet undeniable expressive power of music. Some of the influential works were those of [Helmholtz, 1863],
[Seashore, 1938], and [Meyer, 1956]. However, a significant majority of this research was based on the acoustical phenomena and structural compositional elements. It is only recently that the performance aspect of music has become of interest to this kind of exploration [Juslin and Timmers, 2010]. I will review some of this in the background section.

More recently, from the motivation of expressive performance in computer music, there have been ongoing attempts to quantify and objectify the way in which human expression can be embedded in what otherwise would be a stale performance. The KTH rule-system developed at the KTH Royal Institute of Technology [Friberg et al., 2006], describes a set of rules to employ on a musical score in order for it to sound lively and expressive while being played back by a computer. This research led to an enhanced understanding of how expression is conveyed in the audio of a music performance. But it is reasonable to assume that these realizations might not be limited to the analysis of the acoustical property of music, since many of these insights describe performance rules that can possibly be picked up in the motion of the performer. Thus, the motion can be analyzed using some of the same rules that are employed on musical audio analysis. This is explained in more detail in the previous work and proposed system sections.

The more recent burst of development and utilization of machine-learning algorithms has spawned research in human gesture recognition aimed at the control of musical instruments and audio effects [Dillon et al., 2006],
[Toyoda, 2007], [Höfer et al., 2009], and [Odowichuk et al., 2011]. These publications mostly describe novel controllers and games employing gesture recognition. Utilizing the motion in musical performance has also been attempted in some studies. Friberg [Friberg, 2004] implemented a fuzzy logic analyzer that uses audio data as well as video stream of a performer to map the performance to specific expressions. This is also reviewed in detail in the previous work section.

More recently, Nicholas Gillian and Benjamin Knapp [Gillian et al., 2011a] have developed a gesture recognition toolbox for the EyesWeb\(^1\) environment the enables use of a variety of machine-learning algorithms in real-time. This environment was specially designed to explore interactive multidimensional musical interfaces and displays. For additional related work that has been performed with EyesWeb and this toolbox see [Camurri et al., 2000], [Camurri et al., 2004], [Camurri et al., 2007], [Varni et al., 2010] and [Gillian et al., 2011b]. The gesture-recognition toolbox in EyesWeb is a major infrastructure in the present research. I will go into more detail regarding some of these in the background and previous work sections. However, at this point, it is important to note that all of these developments indicate that we are approaching a point where machines can recognize and detect human intentions in a real-time environment to a level of accuracy that can make the computer an active and live participant in the music-making process.

There is a history of the study of motion of performers in the past century,

\(^1\)http://www.infomus.org/eyesweb_eng.php
but in the recent decades this is traditionally carried out by use of video-tracking systems employing passive or active markers. This setup is referred to as motion capture (mocap) and is mostly used for animation. I will review it in the background section. In addition and in parallel to these, during the past decade, the technology of inertial measurement units (IMUs) has become commercially available and accessible at any budget and scope. Sensors such as accelerometers, gyroscopes, and magnetometers are commonly integrated in almost all mobile devices, and much research is carried out with these sensors in a variety of fields from medical devices, computer games, and musical controllers.

However, to this day, and despite these advances, the use of these novel technologies has not yet been adopted by most musicians, and most performers still use instruments employing technology from decades ago. A common justification to this phenomenon is that many performers lack the bandwidth required to master additional controls [Cook, 2001], or in other words, they have their hands “tied”. This implies that these technologies have still not been implemented in musical controllers in a way that can actually be useful to create music that can be mastered, widely adopted and appreciated.

A possible reason for the deficiency of motion sensor-based controllers could be the knowledge gap in the understanding of the optimal way to employ motion-sensor technology into musical controllers in a way that a greater majority of musicians would accept. A musical controller might be easy to adapt to if instead of having to master it, the device would employ machine-learning
algorithms to detect the musicians’ intentions. This way mastering it would impose a minor requirement on the musician in terms of how much they need to alter their movements. Such a controller could be trained to detect various types of musical intentions such as patterns of playing, expression and musical emotions. Therefore, it seems that the missing link between utilizing motion sensor technology in music controllers, machine learning and interactive computer music composing and performance is some theory in the ability of a machine to detect musical intentions (just like a fellow performer does) in intuitive expressive musical gestures.

But how much information is actually conveyed through a musician’s motions, and how detailed is this information? Are these motions direct expressions of emotions, or are they merely a side effect of the physicality in controlling the instrument? Since it is reasonable to assume that it is not one or the other but rather a combination of the two, by tracking a musician’s motions, how much information can we retrieve about the musical content being displayed? Could we track musical structures such as chords and arpeggios, crescendos and diminuendos? Could we detect emotions such as anger or cheerfulness? At what resolution and accuracy could we detect this information? And moreover, assuming the musician is moved by his or her’s own performance, how would this influence our observations, and how much would this information correlate with the musical interpretation of the piece being played? These are all questions this research attempts to address.
The applications of answering these questions aside from pleasing our curiosity would enable us to design artificially intelligent musical controllers that could interact with the musician by detecting the musician’s intuitive gestures and using them to augment and control the music. This could also be used in music pedagogy and music therapy as a feedback system for expressive performance. Moreover, the evolution path of emotions through a musical piece has interest in other communities such as music informatics/retrieval, musicology, music composition and music psychology. For example, following the emotional path through a Beethoven sonata might help us better understand the tracks of the emotional “rollercoaster” we experience while listening to this exhilarating music.

In this research, I will employ an agglomeration of knowledge from the fields of electrical engineering, digital signal processing, physics, computer music, music performance, music psychology, machine learning, and pattern recognition. I will attempt to combine the knowledge in order to use the rather recent technology of inertial measurement sensors to detect musical expression and emotion in intuitive gestures based solely on kinesthetic data of piano players in real time. While preparing for this research I noticed that much of the published work in the fields of engineering is engaged in the tasks of circuit design, programing, mechanical design, and other technical issues of “getting stuff to work”, which does not leave much time for the research itself. It was therefore my secondary goal to utilize existing technology as much as possible in order to
minimize the time lost on development and maximize that spent on the research.
2

Background

The ordering of the sections in this chapter is one of several possible alternatives that exist. Nevertheless, I have chosen a layout in a bottom-up approach, from the finest physical inspection in the details of particle motion to the large-scale abstract observations and reflections on human emotions. I feel that this should make the reading experience flowing and insightful to the reader.

2.1 Inertial Measurement Units (IMUs)

In this section, I will review the topic of inertial measurement units, describe those used in this research and provide a brief overview of how they operate.

2.1.1 Overview

The inertial measurement units used in this project are body worn monitors named Opals, manufactured by APDM\(^1\). Each IMU consists of a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. The Opals transmit the measurements from these sensors wirelessly to a receiver connected to a computer at a rate of 64 samples per second. These units will be explained in the sections that follow.

\(^1\)For more information about the Opal sensors, see the APDM web site. http://www.apdm.com/.
2.1.2 Accelerometers

Accelerometers typically measure proper acceleration which is the physical acceleration experienced by an object in the frame of an observer in free-fall. Therefore, an accelerometer on the earth’s surface and in no motion will measure a constant positive 9.8 SI (m/sec$^2$) upwards and an accelerometer in free-fall towards earth will measure zero acceleration. A more general explanation to this behavior is the equivalence principle from relativity theory [Einstein et al., 1920] that shows how the effects of gravity on an object are indistinguishable from acceleration. In fact, the design concept of accelerometers is based on this equivalence since what is actually measured in them is force and not position. In order to measure the movement acceleration that is the second derivative of the body position relative to earth, this constant can be corrected by a simple calibration, by subtracting its value from our measurements.

![Diagram of typical accelerometer design](image_url)

Figure 1: **Diagram of typical accelerometer design.** A moving proof mass hung between two tethers and connected to a capacitive plate that moves between two anchored plates creates a capacitance divider. Diagram from [Analog Devices, 2008].

The modern design of accelerometers and also that of our Opals is a micro electro-mechanical system (MEMS). A typical design is based on a proof mass
connected to a tether that is anchored at the edges. The mass is free to move in a constant range between those edges. The proof mass is also connected to some capacitive plate that moves along with it between two anchored capacitive plates. This creates two sets of capacitive plates that modify their capacitance as a result of the mass’s movement. See figure 1.

2.1.3 Gyroscopes

Gyroscopes measure angular velocity in units of degrees per second. The theory behind their operation is employing the phenomena of the Coriolis effect that a mass in motion experiences when angular rotation is applied to it. This force is always perpendicular to both the motion vector and the angular velocity vector \(^2\) [Feynman et al., 1963]. The Coriolis force causes a displacement that can be measured by a capacitive sensing structure, similar to that in a MEMS accelerometer. Typically, in a method called **tuning fork configuration**, two masses are positioned in a way that they constantly oscillate in opposite directions. The masses are attached to capacitive plates that are also in motion with them. Then, when an angular velocity is applied to the two-mass system, the Coriolis force pushes each of the masses in opposite directions. This causes a displacement between the masses that is measured in a change in the capacitance between the plates connected to them.
Figure 2: **Diagram of MEMS gyroscope tuning fork design.** Two masses oscillate in opposite directions. Applying angular velocity to the two-mass system causes them to displace in opposite directions. This displacement is measured in a capacitance change between them. Diagram from [Solid State Technology, 2010].

Figure 3: **Opal monitors on hands.**
2.1.4 The APDM Software Development Kit

The APDM access point is read utilizing a set of library functions in MATLAB in a dedicated SDK. This includes functions for reading the data buffers in the access point per monitor or for all monitors at once, clearing the buffers, setting parameters such as sample rate, streaming configuration modes and data synchronization. The SDK can be used in Matlab or C++. For detailed information on the SDK and its functions see [APDM, 2011] and [APDM, 2012b]

2.2 Motion Analysis and Gesture Recognition
2.2.1 Motion Capture Data

Motion capture (AKA mocap) and motion analysis algorithms have been of deep interest in the fields of computer science and specifically computer animation for military, sports and entertainment applications for over two decades. A majority of the development in this field is image-processing-based and is engaged in detecting human activities in video streams. The most common technique of collecting motion capture [Müller, 2007] is with the use of several dozen (40–50) retro-active optical markers attached to a subject’s suit and a set of 6–12 cameras detecting their location at up to 240 frames per second. Each camera generates a 2D data stream of the marker locations. By knowing the location and orientation of the different cameras, alignment algorithms can collect the multiple sets of 2D data and reconstruct the 3D image of the motion that was tracked. The data set created by such a system can be either used immediately for animation purposes or referred to further processing such as

\[^2\text{it is a vector cross product of the two.}\]
detecting gait deviations and other movements for physical therapy. In some applications, the marker data must be converted to a *skeletal kinematic chain* by using fitting algorithms. This is useful for robustness and enables using different sensor configurations, but it is limited by the skeletal model’s accuracy in its approximation of the human body that does not always account for possible variances.

One way or another, most mocap systems produce some kind of *kinematic chain* that consists of body segments connected by joints. For a proper mathematical description of mocap data, we let $J$ denote a set of joints (such as \{left ankle, right knee\}). Then we define the *motion capture data stream* as a set of frames (also called *poses*). Each pose can be described as a matrix

$$P \in \mathbb{R}^{3 \times |J|}$$  \hspace{1cm} (1)

where $|J|$ is the number of joints. Thus, the $j$th column of $P$ is denoted $P_j$ and is in effect the 3D coordinates of that joint in that frame. So, the complete data stream can be described by

$$D : [1 : T] \to \mathcal{P} \subset \mathbb{R}^{3 \times |J|}$$  \hspace{1cm} (2)

Now $T$ is the number of poses and $\mathcal{P}$ is the set of poses. A subsequence of frames is also referred to as a *motion clip* and the curve described by a single body joint is referred to as *3D trajectory.*
2.2.2 Similarity Measures

While observing mocap data, probably the most fundamental element is finding suitable similarity measures that can be calculated by an algorithm and used to compare two or more motion clips. Now, the definition of similarity of course varies per application according to the requirements on the performance of the system. For example if the system needs to detect a state of running versus walking but disregard the subtleties of gait then we shall define all types of walking as similar as opposed to running. However, if the system is required to detect an emotional intention in motion such as aggressive or shy, then it would have to distinguish between different types of walking but not necessarily between running and walking if they were to fall under the same category (for example, vigorous walking and running would both fall under the aggressive category).

A good general rule for defining similarity is to regard two motions as similar if and only if they can be projected to have the same representation via a global transformation. The simplest example of a similarity such as this is a body performing the same motion at two different locations relative to the center of a room. By projecting both motions to an axis system centered on the body itself, the dissimilarity is eliminated and the motions can be regarded as similar. However, this measure could also be extended to any global rotation about an axis, or, in the case where size and speed are not of interest, (for example in detecting an element in sign language), the transformation could include scaling in time and space, making the similarity measure spatially and temporally
invariant.

In the scope of this project, I intend to detect emotions and intended expressions in piano-playing motion. In mocap nomenclature, this falls under the category of motion style or motion content. Various techniques such as Fourier Expansion, Principle Component Analysis (PCA) and Hidden Markov Models are employed in order to address the complex task of analyzing and synthesizing motion styles. For a comprehensive overview of how motion content and style are treated in literature see [Müller, 2007].

2.3 Bayes Decision Theory

Bayes decision theory is one of the most efficient and straightforward methods in pattern classification. It is a stochastic approach that assumes that the decision problem can be solved based on probabilistic considerations [Duda et al., 1995a]. A Bayes classifier relies on the fundamental Bayes Theorem from probability theory dating back to Reverend Bayes himself. The theorem, [Bayes and Price, 1763] determines the probability of occurrence of event $\omega_j$ given the fact that another event $x$ has occurred. This rule is an inverse version of the conditional probability rule and is depicted in Equation 3.

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$ (3)

Here, the lower case $p$ is the probability density function and the upper case $P$ is the probability mass function. Notice that the denominator in equation 3 does not depend on $\omega_j$ and is actually the sum of all density functions weighted by their
probabilities and is a constant scaling factor. Therefore, because we will only be comparing and not interested in evaluating the probabilities, in practice we can ignore the denominator and use the numerator classify the event $\omega_j$. In order to further understand how we can be apply the Bayes rule to make decisions and predict future events, we can express it in informally in English as in equation 4:

$$
posterior = \frac{likelihood \times prior}{evidence}
$$

(4)

So, by measuring the value of $x$, knowing the probability density of $p(x|\omega_j)$ and knowing the prior probability $P(\omega_j)$ we can evaluate the posterior probability $P(\omega_j|x)$, which is the probability of the occurrence of $\omega_j$ given that the value $x$ has been measured.

Now, let us assume a uniform distribution of the class state $\omega_j$, i.e. without prior knowledge it could be any state at equal probability $^3$. Then, all we need to do is get a good approximation of the likelihood $p(x|\omega_j)$. One way to do this is to assume a distribution function for $x$ (for example a Gaussian distribution), and then estimate its parameters by performing measurements of $x$ in different $\omega_j$ conditions (i.e. for all $j$) [Duda et al., 1995b]. These training measurements will give us estimations of the mean and standard deviation of our variable and that is all we need in order to calculate the Gaussian probability density function for any $x$. And since we already assumed that $P(\omega_j)$ is constant, this Gaussian function will give us the likelihood of event $\omega_j$ occurring given the

---

$^3$More knowledge however, could improve our approximation, and I will discuss this in later chapters.
value of $x$. Now, by choosing the event with the highest likelihood, we can classify which is the most likely event $\omega_j$.

Moreover, by comparing the likelihood of the different events at a given $x$ we can calculate the probability error which is, in the case of two categories, equal to the probability of the category not chosen. Thus, by selecting the category with the higher probability, we de facto minimize the error. Therefore, our final decision and error for the two-category case with equal probabilities can be given by

$$\text{Given } x: \text{ decide } \omega_1 \text{ if } P(x|\omega_1) > P(x|\omega_2); \text{ otherwise decide } \omega_2 \quad (5)$$

and

$$P(\text{error}|x) = \min[P(\omega_1|x), P(\omega_2|x)], \quad (6)$$

I will review the Naive Bayes Classifier in more detail in the proposed system section including a generalization for the multivariate $d$ dimensional feature space and $N$ categories. For a comprehensive mathematical description of Bayesian theory and see [Duda et al., 1995a] and [Duda et al., 1995b].

### 2.4 Expressive Notation

Modern Western music notation has evolved for over 400 years and has realized stages of varying levels of detail in terms of how descriptive composers have conveyed their intentions in scores. The basic score representation describes only the technical aspect of music, namely, which note is played at what time and for how long. Thus, in their primal form, music scores depicted just this. The
Baroque period was in fact characterized with very few expressive markings in the notes and often employed tablature in the form of figured bass. However, as musical instruments evolved and were perfected through the Classical and Romantic periods, as seen in the evolution from the harpsichord to the fortepiano and then on to the piano [Fletcher and Rossing, 1998], the expressive bandwidth of musicians expanded and along with it the expressive markings on piano score sheets. These markings incorporated various forms and will be discussed to an extent in this introduction. For more details and an insightful reading experience about the history and evolution of music notation see [Read, 1979a].

2.4.1 Dynamic Markings

Dynamic markings were introduced in sheet music at the beginning of the seventeenth century. The first indications were the Italian words piano (soft) and forte (loud). Further on, with the development of symphonic writing during the eighteenth century in the works of Mozart and Haydn, the dynamic markings evolved from the binary form to more of discrete degrees of loudness such as mezzo forte and pianissimo. The romantic composers of the nineteenth century (such as Wagner, Tchaikovsky and Mahler) followed by the avant-garde twentieth century composers (e.g. Berio and Stockhausen) with their demand for extreme subtlety ultimately transformed the markings to a practically continuous spectrum of dynamics. This leaves us with a very wide range from fortissississimo (ffff) (used by Tchaikovsky in the 1812 Overture) all the way down to pianissississississimo (pppppp) (used by Tchaikovsky in the Pathetique
Symphony).

In addition to the above absolute and somewhat obscure scale, the more commonly used and comprehensible dynamic markings are the relative markings crescendo and diminuendo which direct a gradual increase or decrease in loudness. When a relatively short increase or decrease is required (over just a few notes), the word markings are often replaced by the dynamic symbols > and <. These symbols are often accompanied by descriptors molto (much) and poco (little).

More specifically, due to its unique design, piano notations are even more complex and are often combined in intriguingly sophisticated ways such as the sign fp on a single note. This marking is a direction to attack with forte and then an immediate piano. This type of marking appears in various forms such as ffpp; ffmp; mfp; and fppp. Finally, accent terms are also often merged in with dynamic notation, forming marking such as fzp which means forzando followed by piano, or sfmp which means sforzando followed by mezzo piano, all variations of the fp sign. See figure 4 for an example on the use of combined dynamic markings. For more details, see [Read, 1979b]

2.4.2 Articulation

A second fundamental variant that makes music more than just a mechanical stream of sounds are the patterns of accents on which the musical expression is conveyed. An accent is an exaggerated stress upon any beat or portion of a beat. Even without extra markings, all Western music has an innate
Figure 4: **First measures of Beethoven Sonata No. 8. Pathetique.** Emphasizing the composers extended use of combined dynamic markings to direct the performer on his precise expressive intention.

pattern of accent or stress that is emphasized by the meter of the piece which is set up by the bar lines. Thus, the initial beat of any measure receives an accent, regardless of the time signature. In order to direct an accent in a place other than this, a composer uses the accent mark (>) below or above the note opposite to where the stem is.

Accent markings are divided into two categories: *percussive* attack and *pressure* attack. Percussive accents are typically used for higher dynamic levels (*mf* and louder) while pressure accents are used with lower levels of dynamics (*mp* and softer). This is comes from pragmatic reasons that accenting a note with a sharp percussive attack will naturally create a loud dynamic level whereas accenting while keeping a low dynamic level will produce a pressured accent with a soft percussive attack. The percussive accents are marked either by > or by a wedge ∧. The wedge typically directs a forceful accent and is used with only high dynamic levels (*f* and up). The pressure accents are usually marked by > or -. 
The accent in this case becomes not a sharp attack but more of a sudden leaning on the note.

It is important to differentiate between the *staccato* and accents. The *staccato* (marked by a ·) does not necessarily mean an accent on the note (in keyboard performance) but merely that the duration is shorter than the actual stem definition of the note. Thus, in fact the *staccato* is actually equivalent to a short note and a rest before the next note. In order to generate a short and accented note, the *staccato* can be combined with the accent.

If accents are intended to emphasize and prominent individual notes, slurs are designed to do the opposite, which is to act in grouping and merging of notes. The actual use of slurs has transformed over the years and varies between instrumental and vocal notation; however, it ultimately functions in guiding the performer to treat a note-sequence as a *unified melodic idea*. Much of this unifying is emotional and visual and instructs that the notes beneath the slur to be played in *legato*, smoothly connected and without breaks between them.

Figure 5: **Excerpt from Schubert Impromptu 142, D 935.** Demonstrating the use of a combination of accents and slurs to group and emphasize individual notes.

Slurs and accents are sometimes combined in a somewhat paradoxical appearance. Nevertheless, this has been used in music notation since Mozart’s piano sonatas. The combination suggests that a sequence of notes should be
played in a *legato-staccato* fashion, thus, stressing each note but also merging them and somehow emphasizing that they are a group. For an example see figure 5. The result of performing such accentuation is more perceptual than physical and therefore it will be quite interesting to see how much of it can be detected with inertial sensors. For more details on accents and slurs see [Read, 1979c].

2.5 Expression and Communication of Emotion in Music Performance

The emotional aspect of music has been studied for over a century and has been approached by researchers in the fields of music psychology, philosophy, musicology, music pedagogy, and music performance. In this section, I will attempt to review and summarize only that research that I intend to employ or observations that seem relevant to me in the context of this dissertation. Namely, the research that involves the understanding of communication of emotions in music performance and specifically that which relates to human motion. For a comprehensive review, see [Juslin and Timmers, 2010], [Gabrielsson and Lindström, 2010], and [Davies, 2010].

2.5.1 Philosophical Problems and Theories of Emotions

Before commencing this short expedition, I feel it is appropriate to point out some philosophical difficulties in the topic itself. The first obstacle that arises immediately when discussing emotions in music is the following: When we say that something expresses an emotion, we mean that it reflects a state that it feels. For example, a person’s tears could express their sadness if they are actually feeling it. So how could music, being merely organized sound and by
definition a non-sentient object, express emotion? There are, of course, several answers to this question, but there are also various consequent questions that would follow [Davies, 2010].

Addressing these problems requires a basic understanding of theories regarding the phenomenon of emotions. The core debates regarding emotions in the world of philosophy and psychology have historically been over the extent of emotions being bodily sensations, a notion dating as far back as Descartes’s awareness of perturbations of animal spirits [Descartes, 1649], or rather cognitive realizations as developed in the cognitive theory in the 20th century.

The first attempts to discuss emotions in the science of the modern age were through a biological-physiological perspective. In Darwin’s “The Expression of the Emotions in Man and Animals” [Darwin, 1874], he discussed the role of emotions in human communication for the purpose of survival. This was then followed later by the James-Lange Theory of Emotions [James, 1884] and [Lange, 1885]. These theorists maintain that emotions are perceptual experiences corresponding to triggered activities in the autonomic nervous system, and that they are caused only by physiological changes in the body. In the words of William James: “we feel sad because we cry, angry because we strike, afraid because we tremble, and neither we cry, strike, nor tremble because we are angry, or fearful, as the case may be”. These theories, and modified versions of them, are still held today to an extent, especially in the fields of neuroscience and in biofeedback research. Since the current research is of proximity to the latter, I
shall refer to it in later sections.

The developments in the fields of biology and psychology during the 20th century lead to a more complex interpretation described in the Cannon-Bard Theory [Cannon, 1929]. It was claimed and even proven to an extent that although physiological responses could cause some of the experienced emotions such as fear through the fight-or-flight mechanism, they could not account for the variety and the rapidness in which emotions are perceived. Moreover, [Bard, 1928] proved that all physiological sensations including motor information had to pass through the thalamus before being processed and interpreted to consequential actions. This makes it impossible for certain sensations to trigger direct physiological responses and then emotions without first being consciously perceived. It was therefore established that cognition generates both the physiological and perceptual manifestations of at least some of the emotions.

Another research that is important to mention here is that which initiated the Two-factor Theory of Emotion [Schachter and Singer, 1962]. This observed the emotional state of subjects that were injected with epinephrine. The epinephrine typically causes a state of arousal and bodily sensations. It was observed that in the presence of emotion evoking cues, (for example, from an actor in the room), the subjects who were unaware of the expected effects of epinephrine, attributed the physiological responses to emotions, while the subjects who were informed about the effect did not, and some of them did not even display any of the typical physiological manifestations of emotions (such as
tremors or increased pulse). This research led to an understanding that although physiological changes play a role in experiencing emotions, the role is to allow a cognitive appraisal to an event, but the interpretation of this event is what defines the subjective emotional experience. Thus, being at a high state of arousal could lead a subject to euphoria just as well as anger, depending on the cues that are available. This research was then followed by [Erdmann and Janke, 1978] and criticized in [Marshall and Zimbardo, 1979].

More topics discussed and specifically regarding emotions experienced in music, are if the existence of emotions requires an external object (real or imaginary) to which they are directed. Moreover, a distinction is made between emotions and moods that I will not go into in this framework. It is however, important to summarize that emotions in and out of the musical context are generally observed as a phenomenon that is more cognitive than mere physiological reactions such as increased heartbeat and tremors, but is not completely and solely a cognitive process. It is also important at this point, to note that even though we are addressing emotions in the musical context, we are still constrained to assign terms such as sad and happy with their usual meanings, otherwise we cannot refer to them as emotions. Therefore, when we describe an emotion expressed in music in terms of regular emotions we must also account for how it relates to its normal application. This constraint is actually quite helpful, because it rules out some approaches to the topic.

It follows that music’s expressiveness can no longer be reduced to simple
technical compositional elements such as a minor chords, since the term *minor key* cannot not equate to *sounds sad* until it is explained how the music’s modality can make true the manifestation of something actually pertaining to sadness and misfortune. Another approach that is ruled out is that which refers to music as a metaphor of expressive nature. This too is not inductive to claim anymore because a metaphor by definition is a linguistic device based on semantic relations. But music in its core cannot contain these semantics unless they are defined that way (i.e. the term *minor key* is not defined as sad even if it can trigger that emotion). Otherwise, it is like suggesting that music is a metaphor by metaphor, which does not really lead anywhere. The third approach that cannot hold is the theory of *sui generis*, which claims that music’s expressiveness is of its own kind. This is now clearly not offering a theory because based on this we cannot refer to the expressions in music as emotions. Of course, music is unique in its expressiveness, which is only to be expected due to its manifestation in a different medium than other communications, but in order to analyze this phenomenon one must address the existence of any equivalence to other biological expressiveness [Davies, 2010].

**Music as a Symbol**

There are however still various possible explanations to our problem. The first one is suggesting that music operates as a symbol [Cooke, 1959]. Thus, it can refer to an emotion and characterize it similar to the way a language does. For this to be possible, music must encompass a sort of vocabulary for expressing
emotions. Although there have been attempts to describe such as system [Jackendoff and Lerdahl, 1983] and it is clear that music is highly formed and organized, it still does not consist of the basic elements required in a meaningful language such as predications and propositional logic. Another symbolic explanation is that music refers to emotions by association due to linking of phrases and sounds to certain texts or ceremonies such as relating the organ to religion and spirituality and the trumpet to majesty and war. This might explain some referral to emotions but still cannot account for how music characterizes emotions. There are more semiotic theories regarding music’s expressive nature, but they all fail in the same way because they lack a proper account for the direct and immediate manner that music affects us. Moreover, [Raffman, 1991] claims that whatever the meaning conveyed in music turns out to be, it will not be the garden-variety of emotions. She supports this with the claim that most traditional musical theories and grammars do not consist of elements attempting to convey emotions, in contrast with linguistics and other semantic languages that are designed to portray the meanings they offer.

Another topic addressed in length, since we have defined that emotion must exist in a sentient being, is who is therefore experiencing the emotions? There are of course three main candidates: the composer, the performer and the listener, and there are theories assigning the emotions to each of these and research supporting and refuting these claims. There is also an interesting theory describing an imagined persona to which we relate the emotions expressed in
music [Walton, 1988]. This theory is widely accepted but also has some objections, again relating to the level of detail that instrumental music can convey to be enough to imagine a story contrary to film or literature. For a full review see [Davies, 2010].

The Contour Theory

This seems to be the most promising theory and it is fortunate because it is highly consistent with the scope of this research. The Contour Theory, [Kivy, 1980], [Kivy, 1989], [Davies, 1994] and also [Nussbaum, 2007] does not attempt to connect the expression in music to occurring emotions. Instead, it suggests that certain patterns, shapes or movements are experienced as expressive without manifesting actual emotions. It is observed that faces and gait can appear happy or sad without actually feeling it or even intending to convey it. Furthermore, even still images of a weeping willow seem to express an emotion. The manner in which this is carried out according to the contour theory is by secondary side effect to a primary emotion that it often accompanies. The weeping willow looks like a person bent over in sorrow. In most cases, this appeals to us as a person actually expressing and feeling sorrow and we are wired to identify to this sorrow because it serves an evolutionary purpose, as I explain in later sections.

The contour theory notion with regard to music is that it can present emotion characteristics in a similar fashion. These manifest in the dynamic structures of music, which resemble those of human behavior and movement that
are related to emotions. This is different from the previous theories, since it does not suggest that music refers or symbolizes something beyond itself. The ability to be expressive is in the features of the music itself because they resemble the features of human motion that in many cases reflect emotion!

The emotion expressed in music is not immediate like the picture of the weeping willow, because music is a temporal phenomenon, and its expressiveness unfolds with the piece. This however, makes it more powerful in expressing emotions because they too, have a temporal aspect to them. Emotions are continuous, and they evolve in time. It is therefore, this innate sensitivity that composers and performers have mastered over the years to control (or one might say, exploit and manipulate), but the expressive nature inherently lies in our ability to animate music characteristics as motion.\footnote{Davies addresses two more problems, which I will not review in this context. For more thoughts and analyses such as these, see [Davies, 2010].}

\subsection{Hevner’s Adjective Circle}

In order to have a solid working definition to the emotions we wish to observe, we must look at some emotion categories. An inclusive and thorough work that laid the ground for further understanding and is still considered a sort of \textit{ground truth} when it comes to musical emotions is Kate Hevner’s \textit{Adjective Circle} [Hevner, 1936]. It classifies the emotions portrayed in music into eight categories located on a circle where the distance between two categories on the circle implies their \textit{proximity} in terms of the emotional meaning. Hence, for example, the categories \textit{merry} and \textit{mournful} are on opposite sides of the circle.
whereas *dreamy* and *serene* are adjacent categories. See figure 6. The circle was

later updated by Schubert [Schubert, 2003] to have more categories, and some of
the terms were adjusted to fit more modern English. Another common
measurement of emotions is using Russell’s *Circumplex Model of Affect* where the
musical emotion is mapped in two-dimensional space of valence vs. arousal
[Russell, 2003]. However, the emotion categories in this model were not tailored
for musical emotions, and some are difficult to convey in music (such as sleepy or
disgust). Moreover, it is possible to map Hevner’s circle to a dimensional model
similar to Russell’s [Gabrielsson and Lindström, 2010]. In this study, I will utilize
this implicit dimensionality by addressing adjacent categories as *close* in the
emotional context.

2.5.3 The GERMS Model

Since the release of the seminal work *Emotion and Meaning in Music* [Meyer, 1956], the study of music and emotion expanded dramatically. This brought with it many advances in our understanding of how and why music moves us the way it does as well methods for measuring and evaluating the emotions aroused by music in listeners and psychological models of the techniques used by composers to stimulate these emotions. However, most of the research over the last century focused on the structure and compositional aspects of music that allow it to convey emotions. More recently and with the rapid development of computational technology that allowed a musical score to be accurately performed by a computer, an understanding has evolved that musical expression is largely in the hands of the performer and in many ways more than the composer or arranger. A computer performing a dictated score in a precise manner will still sound mechanical and unnatural such that it would, in many cases, fail to trigger the musical emotions to the same extent as an expert performer will. This implies that the performer has a crucial role in shaping the musical experience in listeners [Widmer and Goebl, 2004]. The many other manifestations of music are mere coding or other technical representations. This brings us back to the imperative comprehension that “music exists only in the moment of its performance”. Therefore, to fully understand the phenomenon of music experience we must capture and analyze the performance and not only the
The current opinion regarding music expression in performance is that it is a multi-dimensional phenomenon that can be decomposed into subcomponents each influencing the expressive character of the performance. It is very difficult to pinpoint the exact nature of each component and its contribution and their reciprocal relations. Juslin suggests a model known as the GERMS model, [Juslin, 2003] in which he describes expressive performance as deriving from five main sources:

- **Generative rules (G)** are variations in timing, dynamics, and articulations that allow a performer to highlight, accentuate or group notes and harmonic structures in a musically pleasing manner.

- **Emotional expression (E)** is the manipulation of large-scale performance features such as tempo or loudness that the performer uses in order to communicate emotions to listeners.

- **Random fluctuations (R)** reflect human motor precision limitations, even expert performers attempting to play perfectly even intervals will present minor fluctuations in timing [Gilden, 2001].

- **Motion principles (M)**. This is the notion that tempo and dynamic changes should follow natural patterns of human movement and locomotion to convey a musically pleasing shape [Shove and Repp, 1995]

- **Stylistic unexpectedness (S)** is the intentional deviation from stylistic
expectation of the performance. The performer uses this to surprise the audience, thus adding tension and unpredictability [Meyer, 1956]

These components are utilized by the performer and merged together to create an expressive and meaningful performance. It is interesting to note that each component originates from a different source, is characterized by different features and is even processed in a different region of the brain [Juslin, 2003]. Figure 7 describes these observations.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin of pattern</td>
<td>Generative transformations of the musical structure</td>
</tr>
<tr>
<td></td>
<td>Emotion-specific patterns of acoustic cues deriving from vocal expression</td>
</tr>
<tr>
<td>Nature of pattern</td>
<td>Manly overall levels of multiple uncertain, partly redundant cues that are compensatory</td>
</tr>
<tr>
<td>Salient brain regions</td>
<td>Left hemisphere (adjacent to Broca’s area)</td>
</tr>
<tr>
<td></td>
<td>Right hemisphere (the basal ganglia)</td>
</tr>
<tr>
<td>Perceptual effects</td>
<td>Clarifies structure; affects the inherent expression of a piece</td>
</tr>
<tr>
<td>Knowledge dependence</td>
<td>Medium</td>
</tr>
<tr>
<td>Aesthetic contribution</td>
<td>Beauty, order, coherence</td>
</tr>
<tr>
<td>Under voluntary control</td>
<td>Yes, mostly</td>
</tr>
<tr>
<td></td>
<td>Lear</td>
</tr>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Yes, mostly</td>
</tr>
<tr>
<td></td>
<td>Yes, partly</td>
</tr>
</tbody>
</table>

Figure 7: **Five components of performance expression according to the GERMS model.** The components are analyzed by origin, features, processing brain region, perceptual effects, knowledge dependence, aesthetics contribution, and voluntary controllability. From [Juslin, 2003].

### 2.5.4 The Functionalist Perspective

As mentioned above, the fact that music can communicate emotions from a performer to an audience is a complex psychological and anthropological phenomenon and explaining the process that enables this is problematic to say
the least. There have not been many models addressing this issue, nevertheless [Juslin, 1997] suggested a framework inspired by Herbert Spencer [Spencer, 1875]. The reason we call this a functionalist approach is that it assumes the same ideas explaining other non-verbal communications and their function in human survival.

The functionalist approach attributes the expression of emotions in music to two factors. The first factor is evidence of the existence of innate programs for vocal expression of basic emotions. According to this, humans possess an expressive code that originates from involuntary physiological changes that are associated with emotional reactions. These reactions strongly influence different aspects of voice production, see [Juslin and Scherer, 2005]. The notion assumes that the decoding and encoding of emotions is designed to account for a discrete and limited number of emotion categories. The reason for this is the requirement for accuracy in decoding the correct emotion on the expense of high resolution of emotions. The ability to quickly and correctly interpret an emotional expression is a strong advantage for survival because only then can it act as a guideline for action in essential life problems such as danger (fear), competition (anger), loss (sadness), cooperation (happiness) and caregiving (love). It is also reasonable to assume that these vocal expressions of emotion were later used in ancient ceremonies of festivals, funerals, wars, and caregiving and reflected happiness, sadness, anger, and love and through this music obtained its expressive nature. This realization has a significant implication on music performance since it
indicates that these basic emotions will be privileged over other intended emotions in our perception of music due to our biological preparedness for their effective communication. In other words, we are evolutionarily programmed to be very sensitive in detecting these emotions and are constantly searching for their patterns.

The second factor that characterizes music’s emotional expression is related to social learning. This begins in very early childhood and some research suggests even before birth [Parncutt, 2006]. But definitely a process of imprinting happens when mothers talk to their infants while trying to calm them by reducing the tempo and intensity of speech or if they want to scold or warn them using a sharp staccato and louder voice. Later on, expressive skills of actual music performance will also develop but often the performers adapt the basic expressive codes to their performing style [Timmers, 2007]. For a list of consistencies endorsing the functionalist approach of the existence of an innate code for vocal expression of basic emotions, see [Juslin and Timmers, 2010]. Juslin also suggested a model for capturing these functional relationships of decoding and encoding emotions [Juslin, 2000]. This is called the Lens Model and I will review it in the previous work section.

2.5.5 The Standard Paradigm

It is important to note here that when it comes to the exploration of musical performance, the literature dates back as far as the 18th century. Much of the first work during these periods was done by the performers and composers
themselves, stating their personal view on musical expression and even detailed description on how to enhance performance. For example, see [Bach, 1778] who wrote the *Essay on the true art of playing keyboard instruments*. This type of literature describes techniques for expression such as tempo, dynamics and ornamentation and structural composition techniques to enhance musical expression. However, all of this information concludes to personal opinions of musical experts. It is only in the past few decades that scientific research methods have been developed to actually measure and analyze the music-psychological effects of performance.

The most well developed line of research is based on the *Standard Paradigm*. Some claim that this research method originated with Seashore and his famous quote that “deviation from the exact is the medium for the creation of the beautiful, for the conveying of emotion” [Seashore and Metfessel, 1925]. Later on he suggested a paradigm used on actors who were required to express various emotions in speech. Following that, the speech was analyzed acoustically to understand the coding of these emotions. Seashore claimed this could be done for music back in the 1940s but this was not attempted until the early 90s [Juslin and Timmers, 2010].

In the *Standard Paradigm*, performers are requested to play melodies while attempting to express certain emotions selected by the researcher. The performances are recorded and later played to listeners in order to assess if they can recognize the intended emotion. Typically, the performances are presented to
the listeners in random order. The judgments are measured in terms of *forced choice*, *adjective ratings*, *free labeling*, or *continuous response*. The acoustical recording of the performance is then analyzed to discover the cues used by the performer to convey the emotions.

A well accepted technique is to ask the performer play the same melody in different expressions in order to factor out the compositional effect of the melody in conveying the expression and focus on the performance itself. However, due to the interaction of the melody with the performer and the performance, it is not always reasonable to request a performer to play with a certain expression that does not fit the melody. In addition, because this research is mainly interested in the detection of emotion by a computer and not by a human listener, and because the music itself will not be analyzed but rather the motion, the effect of different melodies is less problematic (since no one and nothing is listening to the melody). In this research, I will therefore, employ both techniques. The performers will play the same melodies in different emotions but also play different melodies that match typical emotions in order assure that they can completely engage in the emotion they are expressing.
3

Previous Work

3.1 Computational and Perceptual Models

In this section, I will review some of the computational and perceptual models related to expression of emotions in performance. I have used a combination of the knowledge gained in these models in the final design of my system.

3.1.1 The Todd Model

Presented by [Todd, 1985], this model is one of the first to attempt to design a timing structure in expressive computer performance. The model was entirely based on empirical measurement of performances and model of tonal music generation. Later on the model was modified to reflect relationships between dynamics and tempo [Todd, 1992] and is most well-known for the quote “the faster, the louder, the slower the softer”.

The main assumption in this model is based on the observation that a performer has control only over two characteristics of the music and those are duration and intensity. The pitch and musical structure are controlled by the composer, and the timbre is determined by the physics of the instrument at hand. Nevertheless, the performer manages to create an expressive performance with these two variables. The model therefore assumes that there are direct links between tempo variations (rubato) and dynamics variations to expressive performance and that these links can be drawn out by simple rules. The rules
designed are based on the *Genrative Theory* by [Jackendoff and Lerdahl, 1983]. This theory describes points of stability in the structure of music, and it is on these points that the performer leans in order to emphasize the expressive musical structure to the listener. These points are in many cases the edges of musical groupings, where it is observed that the tempo significance is minimal. This is very similar to the *phrase arch* rule in the KTH rule system described ahead. The rules of this model are rather simple. The tempo is related to dynamics in a power relationship, i.e. intensity is proportional to the squared tempo. The tempo on the other hand varies in a *rubato* according to the phrasing and hierarchical grouping of the piece. Thus, the tempo is minimal at the edges of arches (beginnings and endings of groups) while accelerating toward the center and retarding toward the end. The dynamics, in turn, follow the

![Figure 8: Todd simulation results](image)

Comparison of intensity in computer generated simulation to two human performances. The average intensity in each beat is plotted vs. the beat number (the beats are constant time frames measuring temporal metrical distances from the beginning of the piece). The bold line is the computer-generated intensity and the dotted lines are the human performances. Figure from [Todd, 1992].
tempo in the square rule mentioned above. The rules were implemented in the LISP language to produce artificially generated expressive performances of Haydn and Mozart and compared it to human generated performances. The comparison can be seen in figure 8.

It is also interesting to note that the model analogizes tempo changes and dynamics in musical expression to that of physical movement and extends this to concepts of energy and mass. There is even a suggestion that this is more than an analogy and that it originates from a neuro-physiological source in the inner ear and the vestibular cortex. See [Todd, 1992] for more details.

Todd’s model was later on evaluated by other researchers (see [Widmer and Goebl, 2004] for a summary of this) some of whom showed it was not successful in generating a performance similar to pianists. Regardless, the model was used by some to understand what cannot be explained by these simple rules and to assess some idiosyncrasies in human performance.

3.1.2 The KTH Rule System

In contrast to the empirical nature of the Todd model, the KTH rule system is based on the concept of analysis by synthesis. Hence, the rules are first designed based on theoretical framework, then the music is synthesized with them, then it is evaluated by expert listeners and the rules are adjusted accordingly. The system was initially designed by Gabrielson [Gabrielsson, 1985] and then refined and extended by Friberg and Bresin [Friberg et al., 2006]. It consists of a greater set of rules compared to the Todd model, all intended to
transform a stale technical stream of notes into an expressive musical performance. The rules depict instructions to alter timing, dynamic levels and articulations of specific events in the piece. The top-level rule scheme is a block of performance rules and its input is the nominal score and $k$ values. These $k$ values set the level of expression required for each rule. The rules are grouped in eight categories and are described in Figure 9. Each category consists of a set (one or more) rules that are designed to alter the corresponding aspect of the perceived performance.

It is also interesting to see how these rules fall into the categories of the GERMS model described in the background section [Juslin, 2003]. For example, the *phrasing* rule probably corresponds to the *Generative rules* component, whereas the *Performance noise* rule falls under the *Random fluctuations* category.

In order to generate a specific emotion or expression in a performance, a combination of rules must be applied with appropriate $k$ values. There are no “correct” $k$ values. It is entirely dependent on the requirement of the composer or arranger as to the level of expressivity and accentuation through the piece. For a detailed description of each of the groups see [Friberg et al., 2006]. There has however, been some research done on how to generate certain emotional categories using the rule set. Figure 10 depicts the qualitative ranges on some of the rules to create the emotions *Happy, Sad, Angry,* and *Tender.*
**Phrasing**
- Phrase arch: Create arch-like tempo and sound level changes over phrases
- Final ritardando: Apply a ritardando in the end of the piece
- High loud: Increase sound level in proportion to pitch height

**Micro-level timing**
- Duration contrast: Shorten relatively short notes and lengthen relatively long notes
- Faster uphill: Increase tempo in rising pitch sequences

**Metric patterns and grooves**
- Double duration: Decrease duration ratio for two notes with a nominal value of 2:1
- Inégales: Introduce long-short patterns for equal note values (swing)

**Articulation**
- Punctuation: Find short melodic fragments and mark them with a final micropause
- Score legato/staccato: Articulate legato/staccato when marked in the score
- Repetition articulation: Add articulation for repeated notes.
- Overall articulation: Add articulation for all notes except very short ones

**Tonal tension**
- Melodic charge: Emphasize the melodic tension of notes relatively the current chord
- Harmonic charge: Emphasize the harmonic tension of chords relatively the key
- Chromatic charge: Emphasize regions of small pitch changes

**Intonation**
- High sharp: Stretch all intervals in proportion to size
- Melodic intonation: Intonate according to melodic context
- Harmonic intonation: Intonate according to harmonic context
- Mixed intonation: Intonate using a combination of melodic and harmonic intonation

**Ensemble timing**
- Melodic sync: Synchronize using a new voice containing all relevant onsets
- Ensemble swing: Introduce metrical timing patterns for the instruments in a jazz ensemble

**Performance noise**
- Noise control: Simulate inaccuracies in motor

---

**Figure 9: Summary of KTH rule system.** The rules are grouped into eight categories. Musical expression can be altered by varying combinations of rules. Table from [Friberg et al., 2006].

<table>
<thead>
<tr>
<th>Overall changes</th>
<th>Happy</th>
<th>Sad</th>
<th>Angry</th>
<th>Tender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tempo</td>
<td>somewhat fast</td>
<td>slow</td>
<td>fast</td>
<td>slow</td>
</tr>
<tr>
<td>Sound level</td>
<td>medium</td>
<td>low</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Articulation</td>
<td>staccato</td>
<td>legato</td>
<td>somewhat staccato</td>
<td>legato</td>
</tr>
</tbody>
</table>

**Rules**
- Phrase arch: small to large
- Final ritardando: small
- Punctuation: large
- Duration contrast: large

---

**Figure 10: Portraying emotions with the KTH rules.** Qualitative changes of overall performance and rule quantities for portraying four different emotional expressions. Table from [Friberg et al., 2006].
3.1.3 The Lens Model

In order capture the functional relationship and the manner in which emotion is encoded and decoded in music, [Juslin, 2000] employs a model originally designed for visual perception and cognitive studies based on Brunswik’s *Lens Model* [Brunswik, 1956]. This model must account for some intriguing and somewhat confusing findings related to emotional communication in music. For example, how it is that that different performers using dissimilar instruments manage to successfully communicate emotions despite the fact that the sounds they produce vary drastically and thus offer diverse perceptual sonic cues? The technique suggested is called the *Modified Lens Model* and is shown in figure 11. On the left side of the lens, the model depicts how a performer encodes emotions in performance by manipulating a set of cues (e.g. variations in dynamics, tempo, and timbre). On the other side of the lens, the listener decodes

![Figure 11: Modified Lens Model of communication of emotions in musical performance. Figure from [Juslin, 2000].](image-url)
the expressed emotion by identifying the cues and classifying the emotion they convey. It is important to note that the employment and characteristics of the cues in the performer and listener are probabilistic and redundant. They are probabilistic in that a specific emotion will be expressed by the same cues only to the extent that the performer is consistent in using those cues and that the sampled set of cues is a sufficient representation of the performer’s expressivity that can encapsulate this distribution. They are redundant in that many times more cues are used to express an emotion than necessary in order to guarantee that the message is properly conveyed and not misinterpreted. This redundancy is crucial because of the probabilistic nature of this nonverbal communication.

For example a fast tempo will not always represent happiness, since it could also in some cases occur in anger. The way to distinguish between them is by adding more cues.

Thus, performers and listeners combine several cues together in order to successfully communicate emotions. The more cues, the more reliable the communication. It is interesting to note the analogy of this redundancy to that of the sounds of musical instruments. For example, a strong attack on the piano will produce a loud sound but also a sharper timber. I mentioned this in the difference of accent notations between the case of forte and piano where the former implies a percussive accent and the latter a pressure accent. This correlation and redundancy is shown by Juslin [Juslin, 2000] by calculating a multiple regression analysis between both sides of the lens.
The lens model is also described by some numeral factors that help further understand the nature of this communication. These factors are:

- **Achievement** \((r_a)\) is the correlation between the performer’s intention and the listener perceived emotion. This is an overall measure of the success of the communication.

- **Cue weight** \((\beta_i)\) is the significance of cue \(i\) in the encoding on the performer’s side or decoding on the listener’s side relative to the other cues. In other words, these are the “coefficients” used by the performer or listener in the communication. It is by this weighted combination that these cues are utilized and interpreted.

- **Matching** \((G)\) is the similarity between the use of cues on the performer’s side and the listener’s side. It is calculated by correlating the predicted values of the performer’s regression model to that of the listener’s regression model. This is essentially an upper limit to the **Achievement** rate since if the performer and the listeners are not using the same cues, the communication is bound to fail.

- **Consistency** \((R_e, R_s)\) is the degree of consistency in the performer’s and listener’s use of the cues. This is calculated by a multiple correlation of the performer’s intention or listener’s perception and cues. This too, sets an upper limit to the **Achievement** rate, since an inconsistent behavior on either of the side is likely to be difficult to interpret on the other side unless
by extreme chance the inconsistency is similar on both sides and this factor is dealt with later.

**The Lens Model Equation**

By employing regression models on performers and listeners, Juslin [Juslin, 2000] showed a mathematical relationship between intended and perceived emotions via the *Lens Model Equation* (LME). This equation was originally designed by Hursch and Hammond [Hursch et al., 1964] for describing cognitive judgment. The studies attempt to model the cognitive system as a statistical decision making process. The LME is described in equation 7.

\[
r_a = G R_e R_s + C \sqrt{(1 - R^2_e)} \sqrt{(1 - R^2_s)}
\]  

(7)

The equation relates the factors of consistency and matching to the final achievement in communicating emotions. The first component, also called the *linear* component, is a simple multiplication of the matching factor by the consistencies of the performer and the listener. If the communication is unsuccessful we can look which one of these three variables is the most significant in its influence on the result and then determine whether: (a) the performer and listeners are using different codes (low \( G \) value), (b) the performer is applying the code inconsistently (low \( R_s \) value) or (c) the listener is applying the code inconsistently (low \( R_e \) value). As mentioned above, these three factors set the upper limit of achievement.

The second component, also call the *unmodelled* component, is intended
to correct for the unsystematic and systematic variance that cannot be accounted for in the linear component. This could be the result of inconsistencies, order effects, distractions, memory, omission of relevant cues, and configurationally cue utilization. The \((1 - R^2)\) factors are the residual variances of the regression models and \(C\) is a factor that represents the correlation between these residuals. A high \(C\) value could imply either a common reliance on acoustical cues that was not accounted for in the model, a chance agreement in model errors, cue interaction common to both models, or non linear cue functions common to both models. It turns out however, that the unmodelled matching factor is small in music performance [Juslin and Madison, 1999].

In this research, I will use the lens model to decide on the features to calculate from the pianist motions to decode the emotions. I will then use the Bayesian decision theory to model the cue utilization process that enables the communication of emotions in music.

### 3.2 Real-Time Machine Learning Tools

In this section, I will review some of the existing platforms that allow the use of real-time machine learning algorithms.

#### 3.2.1 The Wekinator

In her doctoral thesis, [Fiebrink, 2011] researched the application of supervised learning algorithms to music performance and composition and implemented the Wekinator, a real-time interactive data mining environment
based on the *Weka*\textsuperscript{1} software from the University of Waikato and the *Chuck*\textsuperscript{2} environment developed at Princeton university.

The Wekinator is a software program that is designed to enable machine-learning algorithms to be trained and employed in real time and with a rather simple interface. It offers two main options to the user, either classification of discrete categories using AdaBoost or K-nearest neighbors or continuous output control using neural networks. The goal of this project was to allow end users to design interactive applications for performance and composition of computer music by employing supervised machine-learning algorithms.

It can handle various input sources such as a web cam, the tilt sensors in a laptop and external inputs such as audio and joysticks. The user trains the program based on selected features that are calculated on the input data and uses this to control audio or video output. The software comes with sample projects such as Max/MSP feature extraction and synthesis, Kinect and GameTrak.

The Wekinator, however, is not a comfortable environment for the development of new applications or exploration of techniques that do not already exist in it. The user has little control over the features and almost no information on the algorithms that are employed. Hence, it is more of an interactive game for playing with controllers and an introduction to machine learning rather than a research and development environment. Therefore, I have chosen not to use it in this project.

\textsuperscript{1}http://www.cs.waikato.ac.nz/ml/weka/

\textsuperscript{2}http://chuck.cs.princeton.edu/
3.2.2 EyesWeb and The Gesture Recognition Toolbox

The EyesWeb environment was developed by a team led by Antonio Cammuri in the infoMus lab at the University of Genoa [Camurri et al., 2000]. It is a visual programming, open software research platform similar to Simulink and Max but dedicated to the design and development of real-time multimodal systems and interfaces. It supports a large number of input systems including audio, video, and motion sensors as well as various standards such as MIDI, OSC, VST plugins and MATLAB.

A rather recent addition to EyesWeb is the machine-learning toolbox that was designed by a team at the Queens University Belfast [Gillian et al., 2011a]. The toolbox features an exhaustive list of machine learning and pattern recognition algorithms including *Adaptive Naïve Bayes, Artificial Neural Networks, Hidden Markov Models, Dynamic Time Warping, Fuzzy C Means, K Means, K-Nearest Neighbor*, and *Support Vector Machines*.

Each algorithm is designed as a set of blocks for data collection, training, and predicting. The inputs to the collection blocks (called *training tool*) are the data features their respective weight vector. This block merely collects the data and stores it to a file upon user request. The training block (called *train*) loads this file upon startup or user request and calculates a training model according to the algorithm it implements. This model is also stored to a file. Finally, the *predict* block loads the model upon startup or user request and classifies newly received data based on the model.
The simplicity of this design and its ease of use makes the gesture recognition toolbox an extremely handy tool while trying to rapidly develop and evaluate pattern-recognition algorithms. The fact that it works in real time makes it even more efficient for research because an algorithm can be optimized by modifying the various parameters while the data is streaming in and the feedback is immediate. An example of a Naïve Bayes Classification layout is shown in figure 12.

![Naïve Bayes Classification in EyesWeb](image)

Figure 12: Naïve Bayes Classification in EyesWeb. Notice the three major blocks marked in red: ANBC training tool, ANBC train, and ANBC Predict.

### 3.3 Previous Applications

In this section, I will describe some implementations and applications that are similar or closely related to the current research.

#### 3.3.1 The Fuzzy Analyzer

The Fuzzy analyzer is a system designed by Andres Friberg [Friberg, 2004] that implements a real-time algorithm for analyzing emotional expression in
music performance and body motion. It is mainly intended for artistic human computer performance. The algorithm uses both acoustical information and motion capture information from a camera to assess the emotional content of the performance.

The audio input is analyzed for cues similar to those in the KTH system such as tempo, sound level and articulation. This is done by calculating the RMS after applying a Hanning window to the audio signal. Then two envelopes are created by filtering this signal with a bank of cut offs of 40Hz and 1Hz. These two filters extract single tones and phrases. The crossing of envelopes defines tone onsets and offsets. On each tone, five cues are calculated: sound level (dB), instant tempo (tones/second), articulation (relative pause duration), attack rate (dB/ms), and high-frequency content (high/low energy). These cues were selected based on multiple regression analysis results performed by [Juslin, 2000].

The motion signal is evaluated for a parameter called Quantity of Motion (QoM) which is an evaluation of the total difference between frames, indicating a change and hence motion. In addition, two other motion cues are calculated, width peak to peak and height peak to peak. These reflect the extent of motion in two dimensions of the video frame.

The second step in the algorithm is a calibration step to account for variances in the instrument and performer. This is merely a normalization that verifies that the mean of the cues is zero and the standard deviation is one.

The third and final step is mapping from cues to emotions. This is done
via a *Fuzzy Logic* rules set based on data collected in previous research [Juslin, 2000]. Fuzzy logic means that the discrete threshold decision process is performed on continuous values of cues. Based on the data collected linking these cues to emotions most of the continuous cues were assigned a value of high or low, however in some cases, they were assigned three discrete levels: high, medium and low. Then, a standard logic (truth table) process is applied to decide on a classification. Figure 13 specifies the complete process on three audio cues only.

![Diagram of the fuzzy logic system](image)

**Figure 13:** **Fuzzy analyzer for estimating emotions.** An audio input is analyzed in terms of tempo, sound level and articulation. The resulting prediction of emotional expression is output in terms of three functions ranging from zero to one. Figure from [Friberg, 2004].

The system detected three categories of emotions: *happiness*, *sadness*, and *anger*. It was demonstrated in a number of installations such as *Ghost in the*
Cave and Expressiball. The system was not evaluated for validity because of difficulties in establishing a ground truth and in simulating the motion in a performance. However, the authors do suggest ways to evaluate the system in future research. See [Friberg, 2004] for details.

3.3.2 Mapping Emotions to Colors in Musical Performance

This research, performed by [Bresin, 2005], explored the relationship between colors and the perceived emotion in musical performance. The aim of the research was to come up with a mapping of HSB (Hue Saturation Brightness) coloring the different emotions aroused in music. This knowledge could help design pedagogic feedback systems for music students by providing them with live expressive evaluation of their performance. Subjects were presented with 12 music performances expressing different emotions and were asked to rate how well each of a set of eight colors and their nuances corresponded with the emotions perceived in those performances.

The emotion categories selected were: happiness, love, contentment, pride, curiosity, indifference, sadness, fear, shame, anger, jealousy and disgust. It is not clear why these emotions were selected since they do not reflect the emotions expressed in musical research and the author does not explain this. Nine musicians were asked to express these emotions in music by Haydn and Brahms played on piano, guitar, and saxophone. Then these recordings were evaluated through listening tests to find the best match for each emotion. The final set of music excerpts were
12 (emotions) \times 3 \text{ (instruments)} \times 2 \text{ (melodies)} = 72 \text{ (performances)}.

The colors presented were red, orange, yellow, green, cyan, blue, violet, and magenta, their brightness and saturation variances (saturation and brightness set to 1 or 0.5 respectively). Thus, \( 8 \times 3 = 24 \) total colors were used.

The subjects were presented with the music and the colors in randomized order and asked to set a slider based on how well the colors were associated with the music on a scale of 1 to 10. The results show some significant correlations between colors and emotional intention in the music. A summary of the results for the hue parameter are displayed in figure 14.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Hue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>0.167 (Yellow)</td>
</tr>
<tr>
<td>Love</td>
<td>0.667 (Blue), 0.75 (Violet)</td>
</tr>
<tr>
<td>Pride</td>
<td>0.167 (Yellow)</td>
</tr>
<tr>
<td>Tenderness</td>
<td>0.75 (Violet)</td>
</tr>
<tr>
<td>Curiosity</td>
<td>0.5 (Cyan)</td>
</tr>
<tr>
<td>Contentment</td>
<td>0.083 (Orange)</td>
</tr>
<tr>
<td>Anger</td>
<td>0 (Red)</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.75 (Violet)</td>
</tr>
<tr>
<td>Fear</td>
<td>0.667 (Blue)</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.75 (Violet)</td>
</tr>
<tr>
<td>Shame</td>
<td>0.083 (Orange)</td>
</tr>
<tr>
<td>Jealousy</td>
<td>0 (Red)</td>
</tr>
</tbody>
</table>

Figure 14: Mapping colors to emotions. Results for hue values that received the highest mean rating for each emotional expression. Figure from [Bresin, 2005].

3.3.3 IMUs in Piano Teaching

Research in the field of music pedagogy has also examined the option of tracking a pianist’s motion with IMUs. With the goal of enhancing piano teaching systems, currently based mainly on MIDI data, [Hadjakos et al., 2008]
explored the use of an accelerometer in a setup quite similar to what this research suggests. They designed a prototype IMU with an accelerometer and a gyroscope. They attached the sensor to three locations: the upper arm, the wrist and the hand of the right arm only. The aim of the research was only to present how piano playing patterns such as trills and scales appear in inertial measurement data.

The research focused on examining the following patterns:

- Rotation patterns displaying supination and pronation (clockwise and counterclockwise rotation of the right ulna), Tremoli and Trills (the rapid repetition of two distant or adjacent notes) and scales.

- Jump patterns displaying vertical forearm motion and flexible or rigid arm.

Some of the interesting results are displayed in figure 15. Trills and Tremoli can be played using only the fingers or using the forearm. Another variant of trills is using the upper arm. All three techniques were clearly observed in the results. Detecting these differences in performance could be useful for music information retrieval because they are often employed in specific settings. For example, finger-tremolo is generally used in soft gentle parts in contrast to forearm tremolo that is used in louder, more dramatic parts.

The data from scales playing is also interesting, the high peaks correspond to thumbs crossings, in which there is either a pronation or supination of the forearm. This typically happens in groups of three and four notes (e.g. in the C
Figure 15: **IMU Data in Piano Playing Patterns.** The dots indicate the *note on* events from the MIDI data. Playing patterns are clearly distinguishable such as pronation and supination, finger, forearm, and upper arm tremolo, scales and jumps. Figure from [Hadjakos et al., 2008].
major scale). Nonetheless, most of the single notes are also observed in the signal. Also, notice the negative drop at the end of the scale. This indicates a supination on the little finger when the player finished the scale, which is a typical expressive gesture pertaining to pianists.

Vertical forearm movement is often used by pianists to connect loud note to a soft note in the case of a resolved dissonance. This is also a very basic expressive gesture and is clearly observed in the data. Although none of this is very surprising or novel, the results of this research seem encouraging in regard to what could be accomplished with the use of inertial measurement units in piano performance.
4

Proposed System

4.1 System Design

In this section, I will review the system structure including the required specifications, the top-level block diagram of the solution, and the system setup with regard to hardware and software implementation platforms and data structures.

4.1.1 System Specifications

The system is required to function as a real-time musical gesture expression classifier that can detect intended musical expression and emotions performed by a pianist. The inputs to the system are kinesthetic data from two wrist-worn IMUs each transmitting 3-axis acceleration and 3-axis angular velocity.

Requirements

- Collect kinesthetic data from APDM sensors at a rate of 64 samples per second.

- Record data to files and use to train an algorithm and enable playback mode for demonstration and algorithm development.

- Detect and Classify in real-time the following common musical structures from each hand:

  1. scales
2. *chords*

3. *arpeggios*

4. *trills*

- Detect and Classify in real-time the following expressions in Western notation [Read, 1979b], [Read, 1979c]:

  1. Tempo: *ritardando* and *accelerando*

  2. Dynamics: *crescendo* and *diminuendo*

  3. Articulation: *staccato* and *legato*.

- Detect and Classify in real-time the following emotion categories from Hevner’s adjective circle [Hevner, 1936]:

  1. *sad/mournful*

  2. *dreamy/tender*

  3. *lyrical/serene*

  4. *humorous/playful*

  5. *cheerful/merry*

  6. *vigorous/dramatic*

- Display all expression evaluations as feedback to the performer in real-time to enable an intuitive interactive musical experience.
4.1.2 Setup and Top-Level Design

In order to perform the required tasks, I have designed a system comprised of the following hardware and software elements. Most of the elements could be substituted with similar functioning parts. However, the real-time performance and stability of the algorithm might be affected by the processing power of the PC and the operating system. Moreover, the IMU data must be synchronized at 64 samples per second with a lag of less than 50 ms to allow for proper tempo alignment and response time.

- **Hardware**
  
  IMU system including 2 Opal sensors and one access point.
  
  Celviano AP-220 Digital Piano (from Casio).
  
  Mini Super computer with 16G RAM and 2 Intel Xeon co-processor CPUs.

- **Software**
  
  Windows 7, 64 bit operating system.
  
  Matlab environment version 7.10 including the APDM SDK.
  
  EyesWeb environment version 5.2.1.

An illustrative block diagram of the system setup is presented in Figure 16.

The Opals [APDM, 2012a] are small wireless wrist worn inertial sensors that feature triaxial accelerometers, gyroscopes, and magnetometers transmitting 9 values 64 times per second. The values transmitted are:
Figure 16: **System Diagram Illustration.** The system is designed to work in real-time or playback mode.
• 3 axis proper acceleration in SI units relative to $x, y, z$.

• 3 axis angular velocity in radians per second units around $x, y, z$.

• 3 axis magnetic field projection in $\mu T$ units on $x, y, z$.

The magnetometer values are ignored in this project because they would make the training data based on orientation relative to the North and thus make the algorithm sensitive to the pianist’s position. Therefore, they will not be discussed here or henceforth.

The data arrives at the receiver via multiple ports. The motion signal data from the sensors is transmitted wirelessly to an access point for synchronization. The access point is read utilizing a set of functions in MATLAB [MATLAB, 2010] in a dedicated SDK. The MIDI data from the piano is transferred via USB to the receiver. The synchronization of Opal and MIDI data is implemented via a combined system in MATLAB and EyesWeb. The MATLAB receives the data from the Opals through the APDM access point and transmits them via OSC messages to EyesWeb. The EyesWeb receiver acquires these data from MATLAB in addition to the MIDI and Audio data directly from the piano. The MIDI and motion data are synchronized at 64 samples per second and stored to a file system. While the data are being collected, it is also passed to the classification algorithms for training and prediction. All of this happens simultaneously on parallel processors and in real-time.
4.1.3 Data Collection and Storage

The MIDI and sensor data are collected in two formats: The EyesWeb native format (.ebf), and a text file. Each row consists of sample of data (IMU measurements and MIDI commands) along with a time stamp. The .ebf files can be read directly in EyesWeb and enable the playback of data just as it was collected simulating a real-time environment. This allows for developing and evaluating the algorithm while running it on real subjects’ data. The text files can be read in any other software platform for additional offline analysis.

4.2 Algorithm Description

Six dimensional data from two sensors are collected by the system at 64 samples per second. These data are buffered into three consecutive frames of one second each. This windowing of data corresponds to the Gestalt theory of music psychology which describes how music is perceived in groups rather than single notes [Narmour, 1992]. Determining the optimal size of these chunks is not trivial and with the aim avoiding a deeper dive into the studies of music perception and auditory scene analysis, it was currently chosen by trial and error. Also, it most likely varies between musical pieces and performances. Nevertheless, a window size of 1 second corresponds to a tempo of 60 BPM, and should allow for capturing several notes in rapid playing or single notes in slower playing. It is also reasonable to assume that in most cases the musical mood will not change during this period. The reason three frames are buffered is due to short term memory considerations. It is assumed that the emotions expressed
and experienced in music preserve some continuity, evolve gradually and are influenced by recent emotional states [Meyer, 1956]. Therefore it make sense to record the data from the recent past in order to make a decision regarding the emotional state of the present. The degree of the past influence on the present is also a parameter of the algorithm and is reflected in the weight assigned to the features in the classifier. The features are calculated on each frame, except for the \textit{tempo} feature which is continuous. The features from each frame are weighed and passed to the classifier for training or prediction. I will discuss the weighing considerations in the sections that follow.

\subsection{Feature Extraction}

For each frame, the following features are calculated on the incoming signals:

- mean motion intensity (per hand per axis)
- RMS motion intensity (per hand per axis)
- variance of acceleration (per hand per axis)
- mean spin intensity (per hand per axis)
- RMS spin intensity (per hand per axis)
- variance of angular velocity (per hand per axis)
- dynamics
- tempo
In the following subsections, I will describe how each of these features is calculated.

**Motion Intensity**

Motion intensity is calculated on each hand over each axis separately via sum of squares over an entire frame, i.e. the average of the sum of squares over all accelerations as described in equation 8, per dimension (per hand per axis). This feature over the \( z \) axis should be a good estimator of the audio intensity since the acceleration is proportional to the force exerted on the keys, which is proportional to the squared velocity of the hammers striking the strings which is proportional to the kinetic energy converted into acoustic energy [Fletcher and Rossing, 1998]. The factor of two compensates for adding the absolute value of the positive and negative accelerations, which should cancel out in the case of zero total distance traveled.

\[
\text{motion intensity} = I_{a_j} = \frac{1}{N} \sum_{i=1}^{N} a_{i,j}^2
\]  

(8)

Where \( N \) is the number of samples in a frame and \( j \) is the dimension column (i.e. which hand which axis).

**Variance of Acceleration**

The variance of acceleration \( \sigma_a^2 \) is calculated in the unbiased estimator definition as described in equation 9. It is an indicator of the variation in intensity during a frame a might point to certain expressive elements such as

- articulation (per hand)
legato vs. staccato playing as well as rapidly changing vs. monotonic performance. The feature is calculated per hand per axis.

\[
\text{variance of acceleration} = \sigma_{a_j}^2 = \frac{\sum_{i=1}^{N} (a_{i,j}^2 - I_{a_j})^2}{N - 1} \quad (9)
\]

**Spin Intensity**

The spin intensity feature is extracted from the angular velocities measured by the gyroscopes. It is calculated via sum of squares over an entire frame, i.e. the average of the sum of squares over all angular velocities as described in equation 10, per dimension. It is an estimation of total supination and pronation as well as forearm tremolo.

\[
\text{spin intensity} = I_{g_j} = \frac{1}{N} \sum_{i=1}^{N} g_{i,j}^2 \quad (10)
\]

**Variance of Angular Velocity**

As with motion intensity, the variance is calculated in the unbiased estimator definition. It is an indication of the change in articulation style within a frame with regards to tremolo, supination, and pronation.

\[
\text{variance of angular velocity} = \sigma_{g_j}^2 = \frac{\sum_{i=1}^{N} (g_{i,j}^2 - I_{g_j})^2}{N - 1} \quad (11)
\]

**Tempo**

For extracting tempo out of motion, one must consider the phenomena of tempo perception in piano playing. Bruno Repp shows some insightful concepts in this specific area [Repp, 1994] and the algorithm is mostly based on his findings. The assumption is that the perception of tempo evolves from an
accumulation of temporal cues derived from the audio in the form of varying intensities. Therefore, using this, a real-time running tempo estimator can be attempted as follows. First, the envelope of the signal is calculated via RMS. Then, two thresholds are calculated based on the standard deviation of a half-second frame. For both thresholds, crossings are detected and their times registered in to a buffer. For the low thresholds the crossings are weighed by the intensity of the peak in that crossing, this is to account for the stronger pulses having a more significant influence on the perception of tempo. Then a weighted average is performed on the buffer. For the high threshold, the timing of the crossing are only registered to buffer and a median filter is applied to account for extreme outliers. Finally, the average of the two tempo estimations is calculated to provide an overall perceived tempo. The implementation of this algorithm in EyesWeb is depicted in Figure 17.
Figure 17: **Tempo from Motion Algorithm**. A perceptual model based on averaging of temporal cues.
It is important to note that the tempo feature is not calculated separately on each hand since both hands contribute to one performance and there should only be one perceived tempo. Thus, the input to this algorithm is a Pythagorean sum of all six accelerations from both hands. Using this approach, the algorithm can detect a joint tempo created with accents from both hands. For example, if a waltz in three fourths is played at 60 BPM playing the bass with the left hand and two accents with the right, then the tempo may vary depending on how this is articulated. If the bass is accentuated significantly stronger than the following notes then only the downbeat will be detected as a threshold crossing and the tempo will be assigned 60 BPM. However, if the right hand part is articulated with equal intensity then the tempo algorithm might detect those as crossings and measure 180 BPM. Intuitively this makes sense because the perceived tempo of the audio would also probably be affected by these differences in articulation and so will the expressive nature of the piece. See [Repp, 1994], [Levitin and Cook, 1996], and [Scheirer, 1998] for a broader review of the problems with definitions and perceptions of tempo and in establishing a ground truth.

Dynamics

The dynamics features are extracted by estimating a corollary to the acoustic energy. Thus, we calculate features that are proportional to mean acoustic energy, gradient of mean acoustic energy, and variance of acoustic energy. This is done by first summing the squares on the two z axis components
of the accelerations from the left and right hand. Then an envelope of that signal is calculated and buffered over time frame of one second. Then we calculate the mean and the variance. In order to detect crescendo and diminuendo we also compute the derivative of the mean. A positive derivative for a long enough time is a continuous increase in energy and therefore a crescendo. Similarly a steady negative derivative is a diminuendo. The EyesWeb implementation of extracting dynamics from motion is shown in Figure 18.

![Dynamics from Motion Algorithm](image)

**Figure 18: Dynamics from Motion Algorithm.** Estimating acoustic energy levels, variance, and gradient from acceleration in the z axis.

**Articulation**

Articulation is defined in [Juslin, 2000] and [Bresin and Battel, 2000] as the ratio between the note duration and the inter onset interval. It is an estimation of the legato in a phrase. To calculate this, two values are defined:

\[ d_{ii} \] the duration from the onset of a note until the onset of the next note.
\( d_{io} \) the duration from the onset of the note to its offset.

Then, note articulation is defined as:

\[
Articulation = \frac{d_{io}}{d_{ii}} \tag{12}
\]

In engineering terms this could be viewed as the *duty cycle* of the note in a phrase. Detecting articulation from motion in this manner would require us to detect precise note onsets and offsets from motion. Since this is difficult to achieve and although intriguing, is not in the context of this research, I have designed a different but similar approach. Instead of finding note onsets we find temporal regions of activity in the \( z \) axis signal. Since the \( z \) axis is in the direction of the key down and up, this should closely correspond to note onset and offset behavior. We count the number of activity occurrences in a time frame of half a second and divide it by the total time. An articulation value close to one represents *legato* and an articulation value close to zero corresponds to *staccato* articulation. The implementation of the articulation algorithm in EyesWeb is described in Figure 19. The display of tempo, dynamics, and articulation in EyesWeb is shown in figure 20.

4.2.2 Classification

Following feature extraction, the features are fed to an Adaptive Naïve Bayes Classifier (ANBC) along with a weight matrix that defines the weight for each feature. The ANBC is a part of the gesture recognition toolbox in EyesWeb [Gillian et al., 2011b]. In this section, I will review how this block operates,
Figure 19: Articulation from Motion Algorithm. Utilizing high and low activity detection on the \( z \) axis as estimations of note onsets and offsets.

Figure 20: Tempo, Dynamics, and Articulation displayed in EyesWeb.
including its features and some of its advantages and drawbacks.

As described in the background section, the design of any classifier employing the Bayes decision theory is structured on the Bayes theorem (equation 3) and some naïve assumptions regarding the probabilistic nature of the features and classes. One assumption is that the features are independent random variables. Another assumption is that their distribution (i.e. their probability density function) can be approximated to be of Gaussian nature. Note that neither of these assumptions is prerequisite for Bayes decision theory to be optimum. Bayes theory is general and optimum for any random variables in any distribution as long as the distribution is correct. However, assuming these distributions and independent variables significantly simplifies calculations and programming complexity.

Rewriting the Bayes theorem for gesture recognition we define the likelihood (or conditional probability) of gesture $g_k$ being the state of nature out of $G$ gestures given an observation of the input feature vector $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ in equation 13.

$$P(g_k | \mathbf{x}) = \frac{p(\mathbf{x} | g_k) P(g_k)}{\sum_{i=1}^{G} p(\mathbf{x} | g_i) P(g_i)}$$

(13)

Notice that just as in equation 3, the denominator is constant because it sums over all gestures and therefore can be ignored for classification purposes. Also note that because $\mathbf{x}$ is a vector of $N$ variables, $p(\mathbf{x} | g_k)$ is a function of $N$ parameters. If we assume that the features abide to a normal distribution then this function becomes the multivariate Gaussian distribution function.
[Duda et al., 1995a] defined in equation 14.

\[ p(x|g_k) \sim \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left( -\frac{1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) \right) \] (14)

Where \( \mu \) is an \( N \) dimensional mean vector of the features in a class defined in equation 15.

\[ \mu \equiv \mathcal{E}(x) = \int x p(x) dx \] (15)

\( \Sigma \) is an \( N \times N \) covariance matrix of the features defined in equation 16.

\[ \Sigma \equiv \mathcal{E}[(x - \mu)(x - \mu)^t] = \int (x - \mu)(x - \mu)^t p(x) dx \] (16)

\( \Sigma \) is always symmetric and semidefinite, \( |\Sigma| \) is its determinant and \( \Sigma^{-1} \) is its inverse matrix. To simplify this, we observe that the components of the covariance matrix are the single pair covariances \( \sigma_{i,j}^2 \). Also, we note that the expected value of a vector is found by taking the expected value of its components. Thus, our individual scalar components are:

\[ \mu_i = \mathcal{E}[x_i] \] (17)

And,

\[ \sigma_{i,j}^2 = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \] (18)

Now, if we assume statistically independent features then the covariance matrix becomes diagonal where \( \sigma_{i,j}^2 = 0 \) when \( i \neq j \). In this case, the multivariate Gaussian distribution simplifies to a multiplication of single variable distributions as in equation 19.

\[ p(x|g_k) \sim \mathcal{N}(x|\mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left( -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \] (19)
Therefore, under these assumptions what we need is to calculate the $\mu_i$ and $\sigma_i^2$ of each feature in each class condition during the training phase. This will give us the conditional probability density function $p(x|g_k)$ for every gesture $g_k$.

Finally, looking back at equation 13, one last assumption is made, and it is that prior distributions of the gestures $P(g_k)$ are equal. This implies that prior to any knowledge there is an equal chance for any gesture to occur. This assumption seems reasonable to make considering the complexity and randomness of emotional responses in music. If this is not the case, and we have some prior knowledge such as the composer’s or performer’s inclination for specific expressive gestures, then adjusting this factor could improve the performance of the classifier.

Therefore, for every input feature vector $x$ we compute the conditional probability density $p(x|g_k)$ for all $G$ gestures. And based on our assumptions and equation 13, this is proportional to the posterior probability $P(g_k|x)$. Thus, the gesture with the highest prior is the one that will be predicted by the classifier.

This is the simplest case of Bayes classification. However, the ANBC block has some additional features to enable flexibility in its application that make it a powerful tool for musical gesture recognition.

**Feature Weighting**

The training block of the ANBC has a weight parameter for every for each feature. This allows certain features to be more significant than others in the classification process. The weighting works as follows. For each gesture class $k$, a
discrimination function is created from vectors of mean, variances, and the weight vector:

$$\Phi_k = \Phi\{\mu_k, \sigma^2_k, \phi_k\}$$ (20)

The function is defined such that if the weight $\phi_n$ is greater than zero, then the Gaussian function is multiplied by the weight, otherwise the function is forced to 1, thus not affecting the classification process. Then, the product of all these functions is computed as before. To prevent underflow of precision due to many multiplications, a log is taken over each distribution, followed by an addition.

In our implementation, the features are weighed such that those from the current frame are weighed 1.0, those from the previous frame are weighed 0.6 and those from the frame before are weighed 0.3. This allows for a diminishing significance of the past features.

**Rejection Threshold**

A second important feature of the ANBC is allowing for a null class without having the user explicitly define and train it. This is carried out by computing a rejection threshold for each gesture under which the classifier will not choose the gesture even if it is with the highest probability. Therefore, if none of the gestures pass the rejection threshold, the classifier will choose the null gesture. This feature is especially important for the case of continuous data that may or may not contain gestures, and this is the case when we are presented with IMUs in piano performance.

The rejection threshold is calculated using a confidence measure in the
form of a log likelihood confidence and standard deviation. The threshold is set by:

\[ \tau_k = \mu_k^* - (\sigma_k^* \gamma) \] (21)

Where \( \mu_k^* \) is the mean log likelihood, \( \sigma_k^* \) is the standard deviation log likelihood, and \( \gamma \) is a scalar set by the user. For an overview on how these are calculated, see [Gillian et al., 2011b].

**Adaptive Real-time Training**

The third powerful feature of the ANBC is the ability to add training samples to improve the model while classifying in real time. This enables an initial relatively small training set, then more data could be recorded for refining the classifier during the performance. The adaptive training feature is controlled by three parameters: maximum training buffer size, model update rate, and the scalar number of standard deviations for the rejection threshold. Based on these parameters, every time the rejection threshold is crossed a new sample is added to the buffer, discarding the oldest one, and a counter is incremented. Once the counter crosses the update rate, the model is recalculated based on the new samples and the counter is set to zero. The adaptive training feature can be powerful for improving accuracy and adapting to performers, however it is also inherently sensitive to errors, especially at the beginning of a trial. If the first samples are classified incorrectly (despite crossing the rejection threshold) then a run-away model will be created which will get worse at every step.
4.3 Post Processing and Visual Feedback

This section describes the operations performed after the Bayes classification. The classifier generates a prediction 64 times per second. These predictions are then further processed for display and feedback.

4.3.1 Post Processing

Since the output of the classifier is a discrete class from 1 to 6 and it is updated every sample, some post processing is required in order to exclude outliers and smooth the transition between emotions. This is meant to create the effect of a human listener perceiving the music and responding emotionally to it.

To do this, two main assumptions are laid:

- Emotions change slowly at intervals of at least several seconds (or several notes) and have a similarly long minimum duration. This is based on short term memory for objects [Miller, 1956] and on phonological working memory research [Baddeley and Hitch, 1974].

- While shifting from one emotion to another, the emotional path should be as continuous as possible and follow the implicit dimensionality in Hevner’s adjective circle [Gabrielsson and Lindström, 2010]. Hence, while jumping from sad and mourning to humorous and playful must first pass through tender and then serene.

These assumptions make the flow of the algorithm more human like in its real-time behavior while following an emotionally varying performance. In order
to account from these assumptions the last few seconds of the classifier output are buffered into a mode filter in order to remove occasional erroneous outliers. Then, then a running average is employed on the output to smooth the transition from one state to another. This creates intermediate categories between the adjectives that can describe transient emotions.

4.3.2 Visual Feedback

The data are collected, trained and classified in real time. The emotions are predicted as discrete integers 1–6 and displayed as text on the screen. To make the display more intuitive and to smooth the behavior of the algorithm, a moving average is computed over the previous several seconds. The working memory for sound, known as the phonological loop is generally known to be 1–2 seconds [Baddeley and Hitch, 1974]. However, following several trials in the development of this display I observed that musical emotions change less rapidly than this and a three second buffer was found to be more appropriate. This creates a continuous flow that is used for two displays as follows.

Piano Roll Display

The piano roll display maps the MIDI note on and note off events to colored locations on an image. The color of the notes is set by the velocity of the note on event. The background color of this piano roll is mapped from the continuous emotion number generated by the classifier. The mapping of emotional musical performances to colors is based on [Bresin, 2005] as described in the background section and was used as a reference for our color mapping.
Displaying the emotions predictions this way allows for the feeling of a continuous flow in the music along with the musical emotion coloring that evolves in a natural rate similar to how we experience emotional responses in music. The emotional piano roll implementation is shown in Figure 21.

![Figure 21: Display of piano roll with emotion colored background.](image)

**Color Wheel Display**

The second display is the Adjective circle projected on an HSV color wheel. The current state of emotion is represented by a moving, circle-shaped object followed by a tracer (tail). The head of the object is located on the current emotion predicted by the algorithm and the tail on the recent trajectory of emotional states. This display provides a feedback that is motion-like and can serve as a motive indication to the pianist similar to a dancer responding to the music. The emotional color wheel implementation in EyesWeb is shown in Figure 22.
Figure 22: **Adjective Circle on HSV Color Wheel.** The head of the circle shaped object tracks the current emotional state tracked by the algorithm, providing the performer with dance-like feedback responding to the music. The emotion-to-color mapping is based on [Bresin, 2005] the adjectives are from [Hevner, 1936].
Evaluation

5.1 Evaluation and Research Goals

The evaluation section has two goals:

- To assess the accuracy in the system’s performance with regards to detecting emotions.
- To provide an understanding of the detail of information that can be obtained from kinesthetic data of musicians and to attempt to identify the reasons and possible solutions for limitations.

Assessing the performance of a classifier is carried out by training it with a training sample set and then classifying using a different known test set. In order to do this, the training and data must include each of the classified categories. This is challenging to accomplish when dealing with live performance since it implies either requesting the musician to perform in a certain fashion (all of the expressions and emotions) or having the musician self-report on the performance. In both cases, the experiment is sensitive to bias and subjectivity.

Moreover, as mentioned earlier, the standard paradigm in experiments such as these is to play the same piece in different emotions thus eliminating the factor of the composition. However, in my pilot study it was observed that it is difficult to express an emotion in a composition that was not meant to convey it. This causes a dissonance in the performance, rendering it more of a technical
attempt at acting an emotion than a successful communication of it. For this reason, I have decided to add to the standard paradigm a test where each emotion is portrayed via a dedicated piece. Because the algorithm is completely deaf to the music the composition factor diminishes in its significance compared to that of analyzing the audio signal.

Finally, it is also in the scope of this project to follow the patterns in which performers shift from one emotion to the other during a performance. This could reveal some interesting observations regarding live performance as well as a musicological understanding of the structure of a piece. Therefore, a third section was added to the test in which the performers were requested to play freely either from a long score or through improvisation while self-reporting on their emotional intentions.

5.2 Subjects

The test subjects for this evaluation were a homogenous group of 13 pianists, with an average playing experience of $\mu = 12.6 \pm 4.8(\text{yrs})$. The age group statistic was $\mu = 21.8 \pm 3.0(\text{yrs})$. Four subjects described their genre as Classical, three as Jazz, four Contemporary, and two Rock/Pop. The subjects were from a variety of academic backgrounds and academic levels, this was not considered an affecting factor in this research since the subjects were not required to perform any academically related task other than piano performance. The details of the subject group are described in Table 1.
Table 1: Test Subjects Description

<table>
<thead>
<tr>
<th>Subject</th>
<th>age (yrs)</th>
<th>experience (yrs)</th>
<th>genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>12</td>
<td>Classical</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>5</td>
<td>Contemporary</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>14</td>
<td>Classical</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>11</td>
<td>Pop/Rock</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>17</td>
<td>Classical</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>15</td>
<td>Contemporary</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>8</td>
<td>Pop/Rock</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>12</td>
<td>Jazz</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>5</td>
<td>Classical</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>23</td>
<td>Jazz</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>12</td>
<td>Jazz</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>13</td>
<td>Contemporary</td>
</tr>
<tr>
<td>13</td>
<td>21</td>
<td>17</td>
<td>Contemporary</td>
</tr>
</tbody>
</table>

5.3 Experiment Procedure

As described above, because the system is required to detect various patterns and due to the complexity of the different affecting factors in this condition, the experiment was divided to three sections.

5.3.1 Stage 1 – The Standard Paradigm, One Piece Different Emotions

In this section, the performers played the first few measures of Bach’s Minuet in G major, BWV 841 from the Notebook of Anna Magdalena Bach. The piece was played six times in the six different emotion categories based on the standard paradigm [Juslin and Timmers, 2010] to create the training data. The performers were only allowed to make variations in intensity, tempo, accents and slight pitch variations (major/minor and decorations). This was repeated twice while randomizing the order for training and testing. The algorithm classified emotion categories were tested against the intended emotion categories.
5.3.2 Stage 2 – Different Pieces, Different Emotions

In this section, the performers chose six different pieces with the intention to convey the six different emotion categories. The hypothesis in this section is that it would be more intuitive for the performers to express the emotions that matched the piece. An example of the six pieces played (per category) by the performers were:

- sad/mournful  Nocturne in E Minor, Chopin
- dreamy/tender  Clair de Lune, Debussy
- lyrical/serene  Songs without Words, Mendelson
- playful/humorous  Maple Leaf Rag, Joplin
- cheerful/merry  Sonata in C Major, Mozart
- vigorous/dramatic  Sonata no. 8 movement 1, Beethoven

The data from this stage was used as training data for stage 3. The primary hypothesis was that training based on this data would achieve improved generality of the classifier compared to that of the standard paradigm in stage 1. This hypothesis was clearly observed while attempting to classify the stage 2 results based on training from stage 1. The reason for this is that training on one song severely limits the expressivity that can be conveyed and thus it yielded poor results in classifying other pieces with a wider range of expressiveness. In
other words, more training data was needed displaying a wider range of expressivity. This proved successful as seen in the Results section.

5.3.3 Stage 3 – Free Playing, Self-report, and Listener Evaluation

In this section, the performers were asked to play freely for several minutes while continuously self-reporting on their intended emotions. They played either from a selected score or improvisation. The algorithm had been trained before based on the previous section. However, in this section the classified emotion categories were tested against the intended emotion categories as well as against listener evaluation. The listener evaluation was carried out by four musically trained listeners. Thus, a three-way comparison was performed between the intended emotion, the algorithm’s classification, and the average of the listeners’ perceived emotion.

5.4 Tools for Classifier Evaluation

There are several methods and techniques to evaluate a classifier that vary with the application, the implementation, and the purpose of the evaluation. In this research, the classification results were evaluated using a Confusion Matrix, Precision, Recall, Specificity, and Accuracy parameters, RMS error, $\chi^2$ test, and Cohen’s $\kappa$ test to assess the statistical validity of the results. The following is a brief overview of these tools and how they were used in the context of this research.
5.4.1 Confusion Matrix

The confusion matrix is the simplest way to provide a qualitative but comprehensive evaluation on the performance of a classifier. It displays the total classification in an $c \times c$ matrix where $c$ is the number of categories. The rows represent the labeled (true) categories and the columns represent the classified (estimated) categories. As the classification improves, the matrix becomes diagonal (with higher values on the main diagonal and zeros in the remaining fields). The confusion matrix is also useful for understanding the limitations and weaknesses of the classifier, i.e. which categories are classified incorrectly and which categories they are confused with, hence the name, confusion matrix.

5.4.2 Precision, Recall, Accuracy, and Specificity

Precision and Recall are single parameters that evaluate the performance of an algorithm in detecting an occurrence of an event [Powers, 2007]. These can be extracted directly from the confusion matrix. In order to define precision and recall, four terms are defined:

- True Positive – the prediction of a category in its presence.
- False Positive – the prediction of a category in its absence.
- False Negative – the non-prediction of a category its presence.
- True Negative – the non-prediction of a category in its absence.
Then, *Precision* and *Recall* are defined in equations 22 and 23 respectively:

\[
Precision = \frac{TP}{TP + FP}
\]  
\[
Recall = \frac{TP}{TP + FN}
\]

Looking again at these definitions, we can observe that *Precision* is an evaluation of the positive predictions while *Recall* is an evaluation of the true positive rate and it is also referred to as *Sensitivity*. In other words, *Precision* is an evaluation of the correctness of classifications generated by the algorithm while *Recall* is an estimation regarding the ability of the classifier to detect an event in its occurrence. A third term that is useful is *Accuracy* and it is defined in equation 24:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

Another term for classifier evaluation is *specificity* as defined in equation 25. It is an estimation of the probability of a negative prediction of a class in its absence.

\[
Specificity = \frac{TN}{TN + FP}
\]

In this research, the *Precision*, *Recall*, *Accuracy*, and *Specificity* parameters are calculated on each of the six categories.

**5.4.3 RMS Error**

This evaluation metric is based on our assumption that there exists a dimensionality in the category space of the adjective circle [Gabrielsson and Lindström, 2010]. Thus, two adjacent emotion categories
are also close in there expressive content. Moreover, the ground truth is set by
the emotion category the performer was requested to play; however, it was
observed that the pianists drifted between categories based on their self-report.
Therefore, the binary relation of true/false is an incomplete description of the
algorithms behavior. Instead, we measure the RMS distance of the classification
from the intended emotion. The RMS error is defined in by:
\[
RMS \text{ error} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\omega_c - \omega_i)^2}
\] (26)
Where \(\omega_c\) is the classified emotion, and \(\omega_i\) is the intended emotion.

Another aspect of the assumed dimensionality is that the categories are
set on circular space. Therefore the difference between categories is the distance
on the circle. In other words, there is a wrapping of the classes, where the actual
distance between two categories is the shortest path around the adjective circle.
For example, the distance from vigorous/dramatic to dreamy/tender is two, and
so is the distance to humorous/playful. The distance error is therefore an angular
RMS error and can be measured in units of degrees.

5.4.4 \(\chi^2\) Test

The \(\chi^2\) test is a common statistical method test used evaluate a null
hypothesis when a sample set consists of several events or categories that are
mutually exclusive and have a total probability of 1. The purpose of the test is to
examine the rejection of the null hypotheses. The null hypothesis in the case of
this research would be that the classification outputs of the algorithm are a
random variable and uniformly distributed with equal probability.

The test is performed by separating the data into category bins and comparing the observed and expected rates at which each bin occurs. Then, the \( \chi^2 \) is calculated by:

\[
\chi^2 = \sum_{i=1}^{M} \frac{(O_i - E_i)^2}{E_i}
\]

(27)

Where

- \( M \) – number of categories or bins.
- \( E_i \) – expected number of occurrences in category \( i \).
- \( O_i \) – observed number of occurrences in category \( i \).

Then the \( \chi^2 \) is mapped to a \( p \) value indicating the probability of the observed data given the null hypothesis. If \( p \) is below a threshold we say that we can reject the null hypothesis.

For the purpose of this research, we shall define a sample as one instance of classification or categorization. Therefore, for every subject, and for every intended emotion category we have \( N \) instances of categorizations into one of six bins (the six emotion categories). The null hypothesis is that the algorithm is randomly guessing with equal probability (uniform distribution one out of six), therefore the expected samples in each of the six bins are \( \frac{N}{6} \). We then plug in our observed classifications for that trial and compute the \( \chi^2 \) based on equation 27. Our number of bins is six and therefore we have five degrees of freedom. This is used to map the \( \chi^2 \) to a \( p \) value. The \( p \) value represents the probability that, had
the null hypothesis been true, we would get our observations. Therefore the lower
$p$ is, the more unlikely the null hypothesis.

5.4.5 **Cohen's $\kappa$ test and weighted $\kappa$ test**

The $\kappa$ test, [Cohen et al., 1960], is a statistical measure of agreement
between two judges in a multicategorization problem. In the context of this
research one judge would be the performer’s intended emotion and the other
judge is the algorithm’s classification. The $\kappa$ test accounts for the
chance-expected agreement, thus evaluating the actual proportion of agreement
after chance is removed from consideration. For this, two quantities are defined:

- $p_o$ - the proportion of units in which the judges agreed.
- $p_c$ - the proportion of units for which agreement is expected by chance.

The purpose of the test is to evaluate the extent to which $p_o$ exceeds $p_c$.
Therefore, $\kappa$ is defined by:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \quad (28)$$

We can compute this directly from the frequencies in the confusion matrices for
each emotion category by:

$$\kappa = \frac{f_o - f_c}{N - f_c} \quad (29)$$

Where, $f_o$ are the observed classifications and $f_c$ are the expected classifications,
i.e $N/6$. Thus, a high $\kappa$ implies a good agreement between the judges. There is
no one criteria scale for interpreting the $\kappa$ values. In this research we use the
scale suggested by [Landis and Koch, 1977].
However, the simple $\kappa$ test assumes that all misclassifications are equally costly. In order to account for categories that are ordered in dimensionality (similar to our case where adjacent categories are considered “close” in the emotional aspect), the *weighted $\kappa$ test* was developed [Cohen, 1968]. In order to compute a the weighted $\kappa$ we define a $c \times c$ weight matrix, where $c$ is the number of categories. Typical weights are linear or quadratically decaying values as a function of the distance from the main diagonal. Then, $f_o$ and $f_c$ are computed by multiplying the frequencies by the weights as in equations 30 and 31.

\[
\text{weighted } f_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} O_{ij} \tag{30}
\]

\[
\text{weighted } f_c = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} E_{ij} \tag{31}
\]

Where $O$ is the observed matrix and $E$ is the expected (chance) matrix.

In this research we will look at the unweighted, linearly weighted, and quadratically weighted $\kappa$ values. The reason for this is that we do not have a priori knowledge regarding the form of dimensionality in the emotional space, or in other words, how close is *sad* to *dreamy* relative to *sad* and *lyrical* or *cheerful* to *playful* and so on. Thus, as shown the results section, experimenting with different weight techniques might give us an understanding of the dimensionality of the emotion space in the adjective circle.
5.5 Results

This section presents the results followed by a brief interpretation. A more thorough inquiry is carried out in the discussion chapter.

5.5.1 Stage 1 - The Standard Paradigm, One Piece Different Emotions

The initial classification results per subject on all categories are summarized in Table 2. The correct and off by one values are calculated as a percentage of the total number of classification samples. Each trial consisted of 3000–6000 classifications, distributed evenly to 500–1000 classifications per emotion. It is important to note that the emotion categories are organized on the adjective circle in a way that adjacent emotions are close to each other in concept and those who are across are different. This is the implicit dimensionality mentioned earlier in [Gabrielsson and Lindström, 2010] and it is for this reason that we do not completely ignore incorrect classifications but also look at those that are one off. The total column shows the sum of correct and 1 off predictions. For example many times the emotions sad/mournful are confused with dreamy/tender. This confuses performers as well as the algorithm and the listeners as observed in the following sections. The percentage of correct classifications is approximately 50% and including those that are off by one we get to over 80% mean total classification accuracy. Also, the results are quite stable (std = 6.7%).

Furthermore, the RMS angular error is displayed in the most right column. This is calculated as explained in the previous section and is an
Table 2: Classification results per subject over all categories in the Standard Paradigm

<table>
<thead>
<tr>
<th>Subject</th>
<th>Correct (%)</th>
<th>Off by 1 (%)</th>
<th>Total (%)</th>
<th>Angular RMS error (◦)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49.37</td>
<td>35.082</td>
<td>84.452</td>
<td>61.884</td>
</tr>
<tr>
<td>2</td>
<td>46.128</td>
<td>39.319</td>
<td>85.447</td>
<td>61.279</td>
</tr>
<tr>
<td>3</td>
<td>59.03</td>
<td>17.893</td>
<td>76.923</td>
<td>63.496</td>
</tr>
<tr>
<td>4</td>
<td>67.531</td>
<td>18.904</td>
<td>86.436</td>
<td>51.321</td>
</tr>
<tr>
<td>5</td>
<td>23.917</td>
<td>45.002</td>
<td>68.919</td>
<td>78.292</td>
</tr>
<tr>
<td>6</td>
<td>39.626</td>
<td>32.532</td>
<td>72.158</td>
<td>76.615</td>
</tr>
<tr>
<td>7</td>
<td>28.828</td>
<td>49.545</td>
<td>78.373</td>
<td>71.144</td>
</tr>
<tr>
<td>8</td>
<td>47.981</td>
<td>27.019</td>
<td>75</td>
<td>81.163</td>
</tr>
<tr>
<td>9</td>
<td>67.899</td>
<td>18.133</td>
<td>86.031</td>
<td>51.616</td>
</tr>
<tr>
<td>10</td>
<td>61.576</td>
<td>15.117</td>
<td>76.693</td>
<td>70.392</td>
</tr>
<tr>
<td>11</td>
<td>59.21</td>
<td>14.983</td>
<td>74.193</td>
<td>74.952</td>
</tr>
<tr>
<td>12</td>
<td>60.97</td>
<td>27.931</td>
<td>88.9</td>
<td>51.028</td>
</tr>
<tr>
<td>13</td>
<td>49.301</td>
<td>39.263</td>
<td>88.564</td>
<td>55.319</td>
</tr>
<tr>
<td>population mean</td>
<td>50.8743</td>
<td>29.2864</td>
<td>80.1607</td>
<td>65.2693</td>
</tr>
<tr>
<td>population std</td>
<td>13.8282</td>
<td>11.8183</td>
<td>6.7427</td>
<td>10.8732</td>
</tr>
</tbody>
</table>

indication of the mean accuracy angle when a classification is made. Thus, while the algorithm is classifying an emotion, this predicts the error range in degrees in which the emotion is expected to be found. The angular distance between adjacent emotions is 60°. We get a mean angle accuracy of approximately 65° which implies that the predictions are within one emotion on the adjective circle.

The overall confusion matrix is calculated by the sum of all confusion matrices and is displayed in Figure 23. The results show high performance in detecting most of the intended emotions. A clear diagonal is observed in the confusion matrix, with the main discrepancies occurring between the sad, dreamy, and serene categories.

In order to understand the confusion matrix more, one should observe it in varying levels of resolution. First, four major blocks are clearly seen, the two
on the main diagonal are bright and the two on the remaining are dark. This is an indication of the algorithm’s strong ability in distinguishing between high and low arousal in emotions. The sad, dreamy, and serene are low arousal and the humorous, cheerful, and vigorous are high arousal categories.

Second, the inner diagonals are observed, this is an indication of the algorithm’s ability in distinguishing valence. The sad and vigorous categories are considered low in valence while the lyrical, humorous, and cheerful are in high valence. The dreamy category is generally assumed neutral in valence (there are good dreams and bad dreams). Perhaps this is why it gets confused the most.

Figure 23: Overall Confusion Matrix Stage 1. A clear diagonal is observed, the main discrepancies are between dreamy/tender, sad/mournful, and lyrical/serene.

Based on the total confusion matrix, the results for Precision, Recall,
Table 3: Precision, Recall, Specificity, and Accuracy per category for Stage 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>sad/mournful</td>
<td>0.5582</td>
<td>0.421</td>
<td>0.8842</td>
<td>0.7647</td>
</tr>
<tr>
<td>dreamy/tender</td>
<td>0.2999</td>
<td>0.4161</td>
<td>0.8208</td>
<td>0.7578</td>
</tr>
<tr>
<td>lyrical/serene</td>
<td>0.4371</td>
<td>0.5263</td>
<td>0.8867</td>
<td>0.8351</td>
</tr>
<tr>
<td>humorous/playful</td>
<td>0.5993</td>
<td>0.5072</td>
<td>0.9311</td>
<td>0.8595</td>
</tr>
<tr>
<td>cheerful/merry</td>
<td>0.4814</td>
<td>0.5203</td>
<td>0.9159</td>
<td>0.8643</td>
</tr>
<tr>
<td>vigorous/dramatic</td>
<td>0.8797</td>
<td>0.8137</td>
<td>0.9813</td>
<td>0.9572</td>
</tr>
</tbody>
</table>

Specificity, and Accuracy are be obtained per category and displayed in Table 3.

The results show that the vigorous emotion has the highest precision, recall, specificity, and accuracy. This is expected since the vigorous playing is very different and easily distinguishable from the other emotions. This is also consistent with the functionalist perspective [Juslin, 1997], i.e. that we are programmed to be sensitive to emotions that can be life threatening and are imperative to our survival.

The humorous and cheerful categories had similar results, both were lower in precision and recall because they were confused between each other. Their accuracy however, is still relatively high because when detected they were not confused with other categories.

The sad and lyrical also had similar results but lower than the other categories because they were often not only confused between each other but also with dreamy. The dreamy category had the lowest achievement in all categories. This too, matches our expectation regarding its neutral valence as explained in the analysis of the confusion matrix.

The results for the $\chi^2$ test per emotion category are displayed in Table 4.
Table 4: $\chi^2$ test results per category for Stage 1.

<table>
<thead>
<tr>
<th>category</th>
<th>$N$</th>
<th>dof</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sad/mournful</td>
<td>38</td>
<td>5</td>
<td>27.895</td>
<td>0.0000382</td>
</tr>
<tr>
<td>dreamy/tender</td>
<td>42</td>
<td>5</td>
<td>32</td>
<td>0.00000594</td>
</tr>
<tr>
<td>lyrical/serene</td>
<td>32</td>
<td>5</td>
<td>21.875</td>
<td>0.000553</td>
</tr>
<tr>
<td>humorous/playful</td>
<td>29</td>
<td>5</td>
<td>19.897</td>
<td>0.001307</td>
</tr>
<tr>
<td>cheerful/merry</td>
<td>29</td>
<td>5</td>
<td>18.655</td>
<td>0.002228</td>
</tr>
<tr>
<td>vigorous/dramatic</td>
<td>25</td>
<td>5</td>
<td>17.32</td>
<td>0.003931</td>
</tr>
</tbody>
</table>

Table 5: $\kappa$ test results for Stage 1. The agreement criteria is taken from [Landis and Koch, 1977].

<table>
<thead>
<tr>
<th>weighting</th>
<th>$p_o$</th>
<th>$p_c$</th>
<th>$\kappa$</th>
<th>agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>unweighted</td>
<td>0.5193</td>
<td>0.1667</td>
<td>0.4231</td>
<td>moderate</td>
</tr>
<tr>
<td>linearly weighted</td>
<td>0.8460</td>
<td>0.6107</td>
<td>0.6044</td>
<td>substantial</td>
</tr>
<tr>
<td>quadratically weighted</td>
<td>0.9361</td>
<td>0.7662</td>
<td>0.7269</td>
<td>substantial</td>
</tr>
</tbody>
</table>

The method of calculation is explained in the previous section. $p < 0.01$ is observed for all categories, clearly rejecting the null hypothesis. Thus, the system is obviously performing much better than guessing one category out of six.

The results for the $\kappa$ test on the complete confusion matrix are displayed in Table 5. The method of calculation is explained in the previous section. The agreement criteria is taken from [Landis and Koch, 1977]. As expected, it is observed that using the weighted $\kappa$ produces a better agreement criteria. However, it is interesting to observe that the quadratic weighting is better than the linear. This will be discussed further in the next section.

5.5.2 Stage 2 – Six Pieces, Different Emotions, testing on Free-playing and Self-report

In this section, the algorithm was trained on the performers playing six different pieces that were chosen to express the six emotion categories. Following that, they were asked to play freely and self-report on their intended emotion.
The first analysis is the overall confusion matrix of the self-reported intended emotion vs. the algorithm classification. The overall confusion matrix is calculated by the sum of all confusion matrices and is displayed in Figure 24. The main diagonal is still observed. However, the sad and dreamy categories were often misclassified as serene. Also, the humorous category was misclassified as cheerful and vigorous. Misclassifying all three low-arousal categories with serene implies that the system in this test performed less accurately in distinguishing valence in the low-arousal categories. Once again, if valence cannot be distinguished, there is really no difference between sad, dreamy, and serene. Misclassifying humorous with cheerful and vigorous implies that there is a dimension in emotions we are not considering. Humor is typically thought of as high in valence but there is also dark and sarcastic humor. Moreover, humor in many cases by definition is a case of contradiction (such as the contradiction between the literal and actual meaning in the definition of irony) this type of contraction can easily throw off an algorithm trained to detect only the literal. Such is the case in many scherzo portions of pieces by Beethoven and Schubert which are often played in a playful manner but convey dark drama attributed to Beethoven’s growing deafness and Schubert’s decaying health [Ringer, 2009]. That said, perhaps there are other ways of detecting hidden meanings in motion that could be explored in future research.

Still, these results are quite promising considering the difficulty in detecting emotions from a much broader range compared to that of the standard
paradigm test in the previous section. The purpose of this section was to evaluate the performance of the algorithm in real-life scenarios where the system is presented with an unknown piece that it had not been trained on, and attempt to predict the intended emotions in it. Considering the difficulty in this, and the fact that it has not been attempted before in such conditions, the results of this section are satisfying.

Figure 24: **Overall Confusion Matrix for training on different pieces and test on free playing and self-report**  Rows represent labeled emotions while columns represent algorithm classified emotions. The main diagonal is still observed. However, the *sad* and *dreamy* categories were often misclassified as *serene*. Also, the *humorous* category was misclassified as *cheerful* and *vigorous*.

Based on the total confusion matrix, the results for *Precision*, *Recall*, *Specificity*, and *Accuracy* are be obtained per category and displayed in table 6. The results for the $\chi^2$ test per category are displayed in table 7. The precision and recall rates are lower than before but the accuracy and specificity are still
Table 6: Precise, Recall, Specificity, and Accuracy per category for Stage 2

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>sad/mournful</td>
<td>0.2992</td>
<td>0.4822</td>
<td>0.8473</td>
<td>0.8039</td>
</tr>
<tr>
<td>dreamy/tender</td>
<td>0.0561</td>
<td>0.189</td>
<td>0.8037</td>
<td>0.7679</td>
</tr>
<tr>
<td>lyrical/serene</td>
<td>0.5604</td>
<td>0.2865</td>
<td>0.89</td>
<td>0.6917</td>
</tr>
<tr>
<td>humorous/playful</td>
<td>0.3202</td>
<td>0.3422</td>
<td>0.886</td>
<td>0.8123</td>
</tr>
<tr>
<td>cheerful/merry</td>
<td>0.4426</td>
<td>0.3547</td>
<td>0.8741</td>
<td>0.7599</td>
</tr>
<tr>
<td>vigorous/dramatic</td>
<td>0.4846</td>
<td>0.4298</td>
<td>0.9264</td>
<td>0.8575</td>
</tr>
</tbody>
</table>

Table 7: χ² test results per category for Stage 2.

<table>
<thead>
<tr>
<th>category</th>
<th>N</th>
<th>dof</th>
<th>χ²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>sad/mournful</td>
<td>34</td>
<td>5</td>
<td>24.471</td>
<td>0.000176</td>
</tr>
<tr>
<td>dreamy/tender</td>
<td>35</td>
<td>5</td>
<td>25.4</td>
<td>0.000117</td>
</tr>
<tr>
<td>lyrical/serene</td>
<td>29</td>
<td>5</td>
<td>18.655</td>
<td>0.002228</td>
</tr>
<tr>
<td>humorous/playful</td>
<td>23</td>
<td>5</td>
<td>14.652</td>
<td>0.011957</td>
</tr>
<tr>
<td>cheerful/merry</td>
<td>33</td>
<td>5</td>
<td>22.818</td>
<td>0.000366</td>
</tr>
<tr>
<td>vigorous/dramatic</td>
<td>25</td>
<td>5</td>
<td>15.4</td>
<td>0.008783</td>
</tr>
</tbody>
</table>

quite high. The results of the χ² still show $p < 0.05$ value, rejecting the null hypothesis. However, it is seen now that the playful category achieved the least significant results ($p = 0.012$). This can be seen in the confusion matrix where playful was often confused with cheerful and vigorous.

The results for the κ test for Stage 2 on the complete confusion matrix are displayed in table 8. The method of calculation is explained in the previous section. The agreement criteria is taken from [Landis and Koch, 1977]. Here too, it is observed that better agreement is achieved with the weighted κ test and especially with the quadratically weighted κ. It seems that there exists a pattern here which could imply the form of the dimensionality in the emotion space. This form is a falling quadratic where adjacent categories are very similar but the similarity drops rapidly after one category. For example, dreamy is very similar
Table 8: $\kappa$ test results for Stage 2. The agreement criteria is taken from [Landis and Koch, 1977].

<table>
<thead>
<tr>
<th>Weighting</th>
<th>$p_o$</th>
<th>$p_c$</th>
<th>$\kappa$</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>0.3466</td>
<td>0.1667</td>
<td>0.2159</td>
<td>Fair</td>
</tr>
<tr>
<td>Linearly weighted</td>
<td>0.7448</td>
<td>0.6122</td>
<td>0.3419</td>
<td>Fair</td>
</tr>
<tr>
<td>Quadratically weighted</td>
<td>0.8723</td>
<td>0.7680</td>
<td>0.4495</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

to sad and lyrical which are immediately adjacent to it but very different from playful and vigorous which are two steps from it. Likewise, playful is similar to cheerful and lyrical but very different from vigorous and dreamy. This fits a falling quadratic form more than a linear form and it is also pleasingly consistent with our initial justification for adding the “one off” column in the results of Stage 1.

5.5.3 Stage 3 – Performer vs. Listener vs. Algorithm

In this section, three pieces from the free-playing self-report part were selected for listener evaluation. The pieces were selected one from each genre, Classical, Jazz, and Contemporary. Four listeners evaluated the perceived emotion and compared to the performers’ intentions and the algorithm classification. This three-way comparison allows for interesting qualitative observations of what the system is doing. First, instead of looking at a 2D confusion matrix we present a 3D confusion matrix or a Confusion Cube. The confusion cube of the total algorithm performance on the three performers is presented Figure 25.

Each of the small cubes represents a bin of classifications corresponding to the emotion a coordinate of the cube. The colors in each cube are mapped the
Figure 25: Three dimensional confusion cube. The intended emotions are compared with the listener perceived emotions and the algorithm classification through a complete piece.
number of classifications via the jet color mapping. The $\alpha$ transparency factor is also adjusted according to the number of classifications. Hence, empty bins will remain colorless, bins with only few classifications in them will be colored in blue transparent colors, and fuller bins will be colored yellow and then in dark red. The main diagonal is still observed but is it wider now, and there are disagreements especially regarding the sad, dreamy, and serene categories. It is still observed though, that most of the classifications occur around the main diagonal and in the center of the cube and most of the remaining edges are empty.

Figure 26: Emotion paths of Performer, Listeners, and Algorithm through 10 seconds of Classical piece played by subject No. 9. The piece was Chopin Nocturne Posthumous in C♯ Minor. The performer, algorithm, and audience in agreement, but the algorithm is detecting vigorous at some point.

Next, while the confusion cubes are fun to look at, they can only tell us about the accuracy of detection at every point in time through the piece.
However, we would also like to observe the emotional path through longer sections of the piece. Thus, we can draw the emotion paths of the performer, listener, and the algorithm projected on the color wheel. This provides some interesting insights regarding communication of emotions during a performance.

Figure 26 shows the emotion path of a 10 second section during the Chopin Nocturne Posthumous in $\text{C}^\#$ Minor. It is observed that the performer, algorithm, and audience are in agreement, but the algorithm is detecting *vigorous* at some point in disagreement with the listeners and performer intentions.

Figure 27 shows a 20 second section of a Jazz improvisation. Here, we observe a general agreement that the section is centered between the homurous and *vigorous* emotions. The algorithm decides on *cheerful* at some point in disagreement with the performer and audience.

Figure 28 portrays the emotion paths through 20 seconds the contemporary piece. Here while the performer is attempting to convey *humorous* and *lyrical*, the *lyrical* is mostly detected by the algorithm and audience, but the *humorous* is perceived as *cheerful* by both. Moreover, the listeners and the algorithm detected a *dreamy* and even *sad* mood that the performer did not intend to convey.
Figure 27: Emotion paths of Performer, Listeners, and Algorithm though one minute of Jazz improvisation played by subject No. 11. There is a general agreement that the piece is centered on the humorous emotion.
Figure 28: Emotion paths of Performer, Listeners, and Algorithm through 20 seconds of a Contemporary improvisation played by subject No. 12. An agreement is observed at the *lyrical* emotion but the humorous is detected as *cheerful* and there is an unintended drift to *dreamy* and *sad*. 
6

Discussion

6.1 System Performance, Limitations, and Possible Improvements

The system presented is a first attempt at musical expression recognition based solely on kinesthetic sensing. It is successful in detecting basic expressions such as dynamics and articulation as well as the performer intended emotions and even listener perceived emotions. Moreover, the system can function as a performance feedback system via its various displays, the piano roll and the emotion color wheel. Using this, a musician can be presented with continuous feedback while practicing for a performance or while composing a piece. Moreover, this type of detection can also be used to augment musical instruments by adding musical and emotional intelligence to them and using the predictions of the algorithm to control various audio parameters in correspondence with the musical mood and performers intentions.

The system performed well in the standard paradigm where it was tested on the same piece played in six different emotions. However, in the second section where a more realistic situation was tested, the system was introduced to an unknown piece and expected to detect the expressions in it. In this case the system still succeeded in detecting the obvious emotions sad, vigorous, and cheerful. However, the system was limited in the prediction of ambiguous emotions such as lyrical and dreamy. Moreover, the humorous emotion is often confused with cheerful and vigorous. This could be explained by the notion that
these emotions differ mainly in valence and are similar in arousal. While it is quite straightforward to detect high and low arousal via motion, it is not the case for valence. There is some equivalence with the findings of [Schachter and Singer, 1962] and their research with epinephrine that is quite intriguing. The epinephrine alters the physiological state of arousal but it does not alter valence, therefore it was difficult in some cases to establish the emotional state. It was thus their conclusion that some emotional experiences could not be generated and via physiological changes because they require cognitive processing. This could be the case of some of the musical emotions as well. When we measure motion, we measure a physiological state. This could be used to carry some of the cues that convey emotions but possibly not enough for those that are complex and require more cognitive processing. Perhaps the dreamy and lyrical emotions fall under this category.

Other limitations of the system and this research include the discrepancy between performer intention and listener perception. Since this is the case, it is difficult to establish a ground truth on which to evaluate the system or assign a goal it must achieve. However, this is precisely the phenomena in which this system can be helpful in addressing, as I will explain in the future research section.

There are many ways in which the system could be improved. First, the obvious way to obtain more information would be to add more sensors. Sensors could easily be added to the legs and torso of a performer, measurements such as
these would add more independent features. It is generally shown that increasing
the number of independent features improves the performance of a classifier
[Duda et al., 1995a]. It is however, important to mention that it also increases
the number of dimensions and can lead to the requirement of a large training
data set. This limitation is referred to as the curse of dimensionality where it is
generally claimed that the required training data grows exponentially with the
dimensionality of the problem. However, this too could be overcome by use of
dimensionality reduction techniques such as Principle Component Analysis and
Multiple Dimension Analysis [Duda et al., 1995b]. Figure 29 shows on the
intensity, tempo, and articulation features per class projected on a
three-dimensional scatter plot. The classification problem is apparent and it is
observed that although the classes occupy different spaces, the shape seems quite
erratic and a decision rule would not take the form of a simple plane or curve.
Figure 30 shows all of the features after they have been projected to a
three-dimensional space using multiple discriminant analysis. Now the clusters
are clearly observed and even though there is still some overlap, the decision
boundaries are now easier to establish.

Second, in this research, the system was only trained per performer. This
is because the performers differed in their ranges and expression styles. However,
this severely limits the generality of the system and therefore its usability in
real-life scenarios. The training data from this research, could be used along with
proper reference to the inter-performer differences to obtain a scaled data set
Figure 29: *Intensity, tempo, and articulation 3D scatter plot*. Although the classes occupy different spaces, the shape seems quite erratic and a decision rule would not take the form of a simple plane or curve.
Figure 30: **3D scatter of linear combination of features projected to three dimensions using MDA.** Clusters are clearly observed and even though there is still some overlap, the decision boundaries are now easier to establish.
such that the system could be trained on multiple subjects. This might prove successful since the number and variety of training samples will grow significantly, and will therefore address the problem of dimensionality mentioned above.

Third, the system was only evaluated using a Bayes classifier. Other classifiers such as Hidden Markov Models could perhaps perform with better results. This assumption is based on the observation that music is a temporal phenomenon, and so are the emotions portrayed and perceived in it. Classifiers such as Hidden Markov Models are specifically designed to handle such time evolving systems in which the current state is influenced by previous states, and thus it is reasonable to assume that they might perform better at this task [Duda et al., 1995b].

Finally, the limiting of kinesthetic data was for the purpose of this research to answer the question, of how much we can achieve by looking only at motion. However, for further applications in which better classification performance is of main interest, using the audio signal in real time could reveal more information. Specifically information regarding major and minor tonality could, in many cases, help in distinguishing valence [Hevner, 1935].

6.2 Research Discoveries and Ideas for Future Research

We have seen that musically relevant information regarding a performance and specifically musically conveyed emotions can be detected by using real-time kinesthetic data. This observation is of importance for several reasons. First, it highlights the manifestation of emotions in musical performance and it serves to
show that they can be perceived and distinguished not only subjectively by humans, but also by a machine. Moreover, since the acoustic detailed musical information regarding pitch and tonality cannot be conveyed in these motions, it is only the envelope that is observed, and it is shown to be enough to classify to some extent. This is consistent with the Contour Theory discussed in [Kivy, 1980], [Kivy, 1989], [Davies, 1994] and also [Nussbaum, 2007].

However, it is also interesting to look at where the classification was less successful. These were generally the categories dreamy and serene that were in many case confused with sad. It was observed that even in cases where there was an agreement between the performer and listeners, the algorithm had difficulty distinguishing between them. This of course could imply a weakness of the algorithm, which as engineers, is always the first assumption. Nevertheless, it could also imply that the fine differences between these categories cannot be explained by the contour theory because they require more detail than the contour can carry. In other words, perhaps the distinction between some emotions requires a higher sample-rate than the musical contour carries. This means that some other carrier is at work for this information which could be the musical grammar and semantics referred to and devised by [Cooke, 1959] and [Jackendoff and Lerdahl, 1983]. Thus, the findings of this research might help determine the boundary in which the contour and the semiotic theories meet. Furthermore, since the semiotic theory generally requires some musical training and the contour theory does not, this could have implications on understanding
the borders on which a trained listener and an untrained listener would be able
to determine the emotional content of a musical piece.

Moreover, this research shows that emotions not only manifest in motion
but also that (for some of the performers) they manifest consistently and
repeatedly in the same gestures such that a machine would be able to detect and
classify them. Also, in the standard paradigm test it was noticed that because
playing an instrument forces the performer to repeat the same motions somewhat
accurately, it is a good tool to detect minor deviations in motion style that could
be used to infer on the emotional and possibly physiological changes. Thus,
tracking a performer playing an instrument could be used as a measurement tool
for subtle physiological responses to emotional states. In addition to this, it could
be used as a non-intrusive bio-feedback setting devised to help a subject regulate
physiological and emotional states.

In a related topic but closer to music pedagogy, we have also seen that the
successful communication of emotion is a challenge for performers. We have
observed that by simply asking the performer to play in a certain emotion or
report the emotion they are expressing we challenge them to not only play and
feel but also control what is felt and be aware of it. This awareness is a primary
key in successful communication, since if one is not aware of what one is
transmitting, how is the information expected to be conveyed to the other? This
is where our system can really come into play. Using a system such as this while
performing or practicing, forces the performer to become aware of the emotion
and mood of the performance and thus adjust the performance accordingly via feedback. In this research, the system was only trained based on the emotions the performers were instructed to play in the first and second sections. However, using the results of the listening tests and performing more listening evaluations on the rest of the recorded data, the algorithm could be trained based on an average audience perception. This could then better represent the audience in a performance feedback scenario. Using this and training on the complete data set could create a system that is trained on many samples and is quite intelligible regarding the audience perception of emotion in a variety of performances.

Moreover, the displays designed in the evaluation of this research are novel ways at observing the phenomena of communication in performance. The idea of the Confusion Cube represents the interaction that exists between performer, listener, and model. This is complementary to the Lens model and Lens equation described in [Juslin, 2000] and [Juslin and Timmers, 2010]. However, in contrast to the lens model and the lens equation that numerically evaluate and describe the overall success rate of a performance, the confusion cube describes which emotion categories fail in communication and which succeed. It could also show where in the “optic” path of lens model the failure occurred, whether it was between the performer and the model, the performer and the listener, or the listener and the model, thus pointing to one of the three factors: $G$, $R_e$ or $R_s$ in the lens equation as described in the background section. Finally, it could show which emotion it was confused with and then possibly allow a correction. The
emotion paths add to this and describe the communication success through a
temporal window during the performance. In this research, a window of 10–20
seconds proved to be a good time to observe deviations between performer,
audience, and algorithm. During a window shorter than this, it was sometimes
difficult to establish a meaningful observation and a longer window seemed to
average out the local deviations that occurred during the performance.

The display of emotion paths is also interesting from a musicological
perspective. If the algorithm is trained to accurately represent the audience in
determining the emotional path, then one could obtain an average musical mood
as well as the general musical path of a piece simply by playing it with the
system. This again, comes down to the matter of awareness. A composer or
songwriter might think that his piece conveys a certain emotion due to the lyrics
or the emotion he is currently feeling. But the algorithm (and perhaps audience)
might think otherwise and this feedback is important to any composer. Even the
mere fact that during the entire piece the algorithm is stuck on one emotion
could alert the composer that the piece is not evolving and might need some
modifications.

Finally, from a music-engineering point of view, an obvious future research
that should be carried out is, now that we know this information regarding our
performance, how do we use this to augment the musical experience in real time?
An immediate application could be displaying objects or videos on a screen in a
performance, corresponding to the musical mood, similar to the background
coloring of the piano roll in our system. Another option could be continuously controlling and adjusting audio effects or synthesized sounds that correspond to detected expressions. Future research should address questions regarding the relationship between certain effects such as *reverb* or *vibrato* and emotions and the use of these correlations in live performance. This would render the system not only emotionally intelligent but also musically competent.
LIST OF REFERENCES


