Computational Methods for Modelling and Explaining Perceived Emotions in Music

Sanga Chaki

Computational Methods for Modelling and Explaining Perceived Emotions in Music

Thesis submitted to
Indian Institute of Technology Kharagpur
for the award of the degree

of

Doctor of Philosophy

by

Sanga Chaki
(Roll No: 15AT91R01)

under the guidance of

Dr. Sourangshu Bhattacharya
(Department of Computer Science & Engineering)

and

Dr. Priyadarshi Patnaik
(Department of Humanities & Social Sciences)

ADVANCED TECHNOLOGY DEVELOPMENT CENTRE
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
October 2023

- To -

- My father, Late Dr. Korak Kanti Chaki -

- My mother, sister, and husband -

- and the music inside me -

CERTIFICATE OF APPROVAL

Certified that the thesis entitled **Computational Methods for Modelling and Explaining Perceived Emotions in Music**, submitted by **Sanga Chaki** to the Indian Institute of Technology Kharagpur, for the award of the degree of Doctor of Philosophy, has been accepted by the external examiner, and the student has successfully defended the thesis in the viva-voce examination held on October 11, 2023.

Prof. Joy Sen
(Chairman)

Prof. Sourangshu Bhattacharya
(Supervisor)

Prof. Priyadarshi Patnaik
(Co-Supervisor)

Prof. Pallab Dasgupta
(Member of the DSC)

Prof. K. S. Rao
(Member of the DSC)

Prof.
(External Examiner)

Advanced Technology Development Centre
Indian Institute of Technology Kharagpur
Kharagpur, India-721302

Certificate

This is to certify that this thesis entitled **Computational Methods for Modelling and Explaining Perceived Emotions in Music** submitted by **Sanga Chaki**, to the Indian Institute of Technology Kharagpur, is a record of bona fide research work carried out under our supervision and is worthy of consideration for award of Ph.D of the Institute.

Prof. Sourangshu Bhattacharya
Supervisor
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
India - 721302.

Prof. Priyadarshi Patnaik
Co-Supervisor
Department of Humanities & Social Sciences
Indian Institute of Technology Kharagpur
India - 721302.

I.I.T. Kharagpur
October 2023

Declaration

I certify that

a. the work contained in the thesis is original and has been done by me under the guidance of my supervisor;

b. the work has not been submitted to any other institute for any other degree or diploma;

c. I have followed the guidelines provided by the Institute in preparing the thesis;

d. I have conformed to ethical norms and guidelines while writing the thesis;

e. whenever I have used materials (data, models, figures and text) from other sources, I have given due credit to them by citing them in the text of the thesis, and giving their details in the references, and taken permission from the copyright owners of the sources, whenever necessary.

_____

Sanga Chaki

Acknowledgment

I express my deep and sincere gratitude to my research supervisor, Dr. Sourangshu Bhattacharya, and co-supervisor, Dr. Priyadarshi Patnaik for providing me with invaluable guidance and constant advice throughout my Ph.D. program. They have guided me through the rigorous process of problem-solving, and critical thinking, taught me to never give up, and trained me to conduct research independently. I give my sincere thanks to my Doctoral Scrutiny Committee members, Prof. Joy Sen, Prof. Pallab Dasgupta, and Prof. K. Sreenivasa Rao for their insightful comments and encouragement. I am thankful to the staff members of the department for providing the necessary support. I am thankful to the Indian Institute of Technology Kharagpur for providing me the financial assistance for carrying out this research work.

I am forever grateful to the family members of my supervisors and the doctoral scrutiny committee for always welcoming me wholeheartedly into their homes and lives. SwastikaDi, Swaralipi, Julie Didi, Pihu, SagarikaDi, Toto, and Runa Didi, some of my best memories of Kharagpur are with you. I am indebted to Prof. Pallab Dasgupta, Prof. Joy Sen, Prof. Priyadarshi Patnaik, Prof Arnab Roy, and Prof Anandaroop Bhattacharya for their constant support and encouragement towards my music.

My lab has given me some of my best friends for life. AntaraDi, SudakshinaDi, ArindamDa, Antonio, Sumana, Sudipa, Madhu, Sanyal, Saha, Anirban, Rumia, Sreejita, Arnab, Sunandan, Ipsita, Sumanta, Briti, each of you have enriched and sustained me in this rather rigorous and difficult journey of PhD. Our tea sessions will remain unmatched and legendary! ChandanDa, thank you for always being there, from administrative complications to how to face interviews. Soumi and Abhisek you have been the ideal younger siblings. Jun, Gouri, RajuDa, Rashmi, Srabani, Chandan and SuchitraDi, thank you for all the collaborations, inputs, and support all through.

Devleena and Debanjana, some relationships are not affected by distances and time differences. You are my sisters-in-arm, Ph.D. journey, and beyond.

Most of all, I am grateful to each member of my family. My father, late Dr. Korak Kanti Chaki, for cultivating my interest in music, and being the best father possible. From life philosophy to food choices, you are always there. My mother, Dr. Suparna Basu Chaki, for showing me how important it is for women to be independent, how to work hard for it, and to believe in myself. My sister, Sreshtha Chaki, for teaching me how to handle difficult life situations with ease and grace. And, my best friend and husband, Dr. Abhishek Ghoshal. Your brilliance, patience, determination, and tolerance set new benchmarks for me every day. You are my rock.

Lastly, I bow to my music, without which I am incomplete.

Sanga Chaki

Abstract

The rapid increase in musical content in various social media and other platforms necessitates data-driven computational modeling of perceived emotions to facilitate applications like music emotion recognition (MER). The description, measurement, collection, processing, and storage of perceived-emotion opinion data form an essential part of such studies, achieved using various emotion representations like discrete and dimensional models. Though significant studies exist based on these models on various music traditions, their effectiveness on Hindustani classical music (HCM) is relatively less studied. On the other hand, the *Nava Rasa* concept in Indian aesthetics provides emotion categories to denote aesthetic emotions, which have been explored sparsely in MER. In this thesis, first, we study an intensity-based, categorical emotion representation called the Emotion-word and Intensity-Value (EWIV) representation, where the emotion-words are taken from the *Nava Rasa* concept. We demonstrate the effectiveness of *EWIV* and validate the quality of self-reported emotions, on existing benchmark clip sets, and a newly introduced set of clips from HCM called the *EmoRaga* clip set for perceived emotion analysis in HCM. We also discuss *representativeness* of EWIV using *goodness-of-fit* measures for statistical models as our metric and demonstrate that it might provide a better fit for perceived emotion data under certain conditions.

Context is one of the key parameters that influence music emotion perception in listeners. We utilize excerpts from the EmoRaga clip-set to explore the influence of musical context on perceived emotions. We demonstrate that change in *immediate intrinsic musical context* changes the perception of musical excerpts, and term this phenomenon *intra-contextual influence*. Using EWIV emotion representation, we show how patterns of such influence emerge for dominantly happy and sad Sitar excerpts. This contributes to the modeling of the subtle nuanced variations in music-perceived emotions with different improvisations of the same music piece.

The dependence of perceived emotions from music on temporally distributed music segments makes most machine learning methods for MER unreliable, calling for human-understandable explanations of the model predictions. We present an attentive-LSTM-based, explainable dynamic emotion prediction model and show that it performs better than existing models on a benchmark dataset, using a benchmark feature set. We also demonstrate that a reduced feature set consisting of Spectral features gives comparable results. We apply the best models to the EmoRaga clip-set to successfully perform dynamic dominant and secondary emotion classifications, music emotion variation detection, and identification of the music segments with high probabilities of perceiving the dominant emotion. These studies also demonstrate the applicability of EWIV representations estimated from self-reported emotion data towards these MER applications. We compare the model-predicted important segments with those annotated with emotion motifs by experts, which yields significant overlap, demonstrating that these are indeed captured by the model, explaining the emotion predictions.

Keywords: Explainable music emotion recognition, emotion representation, MER in Hindustani classical music, emotion motif, intra-contextual influence

Contents

List of Symbols & Abbreviations

List of Symbols

R^2	Coefficient of determination
$\overline{\tau}$	Kendall's τ
AIC	Akaike Information Criterion value
λ	Listener in music perceived emotion experiments
ε	Perceived emotion
I	Perceived emotion intensity
t	Emotion perception timestamp
$\mathcal{E}$	Set of chosen emotion-words
$\boldsymbol{pEWIV}$	EWIV probability vector
$\boldsymbol{CIV^{c,\lambda}}$	cumulative intensity vector
α	Cronbach's Alpha
τ	Typicality
Dom_{ε}	Dominant emotions
Sec_{ε}	Secondary emotions
θ_{ε}	Angular value associated with emotion ε

List of Abbreviations

MIR	Music Information Retrieval
MER	Music Emotion Recognition
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
HCM	Hindustani Classical Music

EWIV	Emotion-Word and Intensity-Value
OE	Other Emotions
DK	Dont Know
F	Fear
A	Anger
S	Sadness
C	Calmness
W	Wonder
R	Romance
H	Happiness
E	Excitement
MEVD	Music Emotion Variation Detection
AIC	Akaike Information Criterion
TEP	Temporal Emotion Patterns
OVL	Overlap Coefficient
A-V	Arousal and Valence
AT	Attention Model
BAT	Backward Attention Model
MAE	Mean Absolute Error
SVM	Support Vector Machines
MFCC	Mel-frequency cepstral coefficients
BTC	Bi-directional Transformer
CQT	Constant-Q transform
STFT	Short-time Fourier transform
ComParE	Computational Paralinguistics Evaluation

List of Figures

List of Tables

CHAPTER 1

Introduction

Music is considered to be one of the cultural universals, with all known human cultures partaking in it. It is often postulated that the origin of music in humans is related to the emergence of human languages, with much debate surrounding whether music arose before, after, or simultaneously with language. Music has facilitated social bonding, organization of cohesive labor, improved communication, reduced conflict, and assisted in perceptual and motor skill development in humans, shaping biological [6] and cultural evolutions [7]. Music is also a form of aesthetic experience, with a great capacity to impact humans psychologically, by communicating and influencing emotions [8].

1.1 Motivation

Emotion evocation is one of the major reasons humans participate in music, from the perspectives of music composers, performers, and the audience alike [9], [10], [11]. Nowadays, modern technology has enhanced the nature of musical experiences, by including emotional components in such experiences. For example, emotion-based music streaming for user-specific actions and situations [12], [13], or simply for mood enhancement [14]. Other prominent applications include mood-based music recommendation [15], [16], and sentiment-based music generation [17].

Yet, despite music-evoked emotions being a fundamental facet of music listening, music-perceived emotions remain one of the most difficult aspects to model computationally. There is an increased awareness of the fact that music technology needs to utilize emotion representations, datasets, features, tools, and algorithms which are aligned with human understanding and perception of music, in order to develop systems more capable of facilitating musical interaction with humans [18], which include building explainable music emotion recognition (MER) models as well.

Most music traditions have an inherent attribution of musical features to specific emotions evoked by music. For example, major and minor notes are associated with happy and sad music respectively. Fast and slow tempos relate to exciting and calming music. The tonality of certain instruments evokes certain emotions. Similarly, musical context also influences perceived emotions in music. For example, in any standard Hindustani classical music (HCM) rendition, the composition consists of arrhythmic phases followed by rhythmic phases of varying tempos, each forming parts of a whole. The perception of one part plays a significant role in how subsequent parts are perceived, building up to an emotional crescendo at the climax. Understandably, musicians learn these rules and develop their intuition of what "works", through years of study and practice. But even the non-initiated audiences are able to perceive, recognize and appreciate the expressed emotions [19], along with changed nuances of emotions in different renditions of the same music. This is especially observed in HCM renditions, where it is a common practice to improvise the same musical phrase in different ways, altering the nuances of perceived emotions. Capturing and utilizing this subtle but rich contextual information and incorporating them into music emotion systems is still a challenge.

Recently, the rapid increase in musical content in various social media and other online platforms has facilitated data-driven studies of perceived emotions in music. The automatic determination and explanation of perceived emotions in music [20] is an active and major area of focus for the music information retrieval (MIR) community. A typical music emotion recognition (MER) pipeline following the machine learning approach involves four vital steps [21]: emotion taxonomy definition, dataset creation, features extraction, and training and evaluation. In this thesis, we attempt to explore and computationally formulate some of the pos-

sible factors underlying human perception of emotions in music and integrate them into each step of these data-driven studies of explainable MER. To this end, first, in our search for intuitive and statistically *representative* emotion representations for MER tasks, we study an intensity-based categorical emotion representation called the *Emotion-Word Intensity-Value* (EWIV) representation in chapter 3 We investigate its applicability in MER tasks, and capability to capture and represent perceived emotions in music using use *goodness-of-fit* measures for statistical models as our metric for *representativeness* over multiple datasets. We also introduce the novel *EmoRaga* clip-set, which consists of Hindustani classical music excerpts annotated with perceived emotion data and *emotion motifs*, which are key musical phrases or features that are discernible by the general audience and provide strong cues to listeners to perceive particular musical emotions. This facilitates the investigation into the dependence of perceived emotions on musical features and explainable MER studies in Hindustani classical music. Next, in chapter 4, we define the concepts of *immediate intrinsic musical context* and *intra contextual influence* to measure and analyze the possible effect musical context has on perceived emotions, with a focus on HCM, taking the first step towards modeling the subtle variation of emotion across different renditions of the same excerpts. Lastly, in chapter 5, we utilize deep learning to build explainable MER models that explain emotion predictions in terms of the musical frames, which provide context for the perceived emotions, relating the frames containing the *emotion motifs* and perceived emotions in HCM excerpts.

1.2 Objectives

Based on the challenges discussed in the above section, this thesis strives to address the following objectives.

1. In chapter 3 we establish how the selection of a proper emotion representation in MER is essential, both to facilitate better results in MER tasks and to improve the quality of emotion data captured. We propose the *EWIV* representation, a dynamic intensity-based categorical emotion representation inspired by the aesthetic concept of *Nava Rasa* [22, 23] in Hindustani

classical music (HCM). We investigate the capability of EWIV to capture perceived emotions in HCM and other music excerpts and statistically compare it with existing emotion representations for goodness-of-fit on emotion data.

2. It is essential to establish a dedicated HCM dataset to develop MER solutions for HCM. To this end, we create the *EmoRaga* excerpt-set, containing metadata and annotated *emotion motifs*. In chapter 3 we define an *emotion motif* as any perceivable musical feature that provides strong cues to listeners to perceive particular emotions. We capture emotion annotations using *EWIV* representation and use statistical analysis to investigate its effectiveness as an emotion representation for HCM excerpts.

3. In chapter 4, we consider a problem specific to *contexts* influencing perceived emotions in music. We use excerpts from *EmoRaga* excerpt-set to explore the effects of *immediate intrinsic musical context* on the perception of any musical piece. We term this possible phenomenon *intra-contextual influence*, and compare it with similar phenomena of emotion induction [24], contagion [10], and affective priming [25].

4. Lastly, in chapter 5, we present an attentive LSTM-based explainable dynamic emotion prediction model, initially using a benchmark MER dataset. We apply this to the proposed *EmoRaga* excerpt-set (chapter 3), to predict dominant and secondary emotions in HCM excerpts. We also compare the model-predicted important *frames* with those annotated with *emotion motifs* by experts, to investigate whether these are indeed captured by the model, explaining the emotion predictions in HCM.

1.3 Key Contributions

In the first work reported in this thesis, two issues pertaining to the first two steps of the MER pipeline [21] are discussed: a) the lack of standard mathematical analysis on the choice of emotion representation models for a particular study, and b) the lack of good quality annotated datasets in music traditions apart from

Western music - for example in Hindustani classical music (HCM). Generally, task-appropriate emotion taxonomy is chosen from psychology studies for representing emotions. The choice of an appropriate emotion representation for a given task is a widely debated, and open research topic since it defines downstream tasks like dataset formats and problem formulations. Though the categorical (e.g. [26]) and dimensional (e.g. [2]) representations are widely used in MER, they have considerable disadvantages. For example, in the categorical representations, the number of emotion classes might be too small to appropriately reflect the richness of music emotions [27]. Increasing the number of classes might lead to an increase in the cognitive load [28]. Again, in the dimensional representations, demarcations between emotions might be fuzzy [29], [30]. To alleviate some of these issues, and to arrive at a general representation of musical emotions that maximizes the information retained from the self-report data under a given modeling assumption, we study a dynamic (time-varying), intensity ratings-based, categorical schema, the *Emotion Word Intensity Values* (EWIV) representation. *Emotion-words* are terms that label our emotion-opinions, and are inspired by the *Nava Rasa* concept [22, 23] of Indian aesthetics. The *intensity* refers to the extent to which an emotion is perceived by a listener while listening to a piece of music. Various emotion estimates are proposed to represent static and dynamic emotion data, and the effectiveness of EWIV is demonstrated on existing datasets. The quality of reported emotion opinions is estimated and EWIV representation is also used to detect clips with ambiguous perceived emotions, establishing the utility of EWIV representation for analyzing perceived emotion data generated from listening to music. We explore the applicability of EWIV representations estimated from self-reported emotion data towards two MER applications - emotion classification, and temporal emotion variation detection. Lastly, we evaluate the *representativeness* of EWIV with respect to other representations, specifically, the dimensional Circumplex representation [2]. We use *goodness-of-fit* measures for statistical models as our metric for representativeness and observe that the *reduced EWIV* - a variant of EWIV - is consistently the best quality representation for perceived emotions, among four competing representations. The second step of the MER pipeline involves dataset creation, using the emotion representation chosen in the previous step. Though several benchmark datasets exist for MER

in Western music [5, 4, 3, 31, 32], some of which are also used in this thesis, to the best of our knowledge, no such dataset exists with Hindustani classical music (HCM) excerpts, perceived emotion annotations, and other relevant metadata. We use the established EWIV representation to annotate selected HCM excerpts with perceived emotions from crowd-sourced surveys. The emotion probability vectors are estimated and various metadata are included to create the *EmoRaga* excerpt-set. Apart from perceived-emotion annotations, it also contains *emotion motif* annotations, which are defined as key musical phrases or features that provide strong cues to listeners to perceive particular musical emotions. *Emotion motifs* form the basis of human-comprehensible musical structures in HCM that might be used for explaining MER results in HCM-related studies. This work is detailed in chapter 3.

In the second work of this thesis, reported in chapter 4, we examine the inherently subjective, personal, and highly contextual nature of music emotion perception through data-driven studies, based on the emotion opinion data collected for some of the excerpts of the *EmoRaga* excerpt-set using the EWIV representation. We specifically observe how changing contexts might change perceived emotions in music. Though *context* [33] has been established as one of the key factors affecting perceived emotions in music, most often external contexts (eg. socio-cultural, situations) have been investigated in the existing literature. The concept of *musical context* and its influence on stimuli-evoked emotions requires investigation in order to explain perceived emotions since music emotion perception is seldom an isolated process. We define the term *immediate intrinsic musical context* as the musical frame of reference within which a listener listens to a particular musical composition. It might include musical structures, notes or note-clusters, and small music excerpts which *immediately* precede the music piece being evaluated for an emotion judgment. The term *intrinsic* refers to the context being an inherent part of the music being heard, and not extrinsic (generated outside the music) like external events, audio, or otherwise. We explore the influence of *immediate intrinsic musical context* on the perception of music-evoked emotions. We term this possible phenomenon as *intra-contextual influence* - since the *context* investigated for being responsible for possible variation in the perception of music-evoked emotions is the preceding music itself. The term *intra* signifies the influence occurring *within* the

reference of a single temporal entity - music. The relevance of scrutinizing possible effects of *intra-contextual influence* stems from the fact that in most previous works with *musical context*, the target musical stimuli are treated in isolation. The effects of various contexts on the isolated stimuli, or the effects the stimuli have on other media are then studied. But in reality, very often musical compositions are created by weaving together numerous smaller pieces. These smaller pieces denote subtle variations in musical structure and emotions, which increase the aesthetic value of the whole. In isolation, each smaller piece might evoke a different emotion than when heard as part of a whole. Thus, understanding if and how the antecedent musical phrases/sequence of notes/musical sections influence the emotion perception in consequent music sections might help in comprehending the perception of music semantics and emotional meaning-making of subtle aesthetic and stylistic components in HCM and other forms of music. The main aim of the study is to explore the possibility of *perceived* musical emotions being *influenced* by *intra-contextual* variations, and include examinations of whether there is any significant effect of *intra-contextual influence* on music emotion perception, and if the possible effects of *intra-contextual influence* on the perception of musical emotions be generalized across music excerpts.

In the last work reported in this thesis, we address some of the issues in the last two steps of the MER pipeline involving relevant feature extraction from audio signals and training and evaluation of machine learning models for automatic emotion recognition. Particularly, we examine a) the dependence of perceived emotions on the intrinsic relationship between temporally distributed music segments and how to track them, and b) how to facilitate human-understandable explanations of the predictions made by black-box deep learning models for emotion recognition from music clips. Though many machine learning models have been proposed over the years which predict emotions with good metric values [20], most are unable to provide human-understandable justifications for their predictions. This gap leads to a loss of confidence in the trained models, which in turn affects their usability in real-life applications. Naturally, there exists an inherent requirement for human-understandable musically relevant explanations for MER models. Utilization of relevant context in the audio sequence is essential for effective dynamic prediction of perceived emotions of music. Existing methods have used LSTMs with mod-

est success. In this work, we describe three attentive LSTM-based approaches for dynamic emotion prediction from music clips. We validate our models through extensive experimentation on a standard dataset [3] annotated with arousal-valence values in continuous time and choose the best performer. We find that the LSTM-based attention models perform better than the state-of-the-art transformers for the dynamic emotion prediction task. We also explore individual smaller feature sets in search of a more effective one and to understand how different features contribute to perceived emotion. Through attention map analysis we visualize how attention is distributed over music clips' frames for emotion prediction. It is observed that the models attend to frames that contribute to changes in reported arousal-valence values and chroma to produce better emotion predictions, effectively capturing long-term dependencies. This work is discussed in chapter 5.

In this thesis, the core aim is to develop computational methods for modeling, predicting, and explaining perceived emotions in the domain of Hindustani classical music, using audience response data analysis and machine learning approaches.

1.4 Thesis Outline

The thesis is organized as follows.

- **Chapter 1**: This is an introductory chapter that presents the motivations and objectives of the thesis and briefly outlines the summary of the major contributions of the work done in the thesis.
- **Chapter 2**: Provides an overview of the basic concepts, relevant literature survey, and background study related to each of the contributions of the thesis.
- **Chapter 3**: In this chapter, we introduce the intensity-based categorical emotion representation for capturing perceived emotions. We show its effectiveness over existing datasets. We also introduce a novel excerpt-set for MER studies in Hindustani Classical Music, called *EmoRaga*, and identify various *emotion motifs* which might explain perceived emotions in HCM. Various statistical tests and goodness-of-fit tests are employed to mathematically justify the choice of emotion representations in MER studies.

- **Chapter 4**: This chapter outlines the concepts of *immediate intrinsic musical context* and *intra-contextual influence*. We use multiple *influenced* surveys over a set of dominantly happy and sad excerpts and compare the results to the baseline using descriptive statistics, ANOVA, and EWIV emotion probability vectors to understand whether *intra-contextual influence* effects perceived emotions in music. We also explore possible patterns of such influence.

- **Chapter 5**: In this chapter, we use deep neural networks to predict and explain dynamic (time-continuous) perceived emotions in the benchmark dataset and *EmoRaga* dataset. Various model architectures and feature sets are explored and attention mechanism is utilized to identify the temporal music frames which might affect perceived emotions, thus contributing to the explanations of emotions perceived.

- **Chapter 6**: This chapter summarises the work done and concludes the thesis while identifying possible directions for future work.

1.5 Publications

The publications from this thesis include the following:

Journal Accepted

1. Sanga Chaki, Priyadarshi Patnaik, Junmoni Borgohain, Raju Mullick, Gouri Karambelkar, Sourangshu Bhattacharya, "Exploring Intra-Contextual Influences in Music Emotion Perception", Psychology of Music, Sage Journals.

Journal Submitted

1. Sanga Chaki, Sourangshu Bhattacharya, Junmoni Borgohain, Priyadarshi Patnaik, Raju Mullick, Gouri Karambelkar, "Applicability of Intensity-based Categorical Emotion Representation in MER: A comparative study".

Conferences

1. Sanga Chaki, Pranjal Doshi, Sourangshu Bhattacharya, Priyadarshi Patnaik, "Explaining Perceived Emotion Predictions in Music: an Attentive Approach", International Society for Music Information Retrieval Conference 2020.

2. Sanga Chaki, Pranjal Doshi, Sourangshu Bhattacharya, Priyadarshi Patnaik, "Attentive RNNs for Continuous-time Emotion Prediction in Music Clips.", AffCON, The AAAI-20 Workshop On Affective Content Analysis, 2020.

3. Sanga Chaki, Sourangshu Bhattacharya, Raju Mullick, and Priyadarshi Patnaik, "Analyzing Music to Music Perceptual Contagion of Emotion in Clusters of Survey-Takers, Using a Novel Contagion Interface: A Case Study of Hindustani Classical Music.", International Symposium on Computer Music Multidisciplinary Research, 2017.

CHAPTER 2

Background & Related Work

Over the past decades, researchers have established that music and emotions are intrinsically linked with one another [28], [10], [11]. This holds true from the perspectives of music composers, performers, and the audience alike [34], [35], [9]. In fact, music-evoked emotion responses are known to constitute one of the main reasons for humans engaging with music [36]. Studies have proved that music may influence feelings [37], change or release emotions, help to experience enjoyment, and comfort, and relieve stress [9], [38]. It may impact our expressions, psycho-physiological reactions, and brain activation [36]. Thus, understanding and explaining music-evoked emotions - both perceived and induced - are challenging and interesting tasks among most researchers of multidisciplinary scientific areas like music psychology, affective computing, music information retrieval (MIR), and cognitive science. Though research on music and emotions has a long history [1], the field has recently received a newfound interest due to the development of modern technologies like machine learning, and digital signal processing, and the easy availability of large volumes of good quality musical content. This has in turn contributed to the growth of direct real-life applications of studies related to music emotions, like music information retrieval (MIR) and its sub-task, music emotion recognition (MER).

The work presented in this thesis builds on existing research in emotion representations, music emotion perception, and explainable music emotion recognition. This chapter aims to survey existing works and present background studies on these topics. In section 2.1, the emotion representations used in music-related studies are discussed, Section 2.2 presents the studies on the psychology of emotion perception in music and the importance of context in music emotion perception. In section 2.3, the existing approaches in MER are reported with reference to explainability in MER algorithms. We also provide a primer on Hindustani classical music (HCM), related emotion representation, and existing MER literature based on HCM in section 2.4.

2.1 Emotion Representations

Psychology literature has provided us with different emotion representations, which encode emotions in a systematic way for the purposes of affect analysis and study. In the context of music emotion recognition, various emotion representation models are used, which are discussed below.

2.1.1 Categorical Representation

The categorical paradigm or discrete emotion model labels music-evoked emotions into a number of discrete classes [27], [26, 39],[1]. The notions of primary emotions in Ekman's Basic Emotions [26, 39], Hevner's adjective checklist [1], and Geneva Emotional Music Scale (GEMS) [40] are some examples. Ekman [26] proposed the existence of six basic emotions: happiness, sadness, fear, disgust, anger, and surprise, based on evolutionary considerations, universal facial expressions of emotions, and semantic distinction. Though this model is widely used in human-emotion studies and has also been applied for MER [41], one drawback is that it becomes rather difficult to represent the nuanced emotions that music-emotion studies demand. Hevner's Adjective Circle [1] is specifically designed for music-evoked emotions and consists of grouped adjectives arranged in the form of a circle. Each group includes adjectives that are close in meaning and form a distinct emotion category. Neighboring groups are emotionally close, while groups on opposite

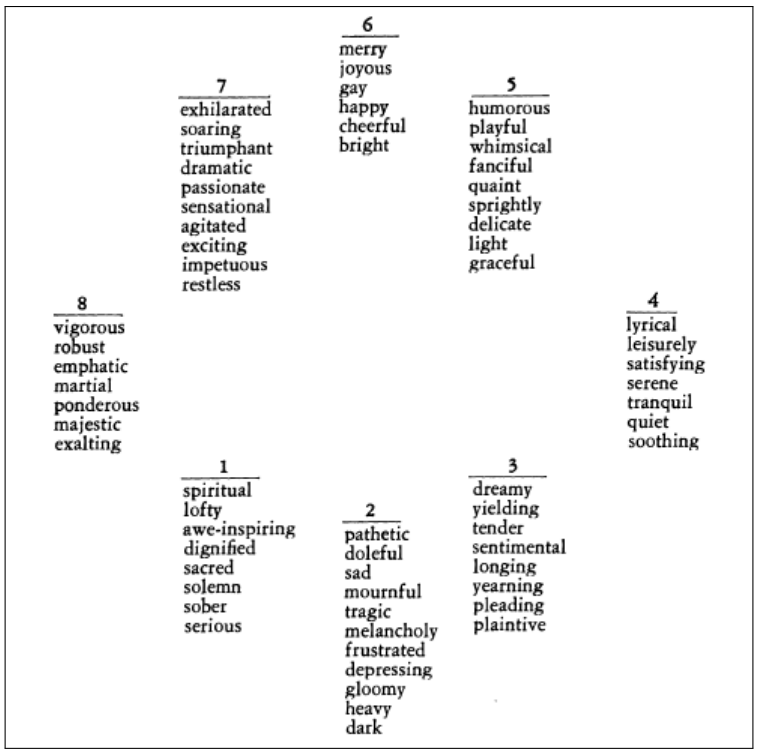


Figure 2.1: Hevner's adjective circle [1]. Each cluster contains adjectives with similar meanings in terms of music emotion. Neighboring clusters represent close emotions.

sides of the circle represent contrasting emotions. The GEMS (Geneva Emotional Music Scale) [40] includes nine categories of musical emotions: wonder, transcendence, tenderness, nostalgia, peacefulness, energy, joyful activation, tension, and sadness. These can be further divided into 45 emotion terms, or clustered into 3 super-factors (sublimity, vitality, and unease). The various categorical models are extensively used in many MIR tasks like music emotion classification [42], dataset creation [43], studies of human perception of music [44], [45], [46] etc. But, it is also recognized that the number of primary emotion classes might be too small to encompass the richness of music-evoked emotions perceived by humans [27, 47]. Increasing the number of classes might not solve the problem [9], and lead to an increase in the cognitive load due to inherent ambiguity in emotion-words, and render the study impractical [28].

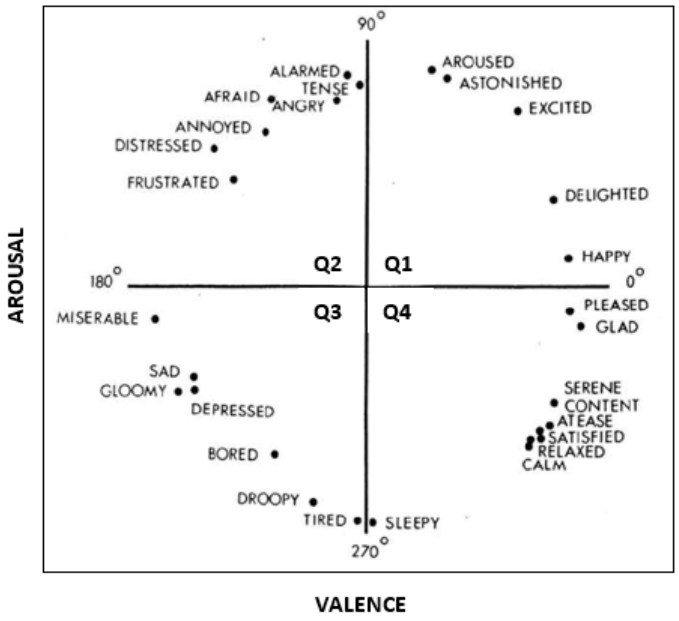


Figure 2.2: 2 Dimensional Circumplex model [2] - Positions of emotion words in the 2D arousal-valence plane.

2.1.2 Dimensional Representation

To overcome the limitations of categorical emotion models, the concept of an emotional space that lies in a continuum was created, resulting in the dimensional models of emotion. The dimensional approach [39, 48, 27] identifies emotions from coordinates of two or three dimensions like *valence*, *arousal* and *dominance*, based on neuro-physiological systems of the brain. In Russell's Circumplex model [2], emotions are characterized by valence and arousal (figure 2.2). Arousal is the level of activation that an event evokes, ranging from calm (low arousal) to excited (high arousal) [49]. Valence is the level of pleasantness (positive valence) or unpleasantness (negative valence) that an event evokes [49]. In many studies, this two-dimensional plane is divided into four quadrants (figure 2.2): a) Q1 (positive valence, high arousal): representing the group of happy, excitement, and other energetic emotions, b) Q2 (negative valence, high arousal): representing anxiety, distress, and other frantic and energetic group of emotions, c) Q3 (negative valence, low arousal): representing depression, melancholy, and other sad group of

emotions, and d) Q4 (positive valence, low arousal): representing contentment, calmness, and other positive emotions [50]. The dimensional model is used extensively in research for various purposes - to understand and model emotions [51], [52], dataset creation [3], [31], music emotion recognition [53], [54], mood prediction [55] etc. But it is also known that dimensional models blur psychological distinctions between some emotions [30], for eg. anger and fear are placed very near in the 2-D plane (fig 2.2) but are very different emotions. The demarcations between certain emotions are fuzzy [29]. Different versions sometimes propose different locations for affective states [56]. Use of the *dominance* dimension might provide a more complete picture of emotions [57], but it might also be more difficult to visualize the emotion concepts [27].

2.1.3 Combined Emotion Representations

While most studies use a single emotion representation model, some prefer to use either both categorical and dimensional models simultaneously, or combinations of characteristics from both models. The motivation for such use is either task-specific or overcoming the disadvantages of any single approach. Eerola et al. [58] was one of the early works to advocate this. Lee et al. [15] used both arousal-valence (A-V) values and mood tags to propose a mood-based music recommendation player *Smoodi*. They used 114 independent mood tags to classify the audio clips into a predetermined number of clusters, within which, the distribution of arousal-valence values was used to recommend music. The combined representation was utilized to corroborate the emotion labeling of individual representations. More recently, Shepstone et al. [16] worked on granularity-adapted emotion classification in audio, used to drive a recommender. They used an evaluation technique, where subjects were expected to rate a set of 12 predefined emotion categories on a scale of 1 to 7, which measured the extent to which the stimulus invoked a certain emotion. They proposed the concept of *adapted class* which is an aggregate of underlying emotion classes, mapping to larger regions in valence-arousal space, from which a list of potentially more similar content items are drawn for the recommendation. The combined representation helped in the conceptualization of the *adapted emotion class*, which aided in exploring emotional

similarity between music clips better. These associations might have gone unnoticed if individual representations were used instead. Parada et al. [59] utilized basic emotions and A-V values to define emotions like *hot anger*, *elated happiness* etc., and created an emotion annotated corpus of a Cappella opera singing for identifying emotions under adverse acoustic conditions. The combined representation made the distinction between real and distractor emotions possible, where the real emotions were those which were effectively expressed by the music. Panda et al. [43] used a 2D quadrant-based approach and mood tags to perform music emotion classification. A similar quadrant-based representation was also utilized by Malheiro et al. [60] for their study on emotionally relevant features for the classification and regression of music lyrics. Hu et al. [61] explored music emotion prediction from users' physiological signals using both mood categories and A-V ratings. A summary of these studies is provided in table 2.1. These studies used concepts from both categorical and dimensional approaches to emotion representation. Thus, it is evident that there is some understanding that these approaches considered in conjunction (in some sense) can carry more information than when used individually.

2.1.4 Category-Intensity Representations

Further, some studies include the concept of intensity annotations along with the categorical approach of emotion representation. It is inherently understood that this provides more information than simple emotion tags. The *intensity* here refers to the extent to which an emotion is either expressed by a piece of music or is perceived by a listener unambiguously, while listening to a piece of music. Emotion intensity annotations are different from arousal annotations of the Circumplex representation model [2]. In order to simplify the difference, we might consider an example of a pair of music excerpts, one that expresses a moderate level of sadness, and the other that expresses a high level of sadness. The first one will be rated with a lower intensity of sadness than the second one by listeners. In contrast, in this context, *arousal* refers to the extent to which one perceives that the music expresses a tense or agitated affective state or a relaxed or calm one. Thus, the second excerpt, though perceived as a higher intensity of sadness,

Table 2.1: Summary of literature survey on studies using combined emotion representations. Column *Study* gives the name of the study referred to. Columns *#Clips* and *#Subs* refer to the number of clips and the number of participants in each study.

Study	#Clips	#Subs	Category Component	Dimension Component	Applications
Eerola et.al. [5]	110	116	5 discrete emotions	Arousal, Valence	Comparing perceived emotion in music using both representations and appropriate dataset creation.
Shepstone et.al. [16]	1260	15	12 classes	Arousal, Valence Focus	Granularity-adapted Emotion Classification of audio. Classification-related studies.
Lee et.al. [15]	446	15	8 Mood categories	Arousal, Valence	Mood based music recommender
Parada et.al. [59]	390	132	10 classes	Emotion intensity labels	MER under adverse acoustic conditions
Panda et.al. [43]	900	-	200 mood tags	Quadrant info	MER and exploration of texture features.
Malheiro et.al. [60]	200	39	92 Mood tags	Quadrant info	Music lyrics emotion recognition and exploration of emotionally relevant features.
Hu et.al. [61]	628	23	10 classes	Arousal score in range (-10, +10)	Study interrelation between emotion responses to music and physiological signals.

might also be a subdued emotional state, with low arousal. Eerola et al. [5] compared the intensity-based discrete emotion model with the dimensional model in their seminal work. Recently, Shepstone et al. [16] used intensity-based emotion category ratings for computing individual valence-arousal focus. It might be noted that, the above studies use *static* intensity ratings for discrete emotions - a single intensity rating for each emotion descriptor over the entire stimulus. Thus, an interesting research question that might remain unanswered is: *Is the information*

content of discrete emotion categories with time-continuous intensity ratings higher than other representations?

As evident from the above discussion, a significant number of studies exist based on these emotion representation models, since the description, measurement, processing, and storage of perceived-emotion opinion data form an essential part of studies regarding music and emotions. They control the information content extracted from the emotion annotations given by the subjects and also define downstream tasks like dataset formats and problem formulations, which can then be used by the affective computing community to develop new technology for specific MIR tasks. Thus, an appropriate choice of emotion representation models is essential, both to facilitate better results in MIR tasks and to improve the quality of emotion data captured. However, this choice of emotion representation for a given task is still a widely debated, and open research topic. Hence, the analysis and comparison of emotion representations for broad applicability in MIR is an important research question, which is addressed in chapter 3 of this thesis. We introduce an intensity-based categorical emotion representation called *EWIV* for capturing perceived emotions. We show its effectiveness over existing datasets. Various statistical tests and goodness-of-fit tests are employed to mathematically justify the choice of emotion representations in MER studies. We also propose a novel HCM clip-set called *EmoRaga*, emotion annotated using EWIV representation, and use it for various experiments.

2.2 Emotion Perception in Music

Emotion in music can be studied under three paradigms [62], [27], [63], which are: a) transmitted/expressed emotions: the emotion that the performer or composer wanted to convey, b) perceived emotions: the emotion that an individual listener identifies when listening to a particular music excerpt, and c) felt/induced emotion: the emotion that an individual listener feels when listening to a particular music excerpt. In this thesis, we focus on the perceived emotions from music.

To examine the inherently subjective and highly contextual nature of music emotion perception, one first needs to comprehend the difficulty of the task arising from various factors [62]. The perception of emotion is innately personal. Differ-

ent people may perceive different and varied emotions when listening to the same song or music excerpt. Even if there is considerable agreement between listeners' emotion opinions, there is often ambiguity in the terms used regarding emotion description and classification. Thus, accurately capturing all these variations, and the relation between music and emotions is a complex problem. The dearth of public, widely accepted, and adequately validated, benchmarks to compare works on music emotion perception in music traditions apart from Western music also hinders progress in the field. Most importantly, it is still not well understood how and why some musical elements and constructs elicit specific emotional responses in listeners. *Musical context* is one such aspect, which is comparatively less explored. Over the past decades, research has begun to study the interrelations of the complex set of different factors influencing musical emotions [58]. However, to the best of our knowledge, there is a dearth of systematic investigation on how music itself can affect emotion perception and eventual meaning-making in successive music. Thus, a data-driven study of musical *context* and how it affects emotion perception in music is detailed in chapter 4 of the present thesis.

2.2.1 Emotional Meaning Making in Music

Among many existing studies that have been conducted over the years to explain musical emotions, mention must be made of theories of cognitive appraisal, musical expectancies [64], and the unified theoretical framework of BRECVEMA - an acronym for 8 mechanisms that attempt to explain music-evoked emotions: Brain Stem Reflex, Rhythmic Entrainment, Evaluative Conditioning, Contagion, Visual Imagery, Episodic Memory, Musical Expectancy, and Aesthetic Judgment [11], [65]. Hargreaves et al. [66], [33] put forward the *reciprocal-feedback* model of responses to music, where three main determinants of musical response were considered: a) music related properties [67], [68], [69], [70], b) listener-specific properties [71], [72], [73], [74], and c) the listening situation (context related). They proposed that any one of these three main determinants of musical response can simultaneously influence the other two, and the influences can work either way. Substantial research on the first two factors has been done. The music-related parameters include mode, meter, tempo [67], [75], [9], [68], compositional structure ,

performance expression [69], temporal variation, intensity, mean centroid, vibrato rate [70], vocal features [76], [77] and so on. The listener-specific traits which have been explored are personality [71], [78], [79], [72], [80], mental disorders [73], socio-demographic factors [74], [81], [82], culture [83], [84], music preference [74], musical training, physical impairments [85] and so on. Ruth et al. [86], [87] further extended the reciprocal feedback model of musical communication [33] along with the *general learning model* (GLM) [88] to propose the *music process model*. In this study, we are particularly interested in the less explored *context* related determinants.

2.2.2 Musical Context and Perceived Emotions

Some recent studies have focused on *context* as a relevant parameter [33] when trying to explain emotions evoked by music. The term *context* might have many meanings and connotations in relation to music-emotion studies. Hargreaves et al. [33] identified four types of *contexts* - socio-cultural [79], [89], everyday situations [90], presence/absence of other person(s) [78], and other activities [91]. Context may refer to the listening location and related activities [91], [92]. The context might also be social [78], where the presence or absence of a close friend/partner is considered to affect emotional reactions to music. Social context may also include family climate, bullying issues, etc. [79]. Coutinho et al. [89] explored the effect of performance setting as context to emotions elicited by a musical performance. Greb et al. [90] found that on the situational level, the context of the main activity while listening to music showed the greatest impact, while on the individual level, the intensity of music preference was most influential on the functions of music listening. Lyrics and types of lyrics have also been explored as possible contexts depending on which music-evoked emotions and thoughts can vary [87]. Thus it is undeniable that various types of *contexts* might have a significant influence on the way music-evoked emotions are perceived.

The concepts of *musical context* and its influence on stimuli-evoked emotions are relatively new. Steffens et al. [93] and Herget et al. [94] explored the influence of *musical context* on video-evoked emotions. In these studies, *musical context* refers to the music being played along with the target video. In some studies

[95], *musical context* includes various earlier musical experiences, music-related memories, and responses to music. Guevara et al. [95] explored this to put forward the *constructionist approach* of emotional meaning attribution to music, drawing from the constructionist theories of emotion and musical meaning [96], [97], [98]. These studies indicate a rather fluid interpretation of the term *musical context*.

2.2.3 Affective Priming, Induction, and Contagion

Discussions of context and emotional responses to stimuli are incomplete without mention of affective priming [25], emotion induction [24], and contagion [10]. In the case of emotion induction [24] only a specific mood state is instigated. Same (or almost similar) emotions are propagated from the emotion induction medium (music) to the mood induction target (listener). Emotional contagion [10], [51] is a process whereby an emotion is induced [24] by a piece of music, because the listener perceives the emotional expression of the music and *mimics* this expression internally. Priming is the procedure that entails that exposure to one stimulus may influence a response to a subsequent stimulus, without conscious guidance or intention. Timmers et al. [99], [100] investigated how listeners' perception of auditory sequences (target) change dynamically, depending on emotional context primed through affective pictures (prime) of people depicting emotional expressions. Huziwara et al. [101] investigated whether affective priming occurs when chords are used as primes, and faces (happy or sad) are used as targets, and how listeners' perceptions of auditory sequences change dynamically depending on emotional context. They came to the conclusion that primed emotion modulates melodic expectations, and that listeners adapt their attention to emotional context. This proved that emotion is indeed an intrinsic part of music perception and not merely a product of the listening experience. Here, the primes were affective pictures, and the target was music. Tay et al. [102], Steinbeis et al. [103], Armitage et al. [104], [105], and Goerlich et al. [106] investigated various aspects of auditory priming and its effect on usage, processing, and perception of words. They explored how auditory priming can influence induced emotions and the use of emotion words. Here the prime was mode and tempo-manipulated music, and the target was emotion words. They showed that major modes and faster tempos

elicited greater responses for positive and high arousal words, while minor modes elicited more high arousal words, highlighting the prominence of affective auditory priming and allowing us to better understand emotive responses to music. Through affective priming paradigm, Steinbeis et al. [103] showed that participants evaluated emotional words congruous to the affect expressed by a preceding chord faster than words incongruous to the preceding chord, for specific musical parameters. Here also, the prime was music, and the target was words. Similar studies were also reported by March [107]. Armitage et al. [104], [105] reported reaction time data to identify word valence, using music as prime to understand music cognition, and to determine which emotional dimension is transferred across modalities. Goerlich et al. [106] showed that both affective music and speech prosody can prime the processing of visual words with emotional connotations, and vice versa. In all these cases, the prime and the target are different media, essentially capturing the effect of affective priming by music on the perception of other stimuli or vice versa. In chapter 4 of this thesis, we aim to study the effect of preceding musical stimuli (as context) on the emotion perception of successive musical stimuli.

2.3 Music Emotion Recognition

The interdisciplinary topic of music information retrieval (MIR) deals with retrieving various relevant information from music, which includes different musical characteristics. MIR has broad applicability in discovering and organizing media collections, searching songs and other audio-based content, and creating audio-based consumer products. Some of the prominent MIR tasks are audio fingerprinting [108], genre recognition [109], music recommendation based on particular musical characteristics [110], and music emotion recognition (MER) [62] among many more.

For the last few decades, the amount of available digital music content has exploded, leading to an ever-increasing demand for easy and effective music information access. Since the emotive aspect of music is undeniable, music organization and retrieval by emotion is considered a reasonable method [34] to access desired musical content. Music emotion recognition (MER) aims at developing algorithms

capable of recognizing the emotional content in music, or the emotional impact of music on a listener. It is an interdisciplinary area drawing inputs from various fields like psychology, affective computing, cognitive science, audio signal processing, machine learning, and natural language processing. Some of the different perspectives of MER include classification of song excerpts [34],[111], emotion variation detection [112], automatic playlist generation [113] and others. MER has broad applicability in search-and-retrieval systems for music, streaming platforms, affective music generation, and recommender systems. Of the significant corpus of research works present currently in the field of MER, we discuss some of the major studies below.

2.3.1 Types of MER Studies

Existing work on MER can be divided into two broad approaches [20]: song-level MER (or static MER) [114], and music emotion variation detection (MEVD, or dynamic MER) [27], [20]. Song-level MER [114] assigns one (or more) emotion labels(s) to the entire song or the entire music excerpt under consideration. MEVD considers music emotion as a dynamic process, and recognizes emotions at each defined time segment of a song, resulting in a series of emotion predictions. Each type of study can use either categorical or dimensional emotion representation models. Accordingly, the objective formulation of the study changes [20]. For example, the MER task can be formulated as a classification problem using categorical representation, or a regression problem using a dimensional approach. Datasets, algorithms, and performance metrics used might all differ significantly in each case. In this thesis, we will primarily focus on dynamic emotion recognition, and discuss in detail a novel method for the same in chapter 5.

2.3.2 General Research Framework for MER

Most MER research nowadays is based on the machine learning paradigm, and as such follow a general four-step framework [20], [21]. The first step consists of the emotion taxonomy definition, where the appropriate emotion representation is chosen for the MER task. As discussed in section 2.1, it is crucial to choose the best-suited taxonomy for the concerned application, since all further steps, in-

cluding problem formulation, machine learning approach used and result metrics depend on this step. In the second step of dataset creation, a labeled dataset is constructed with music excerpts (or songs) and associated emotion opinion ratings and/or labels. The data is generally collected through online crowd-sourced interfaces or controlled listening environments. We discuss some of the existing datasets in MER in section 2.3.3, a few of which are also used in different experiments reported in this thesis. Though many such datasets based on Western and other traditions of music exist in MER literature, there is a dearth of such well-annotated datasets for MER based on Hindustani classical music (HCM). Thus, in chapter 3, we introduce a new set of clips from HCM, called the *EmoRaga* clip set for perceived emotion analysis in HCM. The *EmoRaga* clip set consists of HCM audio excerpts, emotion opinion data, and other metadata related to musical characteristics which will help in MER studies on HCM. The third step comprises feature extraction. In MER, features directly affect the accuracy of emotion recognition. Some features like the signal processing-based audio features and their statistical properties can be extracted from the audio signal directly. Some features are generated by listeners, like physiological data. These features are pre-processed to be used by a machine learning algorithm during training and evaluation. In some learning-based systems, feature extraction is also a part of the machine learning model itself. We discuss the relevant features in section 2.3.4, and use them in chapters 3 and 5 of this thesis for the proposed MER systems. In the last step, training and evaluation of the machine learning model for MER are performed. A machine learning model is trained on the training subset of the annotated dataset. It is validated and evaluated over the validation and test sets of the annotated dataset and evaluated using the appropriate performance metrics. The training and evaluation of the proposed machine learning model in this thesis is detailed in chapter 5

2.3.3 Datasets for MER

An MER dataset is a collection of data, generally of the form *music excerpt-emotion annotation*, that is used to train the machine learning model. Though some MER datasets are available in the literature, due to music copyright restric-

tions, many MER researchers prefer to use self-built and unpublished datasets. One of the earliest music perceived-emotion studies was reported by Eerola et al. [5] in their seminal work, where they collected emotion data in both discrete and dimensional representations for a set of 110 excerpts, and compared the two models. Schubert et al. [4] used six music extracts from film music, each targeting one of six discrete emotions: Excited, Happy, Calm, Sad, Scared, and Angry. They utilized discrete facial expressions as a response interface to capture discrete annotation tags for the extracts. They observed the presence of a second competing emotion in most of their excerpt annotations and explained it as *near miss*, concluding that some emotions might be confused. They attributed this to the adjacent placing of discrete emotion faces or probable improper calibration of selected excerpts to putative target emotions [115]. Soleymani et al. [3] proposed the benchmark *1000 songs for emotional analysis of music* dataset annotated with *static* and time-continuous *dynamic* arousal-valence values. These three datasets are especially relevant for this thesis and have been used in chapters 3 and 5. Other popular datasets include the AMG1608 dataset [31], the Greek music dataset [116], CAL500 [117], DEAM [118], MTurk [119], Emotify [40] and the IADS dataset [32]. To the best of our knowledge, no such dataset exists with Hindustani classical music (HCM) excerpts and dynamic perceived emotion annotations, which hinders MER studies with HCM. We address this issue in chapter 3 of this thesis and propose the *EmoRaga* clip-set which consists of selected HCM excerpts, annotated with emotion opinion data that is utilized to develop dynamic MER models for HCM clips. We also identify various *emotion motifs* particular to HCM, which might explain perceived emotions in HCM.

2.3.4 Relevant Features for MER

Feature extraction is a core concept in MER, as the performance of automatic emotion recognition directly depends on the quality of features. MER tasks are generally found to rely on four kinds of features: audio features, symbolic features, lyric features, and biological features [20]. Of these, audio features are the most extensively used in MER studies. In this thesis, we have used the 2013 Computational Paralinguistics Evaluation (ComParE) tasks feature-set [120]. It contains

6373 features, consisting of various affective low-level descriptors (LLD) of audio signals and their statistical functionals [121, 112]. An acoustic LLD is defined as a parameter computed from a short time frame of a given length, from an audio signal at a given time [122]. Affective audio features (LLD) can be broadly divided into many groups [20] - rhythmic, timbre, spectral, energy-based, melody, etc. In the ComParE tasks feature-set [120], the features are: a) spectral-based, b) energy-based, c) voicing related. These features are generally extracted from audio waveform files with the help of existing tool-kits [123], [124]. For the purpose of this thesis, we have used the *openSMILE* feature extractor [125], and the Librosa [126] feature extractor in chapters 3 and 5 to extract the relevant features from the existing datasets as well as the proposed EmoRaga clip-set.

2.3.5 MER Performance Metrics

The evaluation metrics for MER depend on the type of model used for the task. The metrics commonly used for classification-based MER are accuracy and precision. Accuracy calculates the proportion of correctly classified samples to the total number of samples. Precision gives the proportion of the real positive samples to the total number of samples predicted to be positive [20]. For regression-based MER, the metrics generally used are: Coefficient of determination (R^2), average Kendall's τ per song ($\overline{\tau}$), mean absolute error (MAE), and root mean square Error (RMSE). Since we formulate MER as a regression-based task in both chapters 3 and 5 of this thesis, we detail the metrics for regression-based MER below.

2.3.5.1 Coefficient of determination (R^2)

The determination coefficient (R^2) is a key output of regression analysis, which provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. It can vary between 0 and 1. If a data set has n values marked ($y_1 \dots y_n$), and each associated with a predicted value ($f_1 \dots f_n$). So, R^2 is defined as

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.1}$$

where,

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{2.2}$$

and

$$SS_{tot} = \sum_i (y_i - \overline{y})^2 \tag{2.3}$$

given

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.4}$$

2.3.5.2 Kendall's τ

Kendall's τ per song ($\overline{\tau}$) is a measure of how well the emotional profile of each song is captured by the regressor, as opposed to the overall correlation. It measures the correspondence between two rankings. Values close to 1 indicate strong agreement, and values close to -1 indicate strong disagreement. It is defined as

$$\overline{\tau} = \frac{P - Q}{\sqrt{(P + Q + T) * (P + Q + U)}} \tag{2.5}$$

where, P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in target set $(y_1 \dots y_n)$, and U the number of ties only in predicted set $(f_1 \dots f_n)$.

2.3.5.3 Other Metrics

Root mean square error (RMSE) is an often-used measure of the differences between values predicted by a model or an estimator and the values observed or true values or ground truths. It is defined as the square root of the mean squared error.

The mean absolute error (MAE) is a measure of errors between paired observations, for example, comparisons of predicted versus observed values. MAE is calculated as the sum of absolute errors divided by the sample size

2.3.6 Related Works in MER

In the past, most song-level categorical MER systems used features of timbre, pitch, MFCCs, and/or lyrics and applied classifiers like support vector machines (SVM) [127], [128], K-nearest neighbors, decision tree, naïve Bayes [61], and random forests (RF) [129]. For the song-level dimensional MER systems, the most commonly applied methods were support vector regression (SVR), multivariate linear regression [130], and Gaussian process regression [131, 132]. The most common choices of machine learning models for MEVD tasks were SVM, SVR [133], Gaussian mixture model [134], and MLR [132]. Kim et al. [127], Yang et al. [27] provide an extensive survey of early MER studies.

Current state-of-the-art methods for music emotion prediction are mostly based on deep neural networks like long short term memory based recurrent neural networks (RNN-LSTM) [112] and convolutional (CNN) [135]. In this thesis, we focus on the RNN-LSTM-based models in chapters 3 and 5 for our MER solution. Most of the MER tasks utilizing deep RNN-LSTM structures need a considerable amount of training data to produce good results. With the increase in dataset sizes and related experiments, techniques such as dimension reduction (eg. PCA) started being implemented for better emotion modeling. Coutinho et al. [136] proposed the use of this model for this task. Weninger et al. [112, 137, 53] used RNN-LSTM networks successfully to perform continuous time music emotion regression, using a modified cost function, on the *1000 Songs for Emotional Analysis of Music* dataset [3]. Giamusso et al. [138] used neural networks to predict playlist emotions based on lyrics. Fan et al. [139] performed ranking-based emotion recognition from experimental music. Delbouys et al. [55] used LSTM and ConvNet models on the Million Song Dataset [140] for audio and lyrics-based bimodal music emotion detection. Han et al. [20], provides an extensive survey of current deep learning-based MER studies.

2.3.7 Explainability in MER

The term *explainable machine learning* indicates the methods and techniques to extract relevant knowledge from a machine learning model concerning relationships contained in the data and learned by the models. The terms *explainability*

and *interpretability* are often used interchangeably. *Interpretability* is defined as the degree to which a human can understand the cause of a decision by a predictor model [141] and the degree to which a human can consistently predict the model's result [142]. Explainable methods can be local or global, based on the scope of the application, or they can be model-agnostic or model-specific. Interpretability comes by design in some algorithms where there is a clear and intuitive understanding of the decision-making process, for example, in linear regression there is a simple linear relationship between the input and the output. But most models are *black-box* in nature, and interpretability is not built-in. For these models, different methods like attention [143], LIME [144], Grad-CAM [145], Shapley Values [146] are used.

Many MER models use deep neural network architecture to achieve good performance results [136], [112, 137, 53], [138], [139], [55], [147] (section 2.3.6). The need to ascertain how these black-box models arrive at their decisions has led to research in the area of explaining and interpreting the model predictions in human-understandable ways. Explainable models help build trust in individual predictions as well as the capability of the model as a whole. Recently, Chowdhury et al. [147] proposed a VGG-style deep neural network to predict emotional characteristics of the Soundtracks dataset [58], based on human-interpretable, mid-level perceptual features. These seven mid-level features [148] include melodiousness, articulation, rhythmic complexity etc. and represent musical characteristics that are easily perceived and recognized by most listeners, without any music-theoretical training. Haunschmid et al. [149] took forward the mid-level features-based explanations presented in [147], and used LIME [144] to generate audio-level explanations in terms of image regions of the input spectrogram, derived from the input audio. Berardinis et al. [150] proposed a new computational model (EmoMucs) that used source separation algorithms for better and interpretable valence and arousal regression. The studies reported above use either special annotated mid-level features [147] or source separation of input audio [150] for generating human interpretable explanations for the respective music emotion prediction models. In this thesis, we use the attention mechanism to explain the perceived emotion predictions, in terms of attention distribution over frames of each music clip, for emotion prediction, detailed in chapter 5.

From the above discussion, though it is evident that a considerable body of literature exists on the various perspectives of MER, still some challenges remain. Particularly, we examine the dependence of perceived emotions on the intrinsic relationship between temporally distributed music segments and how to track them, and how to facilitate human-understandable explanations of the predictions made by black-box deep learning models for emotion recognition from music clips in chapter 5 of the present thesis.

2.4 Emotions and Hindustani Classical Music

Traditional classical music in India has two main branches - Hindustani (North Indian) classical music and Carnatic (South Indian) music. In this thesis, we focus on Hindustani classical music (HCM). Some basic concepts of HCM, relevant to the present work are discussed below.

2.4.1 Raga, Tala, and Laya in HCM

HCM is primarily based on the *raga* framework [151]. Each *raga* is characterized by a set of notes, the ascending-descending melodic progression and a specific set of melodic phrases termed *raga motifs* [152]. Ragas are said to be potentially capable of evoking distinct emotions [153]. These *ragas* are themselves considered to be built on a large set of scales, called *thaats* [154]. *Thaats* are the parent scale for *ragas* and form a basis for classification of ragas. Traditionally, ten such *thaats* are prevalent - Bilaval, Kalyan, Khamaj, Bhairav, Kafi, Asavri, Bhairavi, Marwa, Poorvi, Todi. Each *raga* belongs to one *thaat*.

Other primary aspects of HCM include the *tala* (rhythmic cycle) and the *laya* (tempo). In any HCM performance, the *raga*, *tala* and *laya* form the grammatical basis, on which, individual artist apply their interpretation and creativity. Generally, a raga performance has an *alaap* (prelude) section, which is not bound by any rhythmic cycle (*tala*), and a *gat* section, which follows a rhythmic cycle. The *gat* section can be of varying tempos (*laya*) - *Vilambit* (slow tempo), *Madhya* (mid tempo), *Drut* (fast tempo). Although the total number of *ragas* and *talas* are a subject of debate and open to interpretation, it is often considered to be 300

and 108 respectively, not all of which are regularly performed and some are said to be lost. In chapter 3, section 3.4.1.2 we introduce the *EmoRaga* excerpt-set for perceived emotion analysis of HCM. It consists of 48 excerpts, from 23 *ragas* using 4 *talas*. The primary aim was to include a) at least two prominent *ragas* from each of the *thaats* and b) one excerpt for every *raga*, using one of the most prominent *talas* in HCM, - *Tin Taal*. It might be noted that, other combinations of *raga* and *tala* are also possible.

2.4.2 Emotion Representation in HCM

In Indian aesthetics, emotional responses to any art form are said to be one of nine prevalent types, derived from the concept of *Nava Rasa* [23], [155, 22], which translates to Nine (= *Nava*) Emotions (= *Rasa*). The Sanskrit word Rāsā is a concept in Indian arts that describes the aesthetic flavor of any visual, literary, or musical work which is capable of evoking an emotion or feeling in the reader or audience. The *Nava Rasa* theory discusses nine primary emotions or affective states which are as follows: *Shringar* (romance), *Hasyam* (happiness/joy), *Raudram* (anger), *Karunyam* (sadness), *Bibhatsam* (disgust), *Bhayanakam* (fear), *Veeram* (courage), *Adbhutam* (wonder) and *Śāntam* (calmness). These are the set of possible emotions that are perceived by, or evoked in a music listener [155]. Many of the *raga*s are said to be associated with particular emotions, which are elicited through the use of different melodic and other musical constructs particular to HCM. We have termed these as *emotion motifs* in chapter 3 of this thesis. Some of the emotion options used to collect emotion opinion data in 3 are also inspired by the *Nava Rasa* concept. Though significant studies exist based on the popular emotion representation models (section 2.1) on various music traditions, their effectiveness on Hindustani classical music (HCM) is relatively less studied. On the other hand, the *Nava Rasa* concept in Indian aesthetics provides emotion categories to denote aesthetic emotions, which have been explored sparsely in MER. In this thesis, in chapter 3, we study an intensity-based, categorical emotion representation called the Emotion-word and Intensity-Value (EWIV) representation, where the emotion-words are taken from the *Nava Rasa* concept: Fear (F), Anger (A), Sadness (S), Calmness (C), Wonder (W), Romance (R), and Happiness

(H). These seven were chosen as they were found to be most frequently perceived by HCM listeners in our studies. The other two rāsās - *Heroism*, *Disgust* - are excluded as their perception was found to be almost negligible in HCM excerpts of the present study. We also introduce a set of clips from HCM, called the *EmoRaga* clip set for perceived emotion analysis in HCM.

2.4.3 Importance of Musical Context in HCM

In any standard HCM *raga* rendition, the composition consists of an arrhythmic phase followed by rhythmic phases of varying tempo, each forming parts of a whole. The perception of one part plays a significant role in how subsequent parts are perceived, building up to an emotional crescendo at the climax. Compared to Western traditions, HCM renditions focus on the intensification of a cluster of related/congruent emotions. The use of contradicting/incongruent emotions in musical phrases is widespread to enhance the perception of other emotions and increase aesthetic appreciation. In the *Rasa* theory [22], this is termed the intensification of the *sthayee bhava* (dominant emotion) through the use of a *sanchaari bhava* (a glimmer of a contradicting emotion) as an impulse. Very often, hope and despair, romance and sadness, or excitement and calmness are paired in subtle ways, to intensify the perception of emotions. It is also a common practice in HCM renditions to improvise the same music phrase in different ways, evoking subtle nuances of perceived emotions. The study of such aesthetic stylistic components and their influence on emotions perceived by listeners is possible through concepts of *immediate intrinsic musical context* and *intra-contextual influence* described in chapter 4 of this thesis. Though many studies on the effects of Indian music on human psycho-physiological [156], [157] and emotional [153], [158] responses have been undertaken, to the best of our knowledge, no such systematic study exists on the influence of *contexts* on perceived emotions in HCM, compared to its western counterpart.

2.4.4 MER in HCM

From the MIR perspective, significant work has been done in the areas of melodic *motif* based *raga* identification [159, 152], analysis of melodic [160] and rhythmic

components [161], and related corpus creation [162, 163, 164] in HCM. To the best of our knowledge, no systematic study or dataset exists on the perceived emotions in HCM. Non-availability of excerpt scores, the high cost of manual annotations of emotion and related metadata by both general listeners and experts, and inherent dissimilarities between form-fluid HCM and structured Western music (more popular in the MER field) might be possible reasons. In chapter 3 of this thesis, we attempt to bridge this gap by introducing an HCM dataset specifically targeted to solve MER tasks, and a systematic and statistical study of the *Emotion-Word Intensity-Value* (EWIV) emotion representation, based on HCM concepts, followed by an exploration of possible solutions to some popular MIR tasks using this dataset and emotion representation.

CHAPTER 3

A Comparative Data-driven Study of Intensity-based Categorical Emotion Representations for MER

With the rise in freely available, quality musical content, data-driven analysis and modeling of emotions perceived from music have found widespread applications in affective computing, ranging from music recommendation to music therapy. Many emotion representations have been used in the literature mainly arrived at by borrowing ideas from psychology and adapting them to suit the need of the application. However, the relative merits and demerits of these representations in terms of expressiveness, and broad applicability have not received much attention.

On the other hand, while some concepts from Hindustani classical music (HCM), for example, *ragas*, and *layas* (tempo) have found applications in various music information retrieval tasks like *raga* identification and automatic beat tracking, perceived emotions in HCM have been sparsely explored. In this chapter, we introduce and study the effectiveness of an intensity ratings-based, categorical emotion representation called *Emotion-Word Intensity-Value* (EWIV) represen-

tation, where the categories are inspired by the *Rasa* theory of Indian aesthetics. We explore the applicability of EWIV to popular MER of applications of dominant emotion identification, and temporal emotion pattern study. We also perform statistical out-of-sample goodness of fit tests on EWIV and popular emotion paradigms of dimensional and categorical emotion representations to measure their representativeness.

3.1 Introduction

The rapid increase in musical content in various social media and other platforms has facilitated data-driven studies of perceived emotions in music. These studies encompass a variety of applications like mood-based music recommendation [15], [16], sentiment-based music generation [17], music emotion recognition (MER) [42], [54], [60], [112]. It is evident that the description, measurement, processing, and storage of perceived-emotion opinion data form an essential part of such studies, achieved using various *emotion representations* . It controls the information content extracted from the emotion annotations given by the listeners and also defines downstream tasks like dataset formats and problem formulations, which can then be used by the affective computing community to develop new technology for specific MIR tasks. For example, the MER task can be formulated as a classification problem using categorical representation, or a regression problem using a dimensional approach. Datasets, algorithms, and performance metrics used might all differ significantly in each case. The choice of an appropriate emotion representation for a given task is a widely debated, and open research topic. Hence, the analysis and comparison of emotion representations for broad applicability in MIR is an important research question.

Among the various emotion representations, the categorical (e.g. [26]) and dimensional (e.g. [2]) approaches are the most popular, though both have notable limitations like a small number of emotion classes [27], fuzzy demarcations between emotions [29], [30]. To overcome these limitations, many researchers use variants of these representations. They either use features from multiple representation models [58], [15] or incorporate additional measurements in existing representations, e.g. rating scales for discrete emotions [5], or dynamic annota-

tions [3], [112]. Creating high-quality annotated datasets for perceived emotions for different types of music clips is a costly endeavor, the representations of perceived emotions should be carefully determined. Some notable past efforts of such datasets include Soleymani et. al. [3], which uses the Circumplex model of affect [2] for representing emotions on western-origin clips, and Eerola et al. [5], which uses both the discrete and dimensional representations of emotions on film music-clips. Many other studies use the end task as a motivation for the selection of emotion representation, e.g. Lee et.al. [15] used mood categories and arousal-valence for mood-based recommender systems, Shepstone et.al. [16] used 12 categorical components and arousal-valence focus for granularity-adapted emotion classification of audio, Parada et.al. [59] used 10 categorical components of emotion and intensity labels for MER under adverse conditions, Panda et.al. [43] and Malheiro et.al. [60] used mood tags and quadrant information to explore MER relevant feature. Unfortunately, the differences in emotion representations hinder the borrowing of information from existing benchmark datasets into new studies. In this work, we seek to arrive at a general representation of musical emotions, that maximizes the information retained from the self-report data, under a given modeling assumption.

To this end, we propose and study a dynamic (time-varying), intensity ratings-based, categorical emotion representation, inspired by HCM literature (Nava Rasa [22, 23]), called the *Emotion-Word Intensity-Value* (*EWIV*) representation (section 3.3). We demonstrate the effectiveness of *EWIV*, on existing benchmark clip sets from [3], [4], and [5], as well as a newly introduced set of clips from Hindustani Classical Music (HCM), called the *EmoRaga* clip set for perceived emotion analysis in HCM (section 3.4.1.2). Estimations of dominant emotions from self-reported emotions data obtained through crowd-sourced surveys (section 3.4.3), and analyzed using the EWIV format, match the available ground truth provided by original studies for the benchmark datasets, and expert annotation for the EmoRaGa clip-set (section 3.5.1). We also validate the quality of self-reported emotions through the typicality of a clip toward the estimated dominant emotion (section 3.5.2), as well as by measuring the inter-listener agreement through Cronbach's alpha (section 3.5.3). The EWIV representation is also used to detect clips with ambiguous perceived emotions (section 3.5.4). The above exercises validate

the quality of self-reported emotion data on a wide range of clip-sets, as well as establish the utility of EWIV representation for analyzing this data.

Next, we explore the applicability of EWIV representations estimated from self-reported emotion data towards two MER applications using *EWIV* in section 3.6 - (a) emotion classification (section 3.6.1), and (b) temporal emotion variation detection (section 3.6.2). For emotion classification, we consider the tasks of dominant and secondary emotion classification, both in the multi-class as well as multi-label settings. We find that standard LSTM-based classification models [46, 165] achieve high cross-validation accuracy for all tasks. For the temporal emotion variation detection task, we use a segment-wise EWIV representation to identify the high-probability clip segments for perceived dominant emotions. Section 3.6.2 shows a high overlap coefficient between the expert annotated segments and the segments estimated from EWIV.

While the above applications demonstrate the effectiveness of EWIV representations, we are also interested in evaluating the *representativeness* of EWIV with respect to other representations, specifically, the dimensional Circumplex representation [2]. In section 3.7, we use *goodness-of-fit* measures for statistical models as our metric for representativeness. Another problem is that most datasets available in the literature, annotate music clips with single representations only, and emotion annotations using parallel representations are unavailable. The representation of choice varies across different studies. Hence, we use a conversion formula between EWIV and the Circumplex arousal-valence representation (section 3.7.1). While this conversion can incur some loss, we observe (section 3.7.2) that the *reduced EWIV* - a variant of EWIV - is consistently the best quality representation for perceived emotions, among four competing representations, for both converted and original data.

3.2 Background

3.2.1 Emotion Representation Models in MIR

Two emotion representation models are widely used in MIR: categorical and dimensional. The categorical paradigm labels music-evoked emotions into a number

of discrete classes [27], [26, 39, 1]. The dimensional approach [39, 48, 27] identifies emotions from coordinates of dimensions like *valence*, *arousal* and *dominance*, for e.g. Russell's Circumplex model of affect [2] uses arousal and valence. Though both are used extensively in different MER tasks [27, 20], their drawbacks are also much researched [47], [166], [30], [29]. While most studies use a single model, some prefer both models simultaneously or combinations of measures from both models [15],[59],[43],[60],[61]. The motivation is either task-specific or to overcome the disadvantages of any single approach. Eerola et al. [58] was one of the early works to advocate this.

Some studies include the concept of intensity annotations along with the categorical approach of emotion representation. Eerola et al. [5] compared the intensity-based discrete emotion model with the dimensional model in their seminal work. Shepstone et al. [16] used it for computing individual valence-arousal focus. Most often, a song-level (static) single-intensity rating for each emotion word is used. Thus, an the interesting research question which might remain unanswered is: *Is the information content of discrete emotion categories with dynamic intensity ratings higher than other representations?* In the present work, a deeper study of the categorical emotions with *dynamic* intensity score representation is attempted in a systematic manner.

3.2.1.1 Model Quality Estimation

The selection of appropriate emotion representation models should be based on established statistical criteria, since it affects the performance of all downstream tasks in the MER pipeline like data prediction and explanation. The Akaike information criterion (AIC) [167] is a popular measure of the suitability of a statistical model towards a given input data, which measures the loss of information when generating the data from the statistical model. It incorporates the *goodness of fit* by using log-likelihood and a measure of model complexity given by the number of parameters. It ultimately provides an indication of *out-of-sample* prediction accuracy. Given a collection of candidate models for the data, AIC estimates the quality of each model, relative to each of the other models. Let M be a statistical model for some data D, and k_M be the number of *estimated parameters* in M.

Let $\hat{L}_M$ be the maximum *likelihood function* for M. Then the AIC value of M is calculated as:

$$AIC_M = 2k_M - 2\ln(\hat{L}_M) \tag{3.1}$$

Given a set of candidate models $M_1, ..., M_n$ for a given data D, the preferred model is the one with the minimum AIC value.

3.2.2 Perceived-Emotion Tasks in MIR

3.2.2.1 Music-Perceived Emotion Datasets

One of the most common applications of emotion representations in MIR is the creation of music datasets annotated with perceived emotions. Eerola et al. [5] collected perceived-emotion data in both discrete and dimensional representations for a set of 110 excerpts, and compared the two representations. Schubert et al. [4] used six music extracts from film music, each targeting one of six discrete emotions: Excited, Happy, Calm, Sad, Scared, and Angry. After analyzing the collected discrete emotion data, they observed the presence of a second competing emotion (apart from the target) in most of their excerpts and explained it as *near miss*, concluding that some emotions might be confused. Soleymani et al. [3] proposed the benchmark *1000 songs for emotional analysis of music* dataset annotated with *static* and time-continuous (*dynamic*) arousal-valence values. Other popular datasets include the AMG1608 dataset [31], the Greek music dataset [116] and the IADS dataset [32]. To the best of our knowledge, no such dataset exists with Hindustani classical music (HCM) excerpts and perceived emotion annotations.

3.2.2.2 Music Emotion Classification

Music emotion classification is a very popular MIR task, which requires excerpt sets annotated with quality emotion opinion data. Traditionally researchers have used approaches like k-nearest neighbor classification, support vector machines (SVM) [168], [16],[169], or random forest classifiers to classify discrete emotions [44]. Recently, Han et al. [170] used both CNNs and RNNs to create cross-modal emotion embedding framework called EmoBed to leverage the knowledge from other auxiliary modalities to improve the performance of an emotion recog-

nition system. CNNs are used for audio tagging, music classification, speech emotion classification and sound event detection [171], [172]. Xie et al. [173] used attention-based LSTM-RNNs for speech emotion classification, achieving an accuracy of almost 90% in some cases. In most cases, one emotion class represents the entire excerpt under consideration, also termed *static* emotion classification. Music emotion variation detection (MEVD) focuses on the *dynamic* process of music emotion, and studies emotion variation over each predefined time segment of an excerpt [27], [133]. Most of the present literature on MEVD is based on the dimensional arousal-valence emotion annotations [20].

3.2.3 HCM in MIR

Hindustani classical music is one of the two main branches of traditional classical music in India. It is primarily based on the *raga* framework [151]. Each *raga* is characterized by a set of notes, the ascending-descending melodic progression and a specific set of melodic phrases termed *raga motifs* [152] . Other primary aspects of HCM include the *tala* (rhythmic cycle) and the *laya* (tempo). From the MIR perspective, significant work has been done in the areas of melodic *motif* based *raga* identification [159, 152], analysis of melodic [160] and rhythmic components [161], and related corpus creation [162, 163, 164] in HCM. To the best of our knowledge, no systematic study or dataset exists on the perceived emotions in HCM. Non-availability of excerpt scores, the high cost of manual annotations of emotion and related metadata by both general listeners and experts, and inherent dissimilarities between form-fluid HCM and structured Western music (more popular in the MER field) might be possible reasons.

In this chapter we attempt to bridge this gap by a) introducing an HCM dataset specifically targeted to solve MER tasks, b) a systematic and statistical study of the *Emotion-Word Intensity-Value* (EWIV) emotion representation, based on HCM concepts, c) exploring possible solutions to some popular MIR tasks using this dataset and emotion representation.

3.3 Emotion-Word and Intensity-Value Representation

Emotion-words are terms that help us understand, describe and label our emotion-opinions. The *intensity* refers to the extent to which an emotion is perceived by a listener unambiguously, while listening to a piece of music. Hence the name *Emotion-Word and Intensity-Value* (*EWIV*) is coined.

3.3.1 Overview of EWIV Representation

Two components are required to interpret perceived emotions using EWIV: 1) the emotion words, and 2) the corresponding intensities. Throughout the duration of a music excerpt, a listener is expected to *continuously* report perceived emotion-intensity opinions. Statistical analysis of the reported opinions leads to an appropriate emotion representation of the excerpts. In the present study, the choice of emotion-words used in EWIV is inspired by the concept of the *Rāsā* theory [23], a major part of Indian aesthetics [155, 22], which describes the nine primary aesthetic flavours and/or the emotions evoked by any visual, literary or musical art-form (*Nāvā Rāsā*). Seven emotion words are taken from the above list: Fear (F), Anger (A), Sadness (S), Calmness (C), Wonder (W), Romance (R) and Happiness (H). These seven are chosen as they were found to be most frequently perceived by HCM listeners in our studies. Excitement (E) is included as a descriptor of energy. Inspired from the *energetic/lively* emotion term of the Geneva Emotional Music (GEMs) scale [40], Excitement provides an indication of the high arousal perceived from the music clips. Thus, if arousal is high, higher perception of Excitement can be expected. It might be noted that high excitement can be associated with both positive and negative valence. In our experiments, high perceived excitement is associated with higher perceptions of happiness, anger and fear. In order to make the EWIV robust, two more opinion options are included: *Don't Know (DK)* - ambiguity in emotion perception, and *Other Emotions (OE)* - the incompleteness of the set of emotion words. To represent *intensity*, we use the range of $[0, 5]$. Any emotion not perceived by a listener has zero intensity value by default, at any given time during the music excerpt. The maximum in-

tensity that can be perceived and reported is 5. EWIV representation does not normalize the intensities across emotion words. This is because it is possible for a listener (λ) to not express any opinion at a given time, at which point all intensities will be zero. The third inherent component of the EWIV representation is the *timestamp* (t) of any expressed emotion opinion (ε,I). Formally, the tuple (t, ε, I) is defined as an *instantaneous report* of perceived-emotion opinion. If $\mathcal{E}$ is the set of chosen emotion words, then $\mathcal{E} = \{DK, OE, F, A, S, C, W, R, H, E\}$ and $|\mathcal{E}| = 10$. Hence, we interpret each *instantaneous report* as an $|\mathcal{E}|$-dimensional *intensity vector*. From a collection of such *intensity vector*s, the EWIV *probability vector* ($\boldsymbol{pEWIV}$) can be derived, which is the final EWIV representation of perceived music-emotion. The probability vector indicates the probabilities of perceiving the associated emotions during an excerpt. This forms the basis of the EWIV representation for perceived music emotion.

3.3.2 Emotion Estimation

In this section, we discuss the procedure to derive the EWIV *probability vector* ($\boldsymbol{pEWIV}$) from the captured *instantaneous reports* (t, ε, I) of opinion. The following three granularities of music-perceived emotions are considered:

a) **Per listener-Per excerpt**: Quantifies an individual listener's (λ) perceived-emotion opinion over a music excerpt (c). The *intensity vector* is represented by $\boldsymbol{EWIV}^{\lambda,c}$.

b) **Per excerpt**: Estimates the perceived-emotion over an entire excerpt (c) from a set of listener's (Λ) opinions, with normalization across emotions. The *intensity* and *probability vectors* are denoted as $\boldsymbol{EWIV}^{c}$ and $\boldsymbol{pEWIV}^{c}$ respectively. It measures *static* emotion in each excerpt. Both *Per listener-Per excerpt* and *Per excerpt* measures are non-temporal.

c) **Per segment-Per excerpt**: The span of a music excerpt (c) can be divided into predefined temporal *segments* (s). Perceived emotion is estimated for each segment using the same procedure as the *Per excerpt* measure, utilizing the timestamp (t) information. The *probability vector* for each segment is denoted as $\boldsymbol{pEWIV}^{c,s}$. It measures *dynamic* emotion in each excerpt.

Let the number of *instantaneous reports* be $N^{c,\lambda}$ over the span of an excerpt (c) for

a particular listener (λ). Each report (t, ε, I) can mathematically be interpreted as an $|\mathcal{E}|$-dimensional *instantaneous intensity vector* ($\boldsymbol{IIV^{c,\lambda}(n)}$, $n \in \{1, \ldots, N^{c,\lambda}\}$) at the time ($t$), such that all intensities have zero values, except the ε^{th} intensity, which has value I. The *cumulative intensity vector* ($\boldsymbol{CIV^{c,\lambda}}$) is calculated from all the $\boldsymbol{IIV^{c,\lambda}}$s so that each element ($\boldsymbol{CIV^{c,\lambda}_{\varepsilon}}$) is the summation of all intensities of the associated emotion (ε), independent of the other emotions (eq. 3.2). Each intensity in the *per listener-per excerpt* ($\boldsymbol{EWIV^{c,\lambda}}$) measure is estimated by normalizing $\boldsymbol{CIV^{c,\lambda}}$ with respect to $N^{c,\lambda}$ (eq. 3.3). Each intensity in the *per excerpt* ($\boldsymbol{EWIV^{c}}$) measure is estimated by aggregating over the set of listeners Λ (eq. 3.4).

$$\boldsymbol{CIV}^{c,\lambda}_{\varepsilon} = \sum_{n=1}^{N^{c,\lambda}} \boldsymbol{IIV}^{c,\lambda}_{\varepsilon}(n) \tag{3.2}$$

$$\boldsymbol{EWIV}^{c,\lambda}_{\varepsilon} = \frac{\boldsymbol{CIV}^{c,\lambda}_{\varepsilon}}{N^{c,\lambda}} \tag{3.3}$$

$$\boldsymbol{EWIV}^{c}_{\varepsilon} = \sum_{\lambda \in \Lambda} \boldsymbol{EWIV}^{c,\lambda}_{\varepsilon} \tag{3.4}$$

Finally, each probability in the *probability vector* ($\boldsymbol{pEWIV^{c}}$) is calculated (eq. 3.5).

$$\boldsymbol{pEWIV}^{c}_{\varepsilon} = \frac{\boldsymbol{EWIV}^{c}_{\varepsilon}}{\sum_{\varepsilon \in \mathcal{E}} \boldsymbol{EWIV}^{c}_{\varepsilon}} \times 100\% \tag{3.5}$$

The *dominant emotion* (Dom_{ε}) of a music excerpt c is defined as the emotion with the highest probability value in $\boldsymbol{pEWIV^{c}}$. The concepts of *secondary* (Sec_{ε}) and *tertiary* (Ter_{ε}) emotions are similarly defined. It is postulated that the *dominant* emotion will always be perceived from excerpt c under changing physical, mental, and contextual conditions. The same procedure as above is followed to estimate the *per-segment per-excerpt* measure ($\boldsymbol{pEWIV^{c,s}}$), with the additional temporal constraint.

3.4 Data Collection using EWIV

To test the effectiveness of EWIV representation, we collect emotion-opinion data over excerpts of two preexisting datasets and a novel HCM excerpt set. The details of the survey procedure are discussed in this section.

3.4.1 Stimuli

3.4.1.1 Excerpts from Preexisting Datasets

Schubert_6: The six excerpt excerpt-set used by Schubert et al. [4] (section 3.2.2.1). The original discrete emotion annotations are considered ground truth in the current study. *Soleymani_5*: We select five excerpts from the *1000 songs for emotional analysis of music* dataset [3] (section 3.2.2.1). The static arousal-valence annotations are mapped to emotion words [2] and are considered ground truth. We named both these excerpt sets for ease of discussion. Table 3.1 gives the metadata for the excerpts taken from previous studies - *Schubert_6* [4] and *Soleymani_5* [3]. *Excerpt#*, *Origin* and *Dur (sec)* present the excerpt number, the actual song from which the excerpt is taken, and the length of the excerpt. Column *Excerpt Emotion* report the perceived-emotion ground truth. These first four columns present details from the original datasets. The last two columns *#SelfReps* and *#Listeners* indicate the number of self-reports on each excerpt and the number of listeners who took the EWIV surveys for each excerpt. These are metadata from the surveys that we conducted using these excerpts. The corresponding *EWIV probability vectors* of perceived emotion for each excerpt is reported in table 3.2.

3.4.1.2 EmoRaga: An HCM Excerpt-set for MER

The general guidelines for the design of research corpora for computational music studies [164], [162] are followed to introduce the *EmoRaga* excerpt-set for perceived emotion analysis of HCM. For the present study, the excerpt-set comprises 48 HCM audio excerpts, its associated editorial metadata, scores, contextual information on music concepts, and perceived-emotion opinion data. The excerpts and all associated data are identified and substantiated by our HCM experts panel, which consists of five university faculty members and students, who are trained

Table 3.1: Metadata of excerpts used from two pre-exiting datasets - **Schubert_6**[4] and **Soleymani_5**[3]. Excerpt# = Original study's excerpt number, Origin = excerpt origin, Excerpt Emotion = Ground Truth. Dur (sec) = Excerpt duration in seconds. #Self Reps = No. of self-reports. #Listeners = No. of listeners who took the survey

Original Dataset Information				EWIV Survey Information	
Excerpt#	**Origin**	**Excerpt Emotion**	**Dur (sec)**	**#SelfReps**	**#Listeners**
		Schubert_6 [4]			
1	Toy Story: Infinity and Beyond	Excitement	16	402	105
2	Cars: McQueen and Sally	Happiness	16	401	124
3	Finding Nemo: Wow	Calmness	16	296	100
4	Toy Story 3: You Got Lucky	Sadness	21	446	112
5	Cars: McQueen's Lost	Fear	11	294	102
6	Up: 52 Chachki Pickup	Angrer	17	279	108
		Soleymani_5 [3]			
128	Veloma	Sadness / Bored	45	148	59
178	Uno Is Walking	Sadness / Droopy	45	201	78
171	That Kid in Fourth Grade..	Calmness/ Sleepy	45	162	73
191	Variatio 14 a 2 Clav	Happiness / Delighted	45	274	75
294	Chinese Blues ..	Happiness / Delighted	45	342	77

HCM practitioners and musicologists. An overview of the contents of this dataset is reported in table 3.3. To ensure uniformity among the excerpts, the following criteria are maintained: a) The excerpts are each of duration of 30-60 seconds, depending on the natural musical phrasing, b) All excerpts are stereo recordings sampled at 44.1 kHz. c) To avoid possible instrument-based bias, only Sitar (HCM instrument) excerpts are used. To avoid possible pitch and voice-related bias, non-lyrical vocal excerpts of only one accomplished HCM vocalist are used. The excerpts are either sourced from commercially available music releases or are generated by our HCM expert panel. The excerpts belong to 23 different *raga*s [174, 175], four prominent *tala*s, and slow and fast *laya*s. The editorial metadata associated with each excerpt consists of the source of the excerpt, the artists, the musical instruments, and the duration. We used the standardized notation for HCM [176, 177] to annotate each excerpt with relevant scores manually. The contextual information on music concepts includes the *raga*, *tala*, *laya*, pitch, and *emotion motif*s.

Emotion Motifs in HCM: Inspired by the concept of *raga motifs* [152] used for *raga* identification, we define an *emotion motif* in HCM as any key musical phrase or feature that provides strong cues to listeners to perceive particular musical emotions. The *emotion motifs* include, but may not be limited to a) Presence of major

Table 3.2: EWIV survey results over two pre-exiting excerpt-sets - **Schubert_6**[4] and **Soleymani_5**[3], and the **EmoRaga** excerpt-set introduced in section 3.4.1.2. # = Original study's excerpt number, Excerpt Emotion = Ground Truth. Near Miss = Near miss emotion reported in Schubert et al [4]. #Self Reps = No. of self-reports in EWIV surveys. {OE%...E%} = EWIV *probability vector*, where OE=Other Emotions, DK=Dont Know, F=Fear, A=Anger, S=Sadness, C=Calmness, W=Wonder, R=Romance, H=Happiness, E=Excitement. The *dominant, secondary* and *tertiary* emotions are highlighted in blue, gray, and light gray respectively. α = Cronbach's Alpha, τ = Typicality.

#	Excerpt Emotion	Near Miss	#Self Reps	OE%	DK%	F%	A%	S%	C%	W%	R%	H%	E%	α	τ
Schubert_6 [4]															
1	Exc	Hap	402	0.00	0.10	20.92	11.60	6.46	9.01	0.00	0.00	23.03	**28.90**	0.77	0.72
2	Hap	Calm	401	0.00	0.18	5.33	6.21	10.49	25.03	0.00	5.62	**29.69**	17.45	0.79	0.75
3	Calm	Hap	296	0.03	0.37	3.57	3.04	8.54	**39.94**	11.79	10.63	19.53	2.56	0.85	0.95
4	Sad	Calm	446	1.08	3.00	3.39	3.08	**29.88**	25.52	18.53	12.36	3.16	0.00	0.78	0.79
5	Fear	Ang/Exc	294	0.00	0.40	**30.74**	24.63	9.59	0.00	0.00	0.00	10.29	24.35	0.81	0.80
6	Angry	Exc/Fear	279	0.13	0.37	24.37	**31.39**	6.37	3.38	0.00	0.00	6.54	27.45	0.83	0.81
Soleymani_5 [3]															
128	Sad	-	148	2.63	4.09	10.80	1.66	**43.35**	9.99	6.95	9.47	5.16	5.90	0.93	1.12
178	Sad	-	201	3.17	3.15	3.84	1.56	**27.41**	21.54	4.89	22.87	7.90	3.67	0.74	0.77
171	Calm	-	162	0.67	2.43	2.23	0.55	14.40	**36.93**	6.29	20.00	11.17	5.33	0.82	0.83
191	Hap	-	274	1.54	4.00	1.11	5.04	2.44	6.02	15.98	8.78	**28.35**	26.74	0.79	0.78
294	Hap	-	342	1.08	2.53	0.29	1.66	0.56	1.27	7.03	9.38	**42.97**	33.23	0.92	1.06
EmoRaga (section 3.4.1.2)															
1	Hap	-	220	1.58	3.34	0.54	1.09	0.26	14.33	3.42	6.72	**42.11**	26.57	0.96	1.22
2	Hap	-	431	0.29	0.21	0.98	1.65	1.13	6.79	8.56	4.17	**41.55**	34.63	0.98	1.27
3	Sad	-	243	2.26	1.91	3.94	0.98	**54.03**	20.03	5.37	4.42	4.12	2.89	0.98	1.59
4	Sad	-	366	1.30	0.94	7.59	3.85	**56.98**	17.61	4.06	2.03	2.26	3.33	0.99	1.70
5	Calm	-	396	2.09	1.18	0.77	1.82	11.78	**34.93**	5.89	15.57	16.87	9.08	0.93	0.79

or minor notes, b) Faster or slower tempo, c) *Raga* related significant multi-note structures or phrases called *mukhyang*s/*pakad*s [178]/*raga motif*s [152], in exact or broken forms, d) Presence of *raga*-dependent *Vadi* and *Samvadi* notes [179], e) Particular rhythmic cycle (*tala*), f) Presence of particular instruments. In the present work, these *emotion motif*s and their timestamps of occurrence in each excerpt are annotated by the HCM expert panel manually. Discovery of *emotion motifs* should pave the way for efficient and explainable MIR.

Table 3.4 presents the select metadata of the *EmoRaga* excerpts. Each excerpt is assigned an identifying number (*Excerpt#*). The *Raga* (melodic structure), *Tala* (rhythmic cycle), *Laya* (tempo) and duration in seconds are recorded. For those excerpts which are free-form and follow no particular rhythmic cycle, the *Tala* column is left blank. Column *Excerpt Emotion* gives the emotion ground truth annotated by the HCM experts panel. The number of listeners who took the EWIV

Table 3.3: EmoRaga Dataset Content Summary

Topic	Content
Music Genre	HCM
# Excerpts	48
# Listener	500
Emotion Representation	EWIV
Emotion words	$\{F, A, S, C, W, R, H, E\}$
Ambiguity Indicators	OE, DK
Intensity scale	0-5
Excerpt duration	30-60sec
Excerpt selection	Manual
Self report	Perceived emotions
Method of self-report	Dynamic
Annotated by	Experts, General
Emotion Motifs	Annotated
Other metadata	Raga, Tala, Laya, Vocal/Instrumental, Supporting instruments, Pitch

survey for each excerpt, the number of self-reports on each excerpt, and the number of listeners who reported the ground-truth emotion at least once are recorded in columns *#Listeners*, *#SelfReps* and *#Listeners_on_Target* respectively. The *Cronbach's alpha* (α) and *typicality* (τ) measures are calculated from the listener annotation data for each excerpt(sections 3.5.2, 3.5.3) and reported in the last two columns. Table 3.5 presents the EWIV *probability vector* ($pEWIV^c$) for each excerpt in the *EmoRaga* dataset. Columns *Dominant Emotion* and *Secondary Emotion* report the dominant and secondary emotions observed from the EWIV estimates. These are also highlighted in blue and gray respectively, in each *probability vector*. The column *Ground Truth* gives the expert annotated emotion for each excerpt.

3.4.2 Listener-Participants

A total of five hundred general participants took part in the music perceived-emotion annotation surveys. The majority of these participants are students belonging to different courses of the university. Some faculty, staff, and their family

3.4 Data Collection using EWIV

Table 3.4: Detailed metadata of **EmoRaga** excerpts used in the present studies. Excerpt# = Excerpt number, Origin = Excerpt Raga, Tala = Rhythmic Cycle, Laya = Tempo, Excerpt Emotion = Ground Truth as annotated by HCM experts. Dur (sec) = Excerpt duration in seconds. #Self Reps = No. of self-reports. #Listeners = No. of listeners who took the survey. #Listeners_on_Target = No. of listeners who reported the excerpt emotion (ground truth) at least once. α = Cronbach's coefficient. τ = Typicality measure.

Excerpts #	**Origin Raga**	**Tala (Rhythm)**	**Laya (Tempo)**	**Excerpt Emotion**	**Dur (sec)**	**#Self Reps**	**#Listeners**	**#Listeners on_Target**	α	τ
1	Desh	Tin Taal	Fast	Happiness	30	220	82	69	0.96	1.22
2	Hamsadhwani	Tin Taal	Fast	Happiness	40	431	112	97	0.98	1.27
3	Komal Rishabh Asavari	Tin Taal	Slow	Sadness	56	243	93	81	0.98	1.59
4	Marwa	Tin Taal	Slow	Sadness	47	366	109	81	0.99	1.70
5	Khamaj	Dadara	Slow	Calmness	52	396	74	43	0.93	0.79
6	Bhoop	Tin Taal	Medium	Calmness	42	149	53	18	0.87	0.92
7	Adana	-	Slow	Exc(Anger)	30	67	53	13	0.61	0.48
8	Bhoop	-	Slow	Calmness	35	89	45	21	0.84	0.78
9	Sohani	Tin Taal	Medium	Calmness	42	121	45	13	0.39	0.34
10	Adana	Tin Taal	Medium	Calmness	42	137	45	12	0.40	0.35
11	Darbari Kanada	Tin Taal	Slow	Sadness	36	164	55	16	0.76	0.52
12	Darbari	Ektal	Slow	Calmness	45	317	76	33	0.83	0.42
13	Komal Dhaivat Bibhas	-	Slow	Sadness	35	111	45	19	0.84	0.67
14	Desh	Tin Taal	Fast	Happiness	52	318	97	56	0.95	0.93
15	Marwa	Tin Taal	Slow	Sadness	40	192	81	43	0.93	1.07
16	Bhairavi	Tin Taal	Slow	Sadness	48	297	90	48	0.96	1.01
17	Marwa	Tin Taal	Slow	Sadness	58	342	101	63	0.96	1.17
18	Puriya	-	Slow	Sadness	30	66	45	14	0.87	0.95
19	Sohani	-	Slow	Sadness	30	75	45	15	0.78	0.95
20	Darbari Kanada	-	Slow	Sadness	32	88	47	21	0.91	1.24
21	Deskar	-	Slow	Calmness	25	57	51	10	0.62	0.46
22	Puriya	Tin Taal	Fast	Fear	40	101	48	18	0.77	0.58
23	Deskar	Tin Taal	Fast	Happiness	42	158	46	14	0.50	0.31
24	Darbari Kanada	Chautaal	Slow	Sadness	52	270	79	44	0.95	0.87
25	Khamaj	Tin Taal	Slow	Happiness	55	360	74	39	0.92	0.57
26	Malkauns	Chautaal	Slow	Calmness	60	373	76	41	0.93	0.71
27	Malkauns	Ektaal	Slow	Calmness	40	221	60	26	0.86	0.59
28	Mand	Dadara	Slow	Calmness	60	235	62	25	0.83	0.44
29	Mand	Teental	Fast	Happiness	52	398	69	33	0.85	0.49
30	Bhoop	-	Slow	Calmness	27	125	45	21	0.78	0.63
31	Shivaranjani	-	Slow	Sadness	30	127	46	18	0.65	0.51
32	Hansadhwani	-	Slow	Calmness	35	148	47	15	0.54	0.49
33	Gunakri	-	Slow	Sadness	30	130	47	21	0.80	0.53
34	Bairagi	-	Slow	Sadness	30	117	45	17	0.76	0.65
35	Bhoop	Tin Taal	Fast	Happiness	45	157	53	28	0.88	0.73
36	Shivaranjani	Tin Taal	Medium	Sadness	43	134	51	19	0.85	0.65
37	Hansadhwani	Tin Taal	Fast	Happiness	48	126	50	19	0.76	0.33
38	Komal Dhaivat Bibhas	Tin Taal	Medium	Anger	55	164	55	20	0.75	0.51
39	Bairagi	Tin Taal	Slow	Sadness	45	159	54	18	0.59	0.39
40	Gunakri	Tin Taal	Slow	Sadness	46	129	54	15	0.47	0.32
41	Hansadhwani	Tin Taal	Slow	Calmness	25	228	76	43	0.91	0.66
42	Alhaiya Bilawal	Tin Taal	Slow	Sadness	25	222	74	26	0.82	0.47
43	Bibhas	Tin Taal	Slow	Sadness	25	177	72	34	0.91	0.95
44	Bhairav	Tin Taal	Slow	Sadness	25	185	80	38	0.91	0.89
45	Malkauns	Tin Taal	Slow	Sadness	26	280	78	40	0.88	0.74
46	Bhairavi	Tin Taal	Slow	Sadness	25	230	75	27	0.78	0.52
47	Durga	Tin Taal	Fast	Happiness	25	157	70	17	0.66	0.27
48	Shuddha Nat	Tin Taal	Medium	Calmness	25	341	80	33	0.79	0.39

members also volunteered for the surveys. 69.95% of the participants identified as male (μ_{age}=20.21, σ_{age}=4.89, $range_{age}$=[13,56]). 30.05% identified as female

Table 3.5: EWIV survey results for **EmoRaga** excerpts Excerpt# = Excerpt number, {OE%...E%} = EWIV *probability vector*. The *dominant* and *secondary* emotions are highlighted in blue and gray respectively. Ground Truth = Excerpt emotion as annotated by HCM experts.

Excerpt #	OE%	DK%	F%	A%	S%	C%	W%	R%	H%	E%	Dominant Emotion	Secondary Emotion	Ground Truth
1	1.58	3.34	0.54	1.09	0.26	14.33	3.42	6.72	**42.11**	26.57	Happiness	Excitement	Happiness
2	0.29	0.21	0.98	1.65	1.13	6.79	8.56	4.17	**41.55**	34.63	Happiness	Excitement	Happiness
3	2.26	1.91	3.94	0.98	**54.03**	20.03	5.37	4.42	4.12	2.89	Sadness	Calmness	Sadness
4	1.30	0.94	7.59	3.85	**56.98**	17.61	4.06	2.03	2.26	3.33	Sadness	Calmness	Sadness
5	2.09	1.18	0.77	1.82	11.78	**34.93**	5.89	15.57	16.87	9.08	Calmness	Happiness	Calmness
6	2.66	3.89	0.53	3.12	6.33	**38.54**	5.08	15.27	18.54	6.01	Calmness	Happiness	Calmness
7	4.57	5.48	0.91	15.70	19.05	2.13	9.60	4.57	13.95	**24.01**	Excited	Sadness	Anger
8	1.98	3.92	3.93	2.82	18.19	**33.06**	13.20	14.30	10.05	2.46	Calmness	Sadness	Calmness
9	1.32	5.29	6.44	13.02	14.16	**20.74**	8.82	7.03	15.49	7.68	Calmness	Happiness	Calmness
10	8.98	7.57	3.26	6.40	10.67	**20.15**	9.67	11.08	3.03	19.20	Calmness	Excitement	Calmness
11	1.86	4.87	4.63	3.73	**25.05**	24.08	12.46	5.78	5.98	11.54	Sadness	Calmness	Sadness
12	4.94	3.57	3.77	4.08	16.01	**22.69**	8.47	5.90	15.15	15.40	Calmness	Sadness	Calmness
13	2.62	1.95	12.08	2.49	**30.00**	23.82	14.36	1.60	9.69	1.38	Sadness	Calmness	Sadness
14	3.75	2.71	0.41	1.63	1.55	15.39	5.96	15.39	**36.45**	16.70	Happiness	Excitement	Happiness
15	2.45	4.75	12.15	3.53	**41.72**	13.63	10.60	3.50	4.72	2.88	Sadness	Calmness	Sadness
16	1.55	5.23	2.80	2.99	**40.74**	29.07	8.04	4.25	4.56	0.73	Sadness	Calmness	Sadness
17	4.74	2.83	10.79	1.24	**43.32**	23.07	6.10	3.95	2.42	1.51	Sadness	Calmness	Sadness
18	2.32	1.67	3.75	2.09	**38.16**	30.75	7.57	5.35	4.64	3.71	Sadness	Calmness	Sadness
19	0.62	4.67	6.54	14.01	**35.21**	14.56	11.38	6.23	3.24	3.53	Sadness	Calm/Anger	Sadness
20	1.28	4.66	7.92	5.28	**49.30**	13.09	0.76	1.70	7.80	8.18	Sadness	Calmness	Sadness
21	2.5	14.17	1.19	0.00	16.19	**23.41**	9.05	14.29	17.14	2.06	Calmness	Happiness	Calmness
22	0.29	7.84	18.93	7.18	**29.72**	15.89	6.67	6.88	4.18	2.39	Sadness	Fear	Fear
23	8.61	4.54	1.61	8.04	3.94	13.58	11.06	12.92	**20.18**	15.50	Happiness	Excitement	Happiness
24	5.41	3.87	2.12	6.40	**34.86**	33.24	5.11	1.81	5.77	1.40	Sadness	Calmness	Sadness
25	1.26	1.99	1.82	2.85	3.15	14.01	11.75	12.77	**27.26**	23.13	Happiness	Excitement	Happiness
26	2.66	5.03	6.07	5.73	29.67	**30.73**	2.90	4.30	5.38	7.52	Calmness	Sadness	Calmness
27	4.93	1.32	4.19	5.44	12.99	**27.91**	8.45	4.13	6.32	24.33	Calmness	Excitement	Calmness
28	3.53	3.89	1.41	0.58	9.35	**22.98**	7.46	13.85	21.87	15.07	Calmness	Happiness	Calmness
29	4.01	4.18	0.73	2.67	7.77	16.67	7.04	20.70	**25.43**	10.77	Happiness	Romance	Happiness
30	3.84	2.15	4.17	6.68	12.72	**28.73**	18.44	7.39	9.50	6.36	Calmness	Wonderment	Calmness
31	6.62	3.73	11.26	5.09	**26.06**	15.79	8.35	12.37	4.79	5.93	Sadness	Calmness	Sadness
32	6.55	6.39	6.87	10.57	11.42	**23.95**	2.12	9.47	14.06	8.57	Calmness	Happiness	Calmness
33	4.27	2.78	16.99	14.69	**24.72**	20.98	8.91	3.77	1.64	1.24	Sadness	Calmness	Sadness
34	5.58	13.22	10.29	4.53	**28.88**	18.20	6.07	1.96	3.73	7.53	Sadness	Calmness	Sadness
35	2.94	3.14	1.88	4.26	5.70	18.94	6.01	14.47	**32.73**	9.90	Happiness	Calmness	Happiness
36	0.94	2.72	6.79	7.25	**31.16**	25.53	9.00	6.35	4.04	6.22	Sadness	Calmness	Sadness
37	1.11	9.93	0.22	3.69	5.59	16.17	11.73	19.58	**20.58**	11.39	Happiness	Romance	Happiness
38	3.76	4.99	4.35	**25.75**	16.60	9.44	7.09	14.66	7.45	5.88	Anger	Sadness	Anger
39	2.66	11.79	15.01	11.10	**22.02**	12.40	5.95	9.27	1.83	7.95	Sadness	Fear	Sadness
40	5.50	2.32	10.29	10.43	**20.43**	14.25	12.19	11.25	3.84	9.50	Sadness	Calmness	Sadness
41	2.68	2.01	2.78	2.96	14.76	**30.20**	11.48	17.12	10.15	5.84	Calmness	Romance	Calmness
42	3.24	1.02	6.26	3.31	**24.58**	18.44	8.14	9.16	14.48	11.36	Sadness	Calmness	Sadness
43	2.69	3.86	15.16	9.22	**37.56**	14.64	5.16	2.33	4.82	4.55	Sadness	Fear	Sadness
44	5.90	2.85	10.18	6.67	**35.19**	15.26	8.63	5.05	1.39	8.86	Sadness	Calmness	Sadness
45	3.30	8.36	15.91	8.89	**31.93**	12.11	7.00	7.95	1.82	2.72	Sadness	Fear	Sadness
46	2.00	3.46	16.51	10.35	**24.42**	9.60	12.01	6.86	8.37	6.42	Sadness	Fear	Sadness
47	2.38	3.98	3.31	8.10	14.52	17.24	15.27	9.31	**18.40**	7.47	Happiness	Calmness	Happiness
48	2.00	7.57	1.84	6.69	10.99	**21.69**	11.42	17.98	11.85	7.94	Calmness	Romance	Calmness

(μ_{age}=22.64, σ_{age}=6.99, $range_{age}$=[13,59]). All participants are Indian nationals. Participants were informed of the nature and objective of the study prior to the surveys. Participation was voluntary and participants provided online consent before accessing the online survey. Response anonymity and pure academic use of

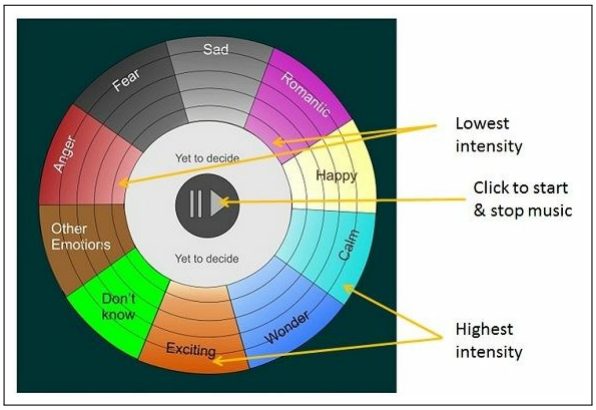


Figure 3.1: Web-based Perceived-Emotion Collection Interface

collected data were guaranteed.

3.4.3 Survey Procedure

The interface presented in figure 3.1 is used to collect opinion responses (t, ε, I) from listeners. Surveys started with an instruction page containing a short description of how to use the interface to report continuous-time perceived emotions during a music excerpt. It also explains the meaning of *perceived* emotion versus *felt* emotion and asked the participants to report "emotions that you perceive or recognize from the music while listening to it and not that which you yourself feel". Each participant was directed to listen to the music and simultaneously respond with the perceived emotion and intensity in the wheel as desired, and as many times as they felt necessary. Through the interface-page each excerpt was presented to the listeners for annotation in isolation, with a time gap of 60 seconds between excerpts. Each round of the survey spanned 20-25 minutes and 10 excerpts were presented to a listener during each round.

3.5 Analysis of Survey Data

In this section, we present the results of various analyses performed using the EWIV emotion data captured in the previous section. The results are used to study the validity and utility of EWIV representation.

3.5.1 EWIV Estimations from Collected Data

The *instantaneous reports* are collected from the surveys and two types of *probability vectors* are estimated for each excerpt: the *per-excerpt* probability vector and the *per segment - per excerpt* probability vectors (section 3.3.2). The static dataset consists of one probability vector, one dominant emotion, and one secondary emotion per excerpt. Table 3.2 presents the *per-excerpt* results for the two existing excerpt-sets, *Schubert_6* and *Soleymani_5*, along with the results for the first 5 excerpts of EmoRaga excerpt-set. The excerpt numbers from the original datasets are retained in the first column (#). Columns *Excerpt Emotion* and *#Self Reps* report the perceived-emotion ground-truth and the number of *instantaneous report*s for each excerpt. The *near miss* for each excerpt in *Schubert_6* are reported in the last column. The EWIV *per-excerpt probability vectors* are reported in columns *OE%*... *E%*. The *dominant*, *secondary* and *tertiary* emotions are highlighted in blue and shades of gray respectively. The *per-excerpt* EWIV vectors for each excerpt are compared with the individual ground truths (*Except Emotion*). For all excerpts in the three datasets (table 3.2) the *dominant* emotions in the EWIV *probability vector*s (highlighted in blue) match the ground truth in column *Excerpt Emotion*. For each excerpt in *Schubert_6* dataset, the *secondary* emotions match the *near miss* [4]. Columns α and τ present Cronbach's alpha and typicality measures respectively, derived from analysis of the *per-excerpt* data, discussed in sections 3.5.2 to 3.5.4.

To estimate the *per segment-per excerpt* emotion probability vectors, uniform, non-overlapping, consecutive *segments* of 1 second duration are considered. Segments with no emotion-word annotations are excluded from the present study. The dynamic dataset thus created consists of 1700 *segments* of music, each associated with a probability vector, one dominant and one secondary perceived-emotion label. This dataset is used for various MER tasks described in section 3.6.

3.5.2 EWIV and Typicality

The *typicality* (τ) of an excerpt to a particular emotion [5] is described as the property by which that emotion is more easily perceived in that excerpt than other emotions. It is estimated as $\tau = \overline{E} - SE - \overline{NE}$ [5], where, $\overline{E}$ and SE are

the mean and standard deviation of the dominant emotion ratings, and $\overline{NE}$ is the mean of non-dominant emotion ratings of an excerpt. The *typicality* values of each excerpt of *Schubert_6*, *Soleymani_5* and the first 5 excerpts of *EmoRaga* excerpt-set to their individual *Excerpt Emotion* are reported in column (τ), table 3.2. It is observed that *typicality* is well reflected in the probability values of the *dominant* emotions, captured by EWIV. The higher the probability of the *dominant* emotion, the higher is the *typicality* of the excerpt for that emotion.

3.5.3 Listener Consensus in EWIV

Cronbach's alpha (α) [5] is used to measure the agreement between the participants about their perceived-emotion opinions for each excerpt. This provides an estimate of the internal consistency and reliability of the reported opinions. The results are reported in table 3.2, column (α). It is observed that most excerpts have a high α value, demonstrating the high quality of reported emotion opinions. Further, it is noted that for highly *typical* excerpts of any *dominant* emotion, the α is also high ($0.9 \leq \alpha$). This is intuitive since a greater number of participants agreeing to a particular emotion in an excerpt lends it to be typical of that emotion. But, low *typicality* does not necessarily mean low consensus. For eg, in the *EmoRaga* excerpt-set, excerpt#8 (α=0.84, τ=0.78) and excerpt#12 (α=0.83, τ = 0.42) have same *dominant* emotion, Calmness. While they both have high consensus (α), excerpt#12 has much lower typicality. This might be explained from the respective *probability vectors*. Excerpt#8 has a markedly *dominant* emotion (Calmness) denoted by a high probability. The probabilities of all the other emotions, including the *secondary* emotion, are notably less. Whereas, in excerpt#12, the probabilities of a number of emotions (Sadness, Happiness, Excitement) are competing with the *dominant* emotion. In this case, the participants highly concede that the excerpt is atypical of any one emotion.

3.5.4 Identifying Ambiguity in Music Excerpts

Two types of ambiguity are identified in the excerpts using α and τ values. *Type 1:* High α, Low τ: e.g. excerpt# 11, 12, 13 of *EmoRaga*. The following are observed from the *probability vectors*: a) The *dominant* emotion might be ambiguous, due

to the presence of at least one other highly perceivable emotion. b) Probabilities of the ambiguity indicators *OE* and *DK* are low (≤ 5). The ambiguity arises from more than one highly perceivable emotion by most listeners. *Type 2:* Low α, Low τ: e.g. excerpt# 7, 9, 10 of *EmoRaga*. In this case, it is observed that a) Probabilities for perceiving multiple emotions are equally low. b) Probabilities for *OE* and *DK* are high (≥ 5). The ambiguity arises as no emotion is perceived well by a large number of participants. Since emotion perception in music is subjective, identifying ambiguity might help to understand generic emotion perception in music better.

3.6 Applications

3.6.1 Dynamic Emotion Classification

In this section, the *dynamic EmoRaga* dataset (section 3.5.1) is used for two emotion classification tasks. First, in the *multi-class classification* task of dominant or secondary emotions, the aim is to classify each music *segment* into one of the 8 emotion classes {F, A, S, C, W, R, H, E} (3.3.1). The second one is joint dominant and secondary emotion labeling - a *multi-label classification* problem. Here the focus is to find the two top-most probable perceived emotions (dominant and secondary) of every *segment* and predict their probabilities of perception.

3.6.1.1 Experimental Setup

The dynamic dataset derived from the *per segment-per excerpt* probability vectors 3.5.1 is used for this task. The *spectral features* of the segments are extracted using the Librosa [126] tool. They denote the distributions of energy over a set of frequencies and have provided state-of-the-art emotion estimates previously [180]. These features consist of Chroma(24), CENs (12) MFCC (20), RMS (1), Mel-scaled spectrogram (128), spectral centroid (1), spectral bandwidth (1), spectral flatness (1), spectral roll-off (1) and zero crossing rate (1). So, the feature set size for each segment is 190. All excerpts are re-sampled to 44100 Hz before feature extraction. The standard scalar normalization is used for preprocessing the data before MIR tasks.

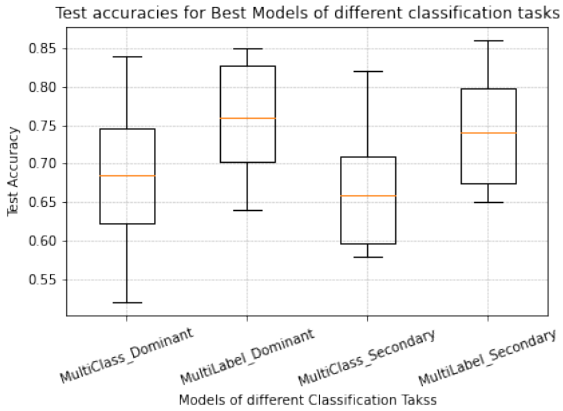


Figure 3.2: Test accuracies across K=10 fold Cross Validation for *dominant* and *secondary* emotion classification using multi-class and multi-label approaches.

The LSTM-RNN [181] is used for the classification tasks. Both *single layer* and *double layer* LSTM models with varied layer sizes are explored and the best suitable architecture is finalized. The best suitable architecture is finalized after an exploration with varying layer sizes and layer numbers. The classification task results are obtained using the best model architecture. K-fold cross-validation is used, with K=10. For the multi-class classification tasks, *softmax cross entropy with logits* function is used to calculate the loss. For the multi-label classification task, *binary cross-entropy* loss function is used. *Adam optimizer* is employed for all the tasks, with a maximum of 50 epochs. All hyper-parameters not explicitly mentioned here are left to their default values as in Tensorflow v2.7.0. The *accuracy* metric is used for presenting the results. All programs were performed in the Linux operating system using Python programming language.

3.6.1.2 Experiment 1: Multi-Class Classification

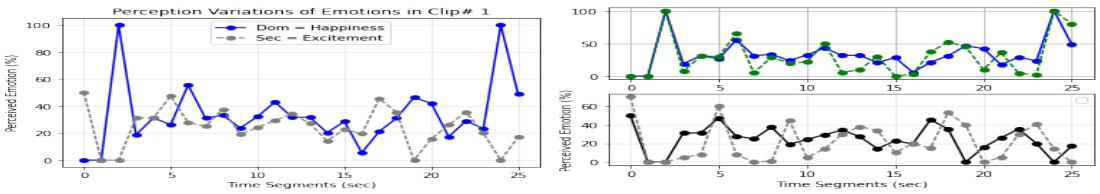
In the single-layer LSTM model, the hidden layer size is varied from 10 to 256 units. For the double-layer LSTM model, the hidden layer sizes are varied as (20,10), (40,20), (64,20), (128,64), (256,64), and (256,128) units. In all the models, the LSTM layers are followed by one *dense layer* with Relu activation and a final *dense layer* of size 8 for the 8 possible classes (emotion words). The corresponding accuracies are compared and the best model is chosen for the multi-class emotion

classification task - the single-layer LSTM model with a hidden layer size of 64 units. The test accuracies for multi-class classification of *segments* into dominant and secondary emotion classes across K(=10) folds are presented in figure 3.2.

Results: The following are observed from this experiment: a) The mean test accuracies for the dominant and secondary emotion classifiers are calculated to be 0.68 and 0.67 respectively. b) EWIV representation can be used to classify *segments* into dominant or secondary emotions with state-of-the-art test accuracies. c) The best classification performance reported in this section is comparable to similar emotion classification results reported in literature [173]. d) The single and double-layer models' performances are comparable.

3.6.1.3 Experiment 2: Multi-Label Classification

The hidden layer sizes of single and double-layer LSTM models are varied and the best model is identified. The model consists of single layer LSTM (size=128 units), followed by a dense layer (size=20 units) with Relu activation and a final dense layer (size = 8 units) with *hard sigmoid* activation. For multi-label classification, joint and individual accuracies of dominant and secondary emotions are calculated first on the raw outputs of the model. It might be noted that the target labels for this task can be considered as *multi-hot* encoded. Since the model outputs a probability value in the range (0,1) for each of the 8 classes, the threshold to consider the presence of an emotion is assumed to be 0.5. All predicted values $\geq$ 0.5 in the output are converted to 1 and all others are replaced with 0s. With these adjusted (corrected) predictions, both the joint and individual accuracies are re-calculated, which represent the actual accuracies produced using the model. *Results:* a) The mean test accuracy across the K folds for the joint prediction of dominant and secondary emotions is 0.50. b) The individual accuracies are calculated to be 0.76 and 0.74 respectively. The individual test accuracies for multi-label classification of *segments* into dominant and secondary emotion classes jointly are plotted in figure 3.2. It is observed that the mean accuracies of the adjusted (corrected) multi-label dominant and secondary emotion classification surpass the multi-class classification accuracies for both dominant and secondary emotions.



(a) Excerpt 1 - Manual TEP (b) Predicted TEP vs Manual TEP

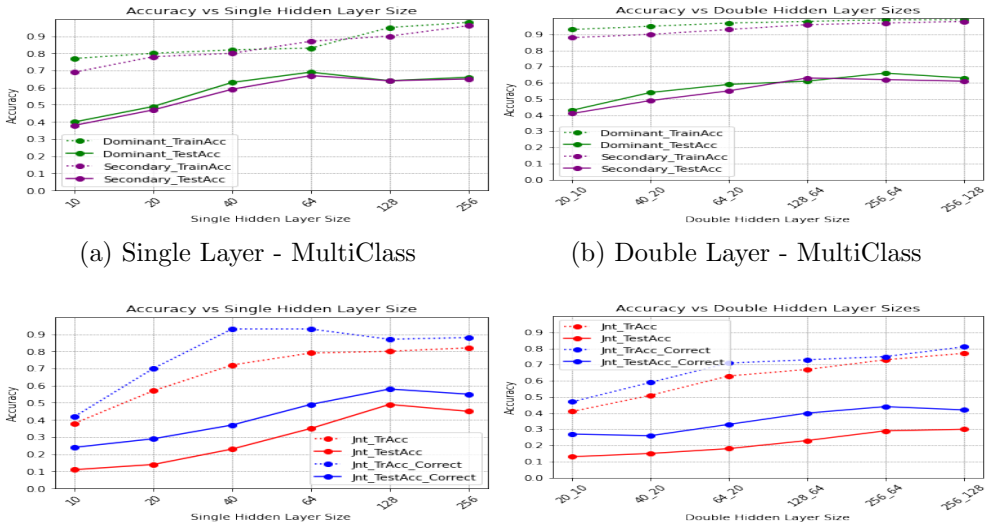
Figure 3.3: Temporal Emotion Patterns (TEP): Variations of *dominant* and *secondary* emotions over excerpt#1 in EmoRaga dataset. Figure(a) shows the ground truth: variations in manual perceived-emotion annotations recorded during the survey. Figure(b) depicts the variations in predicted emotions (dotted graph) in comparison with ground truth (unbroken graph). The top and bottom sub-figures are for dominant and secondary emotions respectively.

3.6.1.4 Illustrative Example

The best multi-label classifier model identified in the previous section is used to predict variations in dominant and secondary emotion perception in individual *EmoRaga* excerpts (MEVD). The ground truths of dominant and secondary emotion probabilities provided by the *per segment-per excerpt* vectors are then compared with the predicted probabilities. Figure 3.3(a) plots the ground truth variations of the *dominant* (Happiness) and *secondary* (Excitement) perceived-emotions in each *segment* of excerpt #1 of the *EmoRaga* dataset. It is observed that excerpt #1 is rated as dominantly happy in the first and last few seconds, although perception probability is generally high ($\simeq 40\%$). Secondary perceptions($\leq 50\%$) of excitement are reported throughout the excerpt. In figure 3.3(b), each sub-graph represents a comparison of the ground-truth perceived-emotion probabilities of *dominant* and *secondary* emotions provided by the *EmoRaga* data and the ones predicted by the multi-label classifier described in section 3.6.1.3. It is observed that the *dominant* emotion prediction fares slightly better than the *secondary* emotion prediction.

3.6.1.5 Detailed Results

To perform multi-class and multi-label classifications, experiments were conducted with various LSTM-RNN architectures and varied layer sizes in order to find the



(a) Single Layer - MultiClass

(b) Double Layer - MultiClass

(c) Single Layer - MultiLabel

(d) Double Layer - MultiLabel

Figure 3.4: Architecture Search - Classification Test Accuracy variations with respect to varying hidden layer sizes of Single and Double layer LSTM models in Multi-Class Emotion Classification Tasks - for Dominant and Secondary emotions.

best-performing model. The train and test accuracies over these architectures are depicted in figure 3.4. Figures 3.4(a) and 3.4(b) give the train and test accuracies of *dominant* and *secondary* emotion multi-class classification using single and double hidden layer models. The best-identified model is used for experiments in section 3.6.1.2. Figures 3.4(c) and 3.4(d) give the joint train and test accuracies - both raw and corrected - of dominant and secondary emotion multi-label classification using single and double hidden layer models. The best-identified model is used for experiments in section 3.6.1.3. The individual classifications of *dominant* and *secondary* emotions are performed on the temporal *segments*, using the best model architectures identified. The train and test accuracies over the 10 folds are reported in the second and third broad columns of table 3.6 - *Multi-Class(Dom)* and *Multi-Class(Sec)*. Various accuracies of the joint *dominant* and *secondary* emotion tagging task are reported in the last broad column - *Multi-Label(Dom+Sec)*. The first set of Train and Test accuracies reported are calculated on the raw outputs of the model. The *Joint* columns indicate the joint classification accuracies of dominant and secondary emotions. The *Dom* and *Sec*

Table 3.6: Details of results from k=10 folds Cross Validation for Dominant emotion classification, secondary emotion classification, and Dominant+Secondary emotion tagging problems

K	Multi-Class(Dom)		Multi-Class(Sec)		Multi-Label(Dom+Sec)											
	Train	Test	Train	Test	Train			Test			Corrected_Train			Corrected_Test		
					Joint	Dom	Sec	Joint	Dom	Sec	Joint	Dom	Sec	Joint	Dom	Sec
1	0.87	0.68	0.83	0.62	0.94	0.97	0.96	0.38	0.46	0.43	0.97	0.99	0.99	0.50	0.74	0.73
2	0.94	0.73	0.91	0.71	0.92	0.94	0.91	0.35	0.50	0.52	0.97	0.99	0.99	0.49	0.77	0.76
3	0.78	0.52	0.78	0.58	0.72	0.79	0.75	0.27	0.44	0.45	0.67	0.89	0.86	0.41	0.64	0.66
4	0.86	0.63	0.81	0.59	0.82	0.86	0.84	0.25	0.47	0.43	0.98	0.99	0.99	0.47	0.69	0.65
5	0.88	0.69	0.89	0.68	0.89	0.90	0.89	0.41	0.52	0.43	0.94	0.98	0.98	0.52	0.79	0.75
6	0.96	0.75	0.94	0.75	0.85	0.88	0.87	0.50	0.67	0.58	0.92	0.97	0.96	0.58	0.84	0.82
7	0.79	0.60	0.80	0.58	0.70	0.77	0.74	0.25	0.33	0.31	0.90	0.96	0.96	0.43	0.69	0.65
8	0.80	0.62	0.85	0.64	0.89	0.90	0.89	0.37	0.47	0.46	0.96	0.99	0.99	0.46	0.75	0.72
9	0.95	0.77	0.89	0.71	0.87	0.92	0.90	0.45	0.55	0.53	0.96	0.98	0.98	0.53	0.84	0.81
10	0.97	0.85	0.97	0.82	0.92	0.94	0.94	0.49	0.60	0.56	0.96	0.98	0.98	0.59	0.85	0.86

Table 3.7: Overlap coefficients (OVL) between a) set of segments with *emotion motif* marked by experts (GT) and set of segments with a high perceived probability of *dominant* emotions in audience response (AR) and b) GT and set of segments with a high predicted probability of *dominant* emotions in model prediction (MP), for the first 4 excerpts of the EmoRaga dataset.

#	**Segments with Expert Annotated Emotion Motifs: Ground Truth (GT)**	**Segments with High Probability of Dom_{ε}: Audience Response (AR)**	**Segments with High Probability of Dom_{ε}: Model Prediction (MP)**	**OVL between GT & AR**	**OVL between GT & MP**
1	2-6,8-10,12,13,18-22	2,4,6-8,10-13,15,18-20,22,24,25	2,4-6,8,10,11,14,17,18,19,21,23,24	0.73	0.64
2	1-4,7-10,14-18,21-27,31-35	1-6,8-11,16,19,24,25,28,29,30,31	1-3,5-8,11-13,25,30,31,34	0.61	0.57
3	1-5,11-18,32-40	2-14,18,21,23,25,26,39-41,43,46,47	2-7,11,12,18,19,25,26,40-43,48	0.64	0.59
4	2-4,6-11,14-17,20,33,37-40,44-47	1-4,6-19,21-26,28-30,33-37,39-43	2-7,10-13,15-19,22,27,33,35,38,44,46	0.71	0.60

columns indicate the classification accuracies of dominant and secondary emotions individually in the multi-label classification scenario. The second set of Train and Test accuracies termed *Corrected_Train* and *Corrected_Test* are calculated on the adjusted (corrected) outputs of the model. It might be recalled that the target labels for this task are multi-hot encoded. Since the model outputs a probability value in the range (0,1) for each of the 10 classes for each segment, we considered the threshold for presence of an emotion to be 0.50. All predicted values ≥ 0.5 were converted to 1 and all others were replaced with 0s. With these adjusted (corrected) predictions, both joint and individual accuracies are calculated.

3.6.2 Detecting Temporal Emotion Patterns and Motifs

The collected *per segment-per excerpt* data of the *dominant* emotions (Dom_{ε}) indicate the presence of some *segments* where the perception probability is sig-

nificantly high (≥30%). The *dynamic* emotion predictions also indicate *segments* with a high predicted probability of the *dominant* emotions (Dom_{ε}). These high-perception *segments* are identified and compared with those that are annotated by experts as containing *emotion motif*s (section 3.4.1) - the ground truth. The Szymkiewicz–Simpson coefficient or Overlap Coefficient (OVL) is used for this comparison, which is given by:

$$OVL = \frac{|A \cap B|}{min(|A|, |B|)} \tag{3.6}$$

where A and B are two finite sets. This is reported in table 3.7 for the first 4 excerpts of the *EmoRaga* dataset. It is observed that the overlap coefficients are high (≥0.50) in most cases, both between ground truth and EWIV collected data and ground truth and model prediction data. This indicates a possible association between these expert-annotated *emotion-motifs* and emotions perceived by listeners. Automatic recognition of such high-perception segments might assist in *emotion motif* detection in HCM excerpts, and help build explainable music emotion recognition models.

3.7 Comparison of EWIV with Arousal-Valence Representation

In this section, we aim to estimate the quality of emotion representations as statistical models, over a given set of perceived-emotion opinion data. We compare the EWIV and the dimensional Circumplex [2] models of music perceived-emotion representation, for expressibility of real-world emotions collected in relevant datasets. We consider both EWIV and the Circumplex representations as probabilistic models for fitting the emotions reported in the self-report data points, and use the Akaike Information Criterion (AIC) [167] for this comparison. The data representation formats for EWIV and Circumplex model are different, one uses the intensity-based categorical format and the other follows the dimensional format. To compare these two representations using AIC, perceived-emotion opinion data is required in both formats for the same target excerpts. Such datasets are rare

in existing literature, with the exception of the seminal work by Eerola et al [5], which provides data in both discrete and dimensional formats over the same dataset. Generally, the opinion data using these models are collected on different music excerpts in different datasets and for different participants. One way to alleviate this issue is to collect new opinion data on excerpts of existing datasets, using the absent format. For example, we have collected EWIV format opinion data on the excerpts of *Soleymani_5*, which is part of a benchmark dataset with dimensional format data 3.4.1. There are two practical constraints to this approach: a) such extensive data collection might be expensive, and b) the participants of the original dataset might not be available for new data collection. To circumvent these problems and to ensure the broader applicability of this comparison in datasets where data of only one format is available, we designed a conversion scheme between these formats. The conversion scheme could possibly be noisy, leading to information loss. Hence we perform both-way conversion and compare the information content of the resulting datasets. In this section, first, we discuss the conversion procedures (section 3.7.1), and next (section 3.7.2), we consider different model estimations with respective AIC calculations and finally report the empirical results over three datasets.

3.7.1 Conversion of Representations

3.7.1.1 EWIV to Circumplex

In the Circumplex 2-D plane [2], each emotion term is associated with an angular value, indicating its location ([2], section *Polar coordinates for the 28 words*). To convert EWIV format data to Circumplex format, we use these: Fear/Scared (100°), Angry (92°), Sad (207.5°), Calm (316.2°), Happy (7.8°), and Excited (48.6°). We assume that a listener responds N^c times during an excerpt c using the EWIV representation, and each time the *instantaneous report* tuple (t, ε, I) is recorded (section 3.3.1). Considering the angular value associated with emotion ε to be θ_ε [2], the corresponding valence v_ε and arousal a_ε values can be determined as:

$$v_\varepsilon = I_\varepsilon \cos\theta_\varepsilon \quad \text{and} \quad a_\varepsilon = I_\varepsilon \sin\theta_\varepsilon \tag{3.7}$$

Table 3.8: AIC results for the first 5 excerpts from Eerola's Dataset [5]. The terms AIC_{EWIV}, AIC_{EWIV_R} and AIC_{AV} represent AIC values for EWIV, Reduced EWIV, and AV models respectively. Columns (1)-(3) represent AICs calculated over emotion data from the original dataset. Columns (4)-(5) give the AICs calculated over converted emotion data. Column (6) presents the AIC of an integrated model, constructed using emotion data from the original dataset. The best model for each excerpt is highlighted in blue, and the second-best in gray.

Excerpt#	Emotion Category Level	Original Data			Converted Data		Integrated Data
		AIC_{EWIV}(1)	AIC_{EWIV_R}(2)	AIC_{AV}(3)	AIC_{EWIV}(4)	AIC_{AV}(5)	$\text{AIC}_{EWIV+AV}$(6)
1	Anger High	466.93	444.44	469.69	613.62	570.89	1116.87
2	Anger High	389.67	388.94	393.99	631.80	536.12	1041.98
3	Anger High	487.46	458.84	486.89	759.68	596.32	1129.84
4	Anger High	463.40	406.24	463.46	727.81	554.93	1264.48
5	Anger High	520.89	494.76	528.40	980.60	578.77	1230.19

The static *per excerpt* estimate of perceived-emotion for excerpt c in arousal-valence terms (V^c_{avg}, A^c_{avg}) can be calculated over all such *instantaneous reports* as:

$$V^c_{avg} = \frac{\sum_{\varepsilon \in \mathcal{E}} v_\varepsilon}{N^c} \quad \text{and} \quad A^c_{avg} = \frac{\sum_{\varepsilon \in \mathcal{E}} a_\varepsilon}{N^c} \tag{3.8}$$

3.7.1.2 Circumplex to EWIV

To convert arousal-valence data to EWIV format, we define a *region* associated with each emotion ε in the 2-D plane. Each *region* is limited by a minimum $(\theta^{min}_\varepsilon)$ and a maximum $(\theta^{max}_\varepsilon)$ angular value from the x-axis. Let $\theta_\varepsilon, \theta_{\varepsilon-1}$ and $\theta_{\varepsilon+1}$ be the angular values associated with emotion ε, and the emotions preceding and succeeding ε in the 2-D Circumplex plane [2]. Then, θ^{min}_ε and θ^{max}_ε are defined as:

$$\theta^{min}_\varepsilon = \theta_\varepsilon - \frac{(\theta_\varepsilon - \theta_{\varepsilon-1})}{2} \quad \text{and} \quad \theta^{max}_\varepsilon = \theta_\varepsilon + \frac{(\theta_{\varepsilon+1} - \theta_\varepsilon)}{2} \tag{3.9}$$

The *region* of emotion ε is thus demarcated by $[\theta^{min}_\varepsilon, \theta^{max}_\varepsilon]$, and can be further be sub-divided into 5 equal *sub-regions*, which we map to the five intensities of the present EWIV. Given an arousal-valence response (a_k, v_k), the angular coordinate θ_k of this point on the 2-D plane is given by:

$$\theta_k = \tan^{-1}(\frac{a_k}{v_k}) \tag{3.10}$$

If $\theta_{\varepsilon}^{min} < \theta_k \leq \theta_{\varepsilon}^{max}$, then θ_k is said to be associated with emotion ε. The intensity is also derived from θ_k, based on the sub-region it is in. It might be noted, the radial coordinate of the point (a_k, v_k) is not used since it is expected to be the unit norm.

3.7.2 Comparison of EWIV and Circumplex Model

We describe the estimation of model parameters and calculation of AIC values for the different representations in section 3.7.2.1. Next, in section 3.7.2.2, we analyze and report the empirical results on three datasets.

3.7.2.1 Model estimation and AIC calculation

For the EWIV representation, we assume that self-reported emotion values for an emotion word ε and an excerpt c follows a normal distribution, with mean $\mu_{\varepsilon,c}$ and standard deviation $\sigma_{\varepsilon,c}$. Hence, each self report of emotion denoted by $EWIV_{\varepsilon}^{c,\lambda}(n)$, where λ = listener, n = response index, (section 3.3.2), can be considered a random sample from the following distribution:

$$EWIV_{\varepsilon}^{c,\lambda} \sim \mathcal{N}(\mu_{\varepsilon,c}, \sigma_{\varepsilon,c}) \tag{3.11}$$

Considering the 8 emotion words of EWIV (section 3.3.1), we have 8 parameters for the mean ($\mu_{EWIV} = [\mu_F, \mu_A, \mu_S, \mu_C, \mu_W, \mu_R, \mu_H, \mu_E]$), and 8 for the variance ($\sigma_{EWIV} = [\sigma_F, \sigma_A, \sigma_S, \sigma_C, \sigma_W, \sigma_R, \sigma_H, \sigma_E]$). Hence, there are k=16 parameters to be estimated. Given a dataset $\mathcal{D}_{\varepsilon,c} = \{EWIV_{\varepsilon}^{c,\lambda}(n) |\ \forall\lambda, n\}$ of all self reports corresponding to emotion ε and excerpt c, the parameters $\mu_{\varepsilon,c}, \sigma_{\varepsilon,c}$ are estimated using standard Gaussian maximum likelihood estimation formulae:

$$\mu_{\varepsilon,c} = \frac{1}{|\mathcal{D}_{\varepsilon,c}|} \sum_{\lambda,n} EWIV_{\varepsilon}^{c,\lambda}(n) \tag{3.12}$$

and

$$\sigma_{\varepsilon,c}^2 = \frac{1}{|\mathcal{D}_{\varepsilon,c}|} \sum_{\lambda,n} (EWIV_{\varepsilon}^{c,\lambda}(n) - \mu_{\varepsilon,c})^2 \tag{3.13}$$

The total log-likelihood for the model M is estimated as:

$$\ln(\hat{L}_c(M)) = \sum_{\varepsilon} \ln(\mathcal{L}(\mathcal{D}_{\varepsilon,c}|\mu_{\varepsilon,c}, \sigma_{\varepsilon,c})) \tag{3.14}$$

We also consider a *reduced* EWIV model ($EWIV_R$), which consists of only the dominant (Dom_ε), secondary (Sec_ε) and tertiary (Ter_ε) perceived-emotions for each excerpt (section 3.3.2). In this case, the estimated parameters (k=6) are μ_{EWIV_R}=$[\mu_{dom}, \mu_{sec}, \mu_{ter}]$, and σ_{EWIV_R}=$[\sigma_{dom}, \sigma_{sec}, \sigma_{ter}]$. The log-likelihoods are estimated using equation 3.14.

For the Circumplex model [2], the self-reported tuple (A, V) is modeled using a Normal distribution $\mathcal{N}_{AV}(\mu_{AV}, \sigma_{AV})$, where, μ_{AV}=$[\mu_A, \mu_V]$, and σ_{AV}=$[\sigma_A, \sigma_V]$. The (k=4) parameters (μ_{AV}, σ_{AV}) and the corresponding log-likelihood ($\hat{L}_{AV}$) are estimated in the same way as the previous models. A hypothetical *integrated model* (EWIV+AV) is also constructed for comparison, where music-perceived emotion is represented using both EWIV and AV formats. The number of parameters estimated is the sum of the parameters of the two parent models (k=20). Finally, using equation 3.1 we calculate the AIC values for each model, on each excerpt, which are represented by AIC_{EWIV}, AIC_{EWIV_R}, AIC_{AV} and $AIC_{EWIV+AV}$ respectively.

3.7.2.2 Results: Comparison of AIC across models

In this section, we compare the calculated AIC values to identify the representation model that fits best, for three different datasets.

Eerola's Dataset [5] (section 3.2.2.1): We sincerely thank the authors of Eerola et al. [5] for allowing us to use their data for our experiments. This dataset contains perceived-emotion opinion data in both discrete and dimensional formats. First, the conversion procedures (section 3.7.1) are used to obtain the converted discrete and dimensional datasets. Next, AIC values are calculated for each excerpt over the original, converted, and integrated data format (section 3.7.2.1). The results for the first five excerpts of *Eerola's Dataset* [5] are presented in table 3.8. Tables 3.9 and 3.10 present results of AIC calculations for all the 110 excerpts from *Eerola's Dataset* [5]. The columns *Excerpt#* and *Emotion Category*

Table 3.9: AIC results for the first 50 excerpts from Eerola's Dataset [5]. The terms AIC_{EWIV}, AIC_{EWIV_R}, and AIC_{AV} represent AIC values for EWIV, Reduced EWIV, and AV models respectively. Columns (1)-(3) represent AICs calculated over emotion data from the original dataset. Columns (4)-(5) give the AICs calculated over converted emotion data. Column (6) presents the AIC of an integrated model, constructed using emotion data from the original dataset. The best model for each excerpt is highlighted in blue, and the second-best in gray.

Excerpt	Emotion Category	Original Data			Converted Data		Integrated Data
#	Level	AIC_{EWIV}(1)	AIC_{EWIV_R}(2)	AIC_{AV}(3)	AIC_{EWIV}(4)	AIC_{AV}(5)	$AIC_{EWIV+AV}$(6)
1	Anger High	466.93	444.44	469.69	613.62	570.89	1116.87
2	Anger High	389.67	388.94	393.99	631.80	536.12	1041.98
3	Anger High	487.46	458.84	486.89	759.68	596.32	1129.84
4	Anger High	463.40	406.24	463.46	727.81	554.93	1264.48
5	Anger High	520.89	494.76	528.40	980.60	578.77	1230.19
6	Anger Moderate	446.51	428.75	466.51	804.05	739.26	1305.72
7	Anger Moderate	454.97	434.83	450.14	763.46	575.67	1201.97
8	Anger Moderate	528.01	393.41	539.93	858.06	780.61	1421.34
9	Anger Moderate	489.79	461.98	476.07	924.23	653.10	1629.10
10	Anger Moderate	441.77	424.16	438.74	856.49	542.97	1151.90
11	Fear High	444.37	463.13	484.23	748.16	591.75	1088.77
12	Fear High	482.07	448.29	487.40	863.01	572.48	1265.72
13	Fear High	477.92	454.62	499.52	841.03	601.59	1161.22
14	Fear High	398.52	445.05	429.49	657.21	527.67	902.06
15	Fear High	453.48	445.36	464.89	777.29	543.91	1021.55
16	Fear Moderate	524.92	455.07	534.26	730.40	668.32	1325.42
17	Fear Moderate	523.62	418.51	532.33	889.09	696.12	1194.15
18	Fear Moderate	515.73	472.02	525.40	860.52	638.28	1198.73
19	Fear Moderate	543.12	473.75	512.68	803.69	761.79	1400.76
20	Fear Moderate	452.68	438.88	441.89	922.40	659.83	1282.56
21	Happy High	450.74	428.98	435.88	850.92	519.01	1227.55
22	Happy High	453.77	414.49	459.97	638.19	600.03	1135.08
23	Happy High	371.89	289.77	536.59	974.20	543.95	735.12
24	Happy High	426.78	326.74	440.89	630.33	598.52	947.70
25	Happy High	459.98	418.22	472.51	883.82	525.10	1020.14
26	Happy Moderate	548.61	416.54	551.75	860.34	582.73	1304.09
27	Happy Moderate	517.58	471.8	521.43	962.48	594.77	1171.19
28	Happy Moderate	525.02	416.05	525.97	807.14	497.91	1776.94
29	Happy Moderate	495.97	433.47	509.34	888.99	516.16	1705.71
30	Happy Moderate	516.30	456.1	513.63	894.26	507.67	1207.24
31	Sad High	587.85	487.73	570.97	977.68	646.24	1422.18
32	Sad High	554.74	433.04	559.49	954.94	759.06	1428.77
33	Sad High	531.76	451.99	528.32	771.98	591.75	1226.78
34	Sad High	571.83	448.58	576.33	837.69	647.46	1336.22
35	Sad High	594.11	392.17	570.95	802.60	784.81	1571.00
36	Sad Moderate	502.75	470.05	513.58	960.17	617.34	1323.73
37	Sad Moderate	564.64	461.78	562.21	859.42	728.48	1440.09
38	Sad Moderate	525.60	439.46	525.61	721.78	662.28	1263.71
39	Sad Moderate	535.06	481.49	536.97	892.52	530.65	1286.38
40	Sad Moderate	575.30	441.41	578.29	853.34	641.75	1292.22
41	Tender High	482.90	448.96	489.94	770.50	496.28	1104.16
42	Tender High	510.93	448.14	504.37	772.32	506.93	1662.04
43	Tender High	482.12	447.25	500.71	861.07	512.73	1078.86
44	Tender High	470.42	460.86	471.23	737.54	519.95	1070.55
45	Tender High	492.55	481.32	501.36	973.64	526.03	1711.26
46	Tender Moderate	543.40	440.46	544.63	700.20	543.22	1276.64
47	Tender Moderate	511.20	444.69	516.93	639.26	498.12	1187.99
48	Tender Moderate	547.99	449.72	559.67	873.46	632.03	1327.32
49	Tender Moderate	514.18	422.69	516.19	846.62	499.83	1713.06
50	Tender Moderate	551.22	453.14	557.53	843.83	577.79	1284.05

Level contain the excerpt numbers and emotion category levels from the original dataset. The columns *Original Data*, *Converted Data*, and *Integrated Data* present the AIC values calculated using different representation models over original, converted, and original integrated emotion data. The subheadings AIC_{EWIV}, AIC_{EWIV_R}, AIC_{AV} and $AIC_{EWIV+AV}$ represent AIC values for various representation models, and are numbered (1)-(6).

Observation 1: Among all the models, $AIC_{EWIV_R}(2)$ is consistently the least and $AIC_{EWIV+AV}(6)$ is consistently the highest. This indicates that model EWIV_R fits the emotion data the best. Though the model (EWIV+AV) has the highest number of parameters, the relative quality of this model is poor, indicating that increasing the number of model parameters does not necessarily make the model a better fit for data. This result holds true for 98% of the excerpts in this dataset. Only in 2 cases EWIV (column (1)) is found to perform better.

Observation 2: For most excerpts (almost 60%), $AIC_{EWIV}(1) < AIC_{AV}(3)$, indicating that EWIV representation model is a better fit. For the rest of the excerpts, the dimensional Arousal-Valence representation model performs better. In some cases, the difference in AIC values of the two competing models is ≤ 2 (e.g. *excerpt#* 3,4), indicating that both models perform similarly.

Observation 3: The AIC values calculated over converted data (columns 4-5), are higher than those calculated over the original emotion data, indicating some loss in information due to the conversion.

***Soleymani_5* Dataset** (section 3.4.1): The original dataset provides data in the arousal-valence format, and EWIV data was collected for the purpose of this study (table 3.2). Table 3.11 contains detailed metadata of the *Soleymani_5* [3] dataset, from the original dataset as well as from the converted and collected EWIV data. A similar procedure was followed for *Eerola's Dataset*. Data format conversions (section 3.7.1) were performed and AIC values were calculated (section 3.7.2.1) for each excerpt over the original, collected, converted, and integrated data The results are presented in table 3.12.

Observation 1: Among all the models, $AIC_{EWIV_R}(3)$ is consistently the least (best fit) and $AIC_{EWIV+AV}(6)$ is consistently the highest.

Observation 2: The second best model varies across excerpts, for some it is EWIV (column $AIC_{EWIV}(2)$), and for others, it is AV (column $AIC_{AV}(1)$).

Table 3.10: AIC results for the last 60 excerpts from Eerola's Dataset [5]. The terms AIC_{EWIV}, AIC_{EWIV_R}, and AIC_{AV} represent AIC values for EWIV, Reduced EWIV, and AV models respectively. Columns (1)-(3) represent AICs calculated over emotion data from the original dataset. Columns (4)-(5) give the AICs calculated over converted emotion data. Column (6) presents the AIC of an integrated model, constructed using emotion data from the original dataset. The best model for each excerpt is highlighted in blue, and the second-best in gray.

Excerpt	Emotion Category	Original Data			Converted Data		Integrated Data
#	Level	AIC_{EWIV}(1)	AIC_{EWIV_R}(2)	AIC_{AV}(3)	AIC_{EWIV}(4)	AIC_{AV}(5)	$\text{AIC}_{EWIV+AV}$(6)
51	Valence Pos High	667.54	431.28	698.20	977.81	672.91	1641.81
52	Valence Pos High	649.97	417.77	676.56	1315.61	665.63	1561.81
53	Valence Pos High	685.20	433.81	680.90	1123.54	741.00	1653.21
54	Valence Pos High	749.52	451.16	754.08	1431.18	745.06	1836.83
55	Valence Pos High	647.02	355.19	677.54	1102.78	764.17	1151.34
56	Valence Pos Moderate	696.20	411.89	700.47	1198.89	865.63	1171.05
57	Valence Pos Moderate	729.78	360.78	727.03	1279.30	953.54	1346.99
58	Valence Pos Moderate	726.64	441.72	737.45	1477.60	823.18	1270.06
59	Valence Pos Moderate	673.62	428.44	700.13	1381.13	738.17	1207.89
60	Valence Pos Moderate	727.97	357.42	734.73	1404.83	969.21	1314.46
61	Valence Neg Moderate	762.05	385.96	747.95	1128.76	1000.76	1299.62
62	Valence Neg Moderate	663.36	426.49	654.38	1340.19	847.18	1636.91
63	Valence Neg Moderate	745.94	419.82	742.73	1340.68	1016.17	1310.64
64	Valence Neg Moderate	665.11	377.09	661.81	1299.78	844.26	1165.7
65	Valence Neg Moderate	716.50	366.56	732.92	1364.15	998.29	1336.48
66	Valence Neg High	481.34	447.22	536.34	1362.52	737.60	989.74
67	Valence Neg High	703.69	406.05	727.21	1272.81	1035.99	1285.97
68	Valence Neg High	659.70	426.82	666.47	1244.64	789.45	1613.77
69	Valence Neg High	583.30	437.49	588.15	1191.51	710.79	1580.15
70	Valence Neg High	544.00	398.16	730.84	1313.44	749.81	1055.00
71	Energy Pos High	588.75	339.64	589.86	1022.27	696.46	1139.25
72	Energy Pos High	443.07	206.33	604.34	968.47	814.67	833.55
73	Energy Pos High	488.86	383.56	679.87	1041.96	968.02	966.58
74	Energy Pos High	691.64	402.41	699.96	1346.17	1005.66	1292.9
75	Energy Pos High	491.79	310.72	656.62	1032.59	741.06	931.24
76	Energy Pos Moderate	684.85	404.48	687.11	1267.32	706.61	1556.91
77	Energy Pos Moderate	608.91	340.53	744.32	1234.61	939.48	1127.12
78	Energy Pos Moderate	770.54	424.20	772.60	1538.18	1052.23	1402.3
79	Energy Pos Moderate	421.07	297.43	673.40	971.25	876.41	875.83
80	Energy Pos Moderate	661.84	346.41	734.70	1292.16	1048.70	1166.74
81	Energy Neg Moderate	661.02	395.8	696.69	1256.87	754.26	1394.15
82	Energy Neg Moderate	621.87	406.6	744.78	1179.43	842.11	1119.68
83	Energy Neg Moderate	734.33	448.01	750.41	1520.44	787.90	1944.62
84	Energy Neg Moderate	555.13	408.73	728.42	1041.01	824.11	1050.71
85	Energy Neg Moderate	746.64	408.37	744.09	1239.52	1053.12	1268.61
86	Energy Neg High	678.79	393.85	721.76	1342.04	1014.32	1176.52
87	Energy Neg High	768.02	422.3	768.28	1564.27	1008.61	1817.25
88	Energy Neg High	719.66	360.68	722.99	1347.13	813.39	1601.85
89	Energy Neg High	608.63	419.5	775.39	1180.94	1009.54	1146.83
90	Energy Neg High	731.03	413.44	740.22	1270.97	1016.17	1797.94
91	Tension Pos High	654.12	428.56	651.75	1410.03	805.03	1051.8
92	Tension Pos High	655.44	428.51	666.21	1458.62	850.97	1589.42
93	Tension Pos High	646.11	442.9	669.50	1358.31	787.18	1605.81
94	Tension Pos High	659.36	438.44	674.49	1220.20	866.46	1163.43
95	Tension Pos High	693.64	440.04	695.40	1606.19	824.52	1707.55
96	Tension Pos Moderate	735.60	395.75	749.80	1499.07	1027.66	1354.9
97	Tension Pos Moderate	572.71	426.44	711.43	1058.72	986.61	1074.16
98	Tension Pos Moderate	737.20	407.36	751.37	1378.84	973.09	1384.49
99	Tension Pos Moderate	676.50	401.83	674.36	1315.17	1030.85	1324.22
100	Tension Pos Moderate	657.38	366.01	652.31	1114.75	827.18	1407.59
101	Tension Neg Moderate	668.59	425.53	655.34	1500.82	677.99	1537.68
102	Tension Neg Moderate	686.13	438.32	713.33	1421.53	856.23	1175.24
103	Tension Neg Moderate	771.86	411.58	785.29	1242.87	860.08	1355.4
104	Tension Neg Moderate	685.61	438.94	738.93	1393.04	737.90	1183.95
105	Tension Neg Moderate	587.09	270.15	622.39	1082.65	674.35	1173.94
106	Tension Neg High	670.39	378.88	674.97	1216.21	687.54	1446.91
107	Tension Neg High	629.87	431.47	616.59	1319.16	672.26	1579.08
108	Tension Neg High	557.82	425.43	698.66	1082.22	750.76	1042.64
109	Tension Neg High	523.24	391.91	702.30	1209.96	732.38	1009.78
110	Tension Neg High	675.46	435.9	683.78	1569.61	687.39	1675.17

Observation 3: AIC values calculated over converted data are higher than those calculated over the original emotion data, indicating some loss in information due to the conversion.

Table 3.11: Detailed metadata of excerpts from *Soleymani_5* [3] dataset. *Excerpt#*=Clip number from original dataset. The Circumplex representation metadata from the original dataset are given by A_{Stat}=Static arousal, V_{Stat}=Static valence, A_{Dyn_Avg}=Dynamic arousal average, V_{Dyn_Avg}=Dynamic valence average, Θ=Calculated angular value, *Emotion Region*=Corresponding region on 2D plane. The *dominant* and *secondary* emotions observed from the EWIV surveys (collected) and converted are presented in the last four columns.

Excerpt#	Circumplex Metadata						EWIV (Collected)		EWIV (Converted)	
	$\mathbf{A}_{Stat}$	$\mathbf{V}_{Stat}$	$\mathbf{A}_{Dyn\_Avg}$	$\mathbf{V}_{Dyn\_Avg}$	Θ	Emotion Region	Dominant Emotion	Secondary Emotion	Dominant Emotion	Secondary Emotion
128	3.6	4	-0.32	-0.23	234.29°	Bored	Sadness	Fear	Sadness	-
178	2.8	3.7	-0.51	-0.13	255.70°	Droopy	Sadness	Romance	Sadness	Calmness
171	2.1	3.6	-0.45	-0.19	247.11°	Sleepy	Calmness	Romance	Calmness	Sadness
191	5.1	6.3	0.09	0.32	15.71°	Delighted	Happiness	Excitement	Happiness	Excitement
294	5.5	6.5	0.20	0.42	25.46°	Delighted	Happiness	Excitement	Happiness	Excitement

Table 3.12: AIC results for the Soleymani_5 [3] excerpts. The terms AIC_{EWIV}, AIC_{EWIV_R} and AIC_{AV} represent AIC values for EWIV, Reduced EWIV, and AV models respectively. Column (1) gives the AIC calculated over emotion data from the original dataset. Columns (2)-(3) give the AIC calculated over emotion data collected for these excerpts in the EWIV format. Columns (4)-(5) give the AIC calculated over converted emotion data. Column (6) presents the AIC of an integrated model, constructed using emotion data from the original and collected dataset. The best model for each excerpt is highlighted in blue, and the second-best in gray.

Excerpt#	Emotion Category		Original Data	Collected Data		Converted Data		Integrated Data
	Dominant	Secondary	AIC_{AV}(1)	AIC_{EWIV}(2)	AIC_{EWIV_R}(3)	AIC_{EWIV}(4)	AIC_{AV}(5)	$\text{AIC}_{EWIV+AV}$(6)
128	Sadness	Fear	206.97	210.31	115.83	217.79	213.01	508.96
178	Sadness	Romance	245.15	273.94	170.26	312.73	276.33	510.37
171	Calmness	Romance	237.12	238.28	178.78	297.10	289.10	481.41
191	Happiness	Excitement	201.98	196.57	132.90	254.52	247.74	411.53
294	Happiness	Excitement	237.29	231.67	157.41	307.12	301.77	492.27

***EmoRaga* Dataset** (section 3.4.1.2): The dataset provides EWIV format data, which was converted to dimensional format (section 3.7.1.1), and AIC values were calculated (section 3.7.2.1) for each excerpt over the collected and converted datasets. The results for the first five excerpts are presented in table 3.13. Table 3.13 presents results of AIC calculations for all the 48 excerpts from *EmoRaga Dataset*.

Observation 1: The best model is consistently observed to be EWIV_R (column $AIC_{EWIV_R}(2)$).

Observation 2: The second best model varies across excerpts, for some (almost 60%) it is EWIV (column $AIC_{EWIV}(1)$), and for others, it is AV (column $AIC_{AV}(3)$).

3.8 Discussion and Conclusion

In this chapter, we introduce a dynamic intensity rating-based categorical emotion representation adapted for perceived-emotion studies in HCM, called the Emotion-Descriptor Intensity-Value (EWIV) representation. We discuss the choice of *emotion-descriptors* and establish a mathematical procedure to estimate music perceived-emotion as a *probability vector* using *EWIV*. We present the *EmoRaga* dataset, dedicated to perceived-emotion studies in HCM, and introduce the term *emotion motifs* in HCM to indicate any musical features that can possibly cue the perception of certain emotions in HCM. Using EWIV-based self-report survey results on benchmark and *EmoRaga* datasets, we validate the application of EWIV representation and study *typicality*, *consensus*, and *ambiguity* in emotion opinion data. In order to understand the extent of EWIV's applicability in MER, we perform classification, emotion variation detection, and contextual influence measurements and obtain satisfactory results. Finally, we evaluate the quality of EWIV and other emotion representation models using a statistical goodness-of-fit measure of AIC.

Table 3.13: AIC results for the *EmoRaga* dataset. The terms AIC_{EWIV}, AIC_{EWIV_R}, and AIC_{AV} represent AIC values for EWIV, Reduced EWIV, and AV models respectively. Column (1)-(2) gives the AIC calculated over emotion data collected for these excerpts in the EWIV format. Column (3) give the AIC calculated over emotion data converted from EWIV to AV representation. The best model for each excerpt is highlighted in blue, and the second-best in gray. Columns A_{avg}, V_{avg}, and Θ represent the average arousal, valence, and calculated angle of the converted AV representation.

Excerpt#	Emotion Category		Collected Data		Converted Data			
	Dominant	Secondary	AIC_{EWIV}(1)	AIC_{EWIV_R}(2)	A_{avg}	V_{avg}	Θ	AIC_{AV}(3)
1	Happiness	Excitement	540.33	186.12	0.41	1.99	78.26	592.91
2	Happiness	Excitement	712.63	200.44	0.73	1.75	67.30	724.69
3	Sadness	Calmness	719.46	192.92	-0.92	0.91	224.40	688.13
4	Sadness	Calmness	712.63	203.81	-0.71	-1.04	235.82	742.6
5	Calmness	Happiness	723.54	160.13	-0.48	0.74	302.92	773.06
6	Calmness	Happiness	782.2	211.5	-0.48	1.05	294.63	780.28
7	Excitement	Sadness	628.75	205.95	0.78	0.42	28.45	644.83
8	Calmness	Sadness	808.57	197.47	-0.62	0.42	325.91	812.6
9	Calmness	Happiness	618.09	205.37	0.15	0.46	77.11	577.64
10	Calmness	Excitement	726.79	176.74	0.06	0.40	81.41	691.27
11	Sadness	Calmness	860.26	226.82	-0.27	0.29	313.17	857.67
12	Calmness	Sadness	735.76	150.34	-0.12	0.57	282.72	737.34
13	Sadness	Calmness	828.47	229.12	-0.48	0.005	359.39	788.16
14	Happiness	Excitement	719.13	175.88	0.19	1.36	81.79	730.11
15	Sadness	Calmness	417.39	191.99	-0.36	-0.70	242.71	630.00
16	Sadness	Calmness	727.50	191.09	-0.86	-0.38	203.87	731.01
17	Sadness	Calmness	718.4	194.04	-0.64	-0.56	221.63	777.85
18	Sadness	Calmness	877.14	251.43	-0.97	-0.25	194.41	875.43
19	Sadness	Calmness	647.61	246.93	-0.13	-0.65	258.52	604.11
20	Sadness	Calmness	824.27	202.7	-0.40	-0.78	242.62	826.75
21	Calmness	Happiness	838.05	202.18	-0.47	0.61	307.81	808.32
22	Calmness	Fear	804.77	174.11	0.07	-0.35	102.39	868.84
23	Happiness	Excitement	824.29	188.79	0.27	0.80	70.95	822.96
24	Sadness	Calmness	737.71	199.29	-0.76	-0.01	180.75	735.09
25	Happiness	Excitement	721.99	161.08	0.21	0.91	76.84	724.23
26	Calmness	Sadness	729.52	193.37	-0.56	0.07	352.30	740.77
27	Calmness	Excitement	761.65	192.39	-0.03	0.73	272.55	761.87
28	Calmness	Happiness	758.99	182.67	-0.24	0.79	286.96	732.59
29	Happiness	Romance	729.87	173.39	-0.11	0.65	280.35	733.38
30	Calmness	Wonder	783.56	186.47	-0.25	0.56	294.19	702.66
31	Sadness	Calmness	569.62	177.94	-0.07	-0.14	242.60	571.33
32	Calmness	Happiness	639.27	183.57	-0.01	0.66	271.58	659.65
33	Sadness	Calmness	816.97	193.83	0.22	-0.17	142.47	869.64
34	Sadness	Calmness	533.27	196.44	-0.26	-0.36	233.93	534.7
35	Happiness	Calmness	748.82	166.11	0.03	0.97	87.77	725.3
36	Sadness	Calmness	780.1	201.77	-0.48	-0.18	200.60	772.81
37	Happiness	Romance	770.4	159.6	0.008	0.60	89.26	794.03
38	Anger	Sadness	498.61	168.49	0.41	-0.01	178.11	501.13
39	Sadness	Fear	540.41	167.8	0.51	-0.18	131.42	405.75
40	Sadness	Calmness	564.21	172.39	0.11	0.006	3.11	588.47
41	Calmness	Romance	722.80	158.93	-0.37	0.34	317.85	765.15
42	Sadness	Calmness	736.40	175.7	-0.13	0.17	306.67	732.63
43	Sadness	Fear	757.94	196.92	-0.05	-0.53	264.00	791.74
44	Sadness	Calmness	580.89	234.56	-0.14	-0.49	253.93	588.5
45	Sadness	Fear	740.12	178.02	0.10	-0.51	101.68	749.6
46	Sadness	Fear	707.58	231.65	0.29	-0.18	148.48	754.1
47	Happiness	Calmness	637.24	196.87	0.04	0.54	85.39	639.53
48	Calmness	Romance	615.89	206.21	-0.11	0.37	285.72	600.63

CHAPTER 4

Exploring *Intra-Contextual Influences* in Music Emotion Perception

Context is one of the key parameters that influence music emotion perception in listeners. The current study systematically investigates the influence of *immediate intrinsic musical context* on the perception of music-evoked emotions. Four dominantly happy and four dominantly sad Hindustani classical music excerpts were chosen and rated for perceived emotions in two types of listening experiments. In the first experiment, *baseline* ratings of these excerpts were collected from general participants to establish the emotions perceived without any kind of influence. In the second experiment, pairs of excerpts were presented in succession, without any breaks, and rated continuously for the emotions perceived. The perceived emotion responses for the second excerpt in the pair were considered for possible *influence* by the preceding music excerpt. The baseline and *influenced* ratings were compared to understand the effect of immediate intrinsic musical context - the first excerpt in this case - on emotion perception. Significant differences in terms of intensity of perceived emotions were found between baseline and *influenced* conditions for both happy and sad excerpts, either in the form of changes

in perceived dominant emotion or in perceived nuance, and this phenomenon was termed *intra-contextual influence*. The results reflect the importance of immediate intrinsic musical context on emotion perception of musical excerpts and might be used to understand music semantics better.

4.1 Introduction

Music and emotions are known to have an intrinsic link with one another [28], [10], [11], from the perspectives of music composers, performers and the audience alike [34], [35], [9]. Thus, understanding and explaining music-evoked emotions is an important research topic. Over the past decades, research has begun to explore the interrelations of the complex set of different factors influencing musical emotions [58]. However, to the best of our knowledge, there is a dearth of systematic investigation on how music itself can affect emotion perception and eventual meaning-making in successive music.

Among many existing studies to explain musical emotions, mention must be made of theories of cognitive appraisal, musical expectancies [64], and the unified theoretical framework of BRECVEMA - an acronym for 8 mechanisms: Brain Stem Reflex, Rhythmic Entrainment, Evaluative Conditioning, Contagion, Visual Imagery, Episodic Memory, Musical Expectancy, and Aesthetic Judgment. [11], [65]. Hargreaves et al. [66], [33] put forward the *reciprocal-feedback* model of responses to music, where three main determinants of musical response were considered: a) music related properties [67], [68], [69], [70], b) listener-specific properties [71], [72], [73], [74], and c) the listening situation (context related). In this study, we are particularly interested in the *context* related determinants.

Hargreaves et al. [33] identified four types of *contexts* - socio-cultural [79], [89], everyday situations [90], presence/absence of other person(s) [78], and other activities [91]. The concepts of *musical context* and its influence on stimuli-evoked emotions are relatively new. Steffens et al. [93] and Herget et al. [94] explored the influence of *musical context* on video-evoked emotions. In these studies, *musical context* refers to the music being played along with the target video. In some studies [95], *musical context* includes various earlier musical experiences, music-related memories, and responses to music. These studies indicate a rather fluid

interpretation of the term *musical context*.

Discussions of context and emotional responses to stimuli are incomplete without mention of affective priming [25], emotion induction [24], and contagion [10]. In the case of emotion induction [24] only a specific mood state is instigated. Same (or almost similar) emotions are propagated from the emotion induction medium (music) to the mood induction target (listener). Emotional contagion [10], [51] is a process whereby an emotion is induced [24] by music, because the listener perceives the emotional expression of the music and *mimics* this expression internally. Priming is the procedure that entails that exposure to one stimulus may influence a response to a subsequent stimulus, without conscious guidance or intention. Timmers et al. [99], [100] investigated how listeners' perception of auditory sequences (target) change dynamically, depending on emotional context primed through affective pictures (prime) of people depicting emotional expressions. Huziwara et al. [101] investigated whether affective priming occurs when chords are used as primes, and faces (happy or sad) are used as targets. Tay et al. [102], Steinbeis et al. [103], Armitage et al. [104], [105], and Goerlich et al. [106] investigated various aspects of auditory priming and its effect on usage, processing, and perception of words. In all these cases, the prime and the target are different media, essentially capturing the effect of affective priming by music on the perception of other stimuli or vice versa. In the present work, we aim to study the effect of preceding musical stimuli (as context) on the emotion perception of successive musical stimuli.

4.1.1 Immediate Intrinsic Musical Context

In this study, we define the term *immediate intrinsic musical context* as the musical frame of reference within which a listener listens to a particular musical composition. It might include musical structures, notes or note-clusters, and small music excerpts which *immediately* precede the music piece being evaluated for an emotion judgment. The term *intrinsic* refers to the context being an inherent part of the music being heard, and not extrinsic (generated outside the music) like external events, audio, or otherwise. We explore the influence of *immediate intrinsic musical context* on the perception of music-evoked emotions. We term

this possible phenomenon as *intra-contextual influence* - since the *context* investigated for being responsible for possible variation in the perception of music-evoked emotions is the preceding music itself. The term *intra* signifies the influence occurring *within* the reference of a single temporal entity - music. The relevance of scrutinizing possible effects of *intra-contextual influence* stems from the fact that in most previous works with *musical context*, the target musical stimuli are treated in isolation. The effects of various contexts on the isolated stimuli, or the effects the stimuli have on other media are then studied. But in reality, very often musical compositions are created by weaving together numerous smaller pieces. These smaller pieces denote subtle variations in musical structure and emotions, which increase the aesthetic value of the whole. In isolation, each smaller piece might evoke a different emotion than when heard as part of a whole. Thus, understanding if and how the antecedent musical phrases/sequence of notes/sections influence the emotion perception in consequent music sections might help in comprehending the perception of music semantics. An analogy might be drawn with sentences in a paragraph. If we take individual sentences out of the *immediate context* of the paragraph, the (emotional) meaning of the sentences might change. To understand the effect of each sentence, the *immediate context* is important.

4.1.2 Hindustani Classical Music

Though the concepts of *immediate intrinsic musical context* and *intra-contextual influence* are arguably applicable to all music traditions, we focus on the Hindustani classical music (HCM) of India in the present work. HCM is one of the two main branches of traditional Indian music. It is primarily based on the *raga* framework [151], characterized by a set of notes, the ascending-descending melodic progression, and a specific set of melodic phrases [152], the *tala* (rhythmic cycle) and the *laya* (tempo). In any standard HCM *raga* rendition, the composition consists of an arrhythmic phase followed by rhythmic phases of varying tempo, each forming parts of a whole. The perception of one part plays a significant role in how subsequent parts are perceived, building up to an emotional crescendo at the climax. Compared to Western traditions, HCM renditions focus on the intensification of a cluster of related/congruent emotions. The use of contradict-

ing/incongruent emotions in musical phrases is widespread to enhance the perception of other emotions and increase aesthetic appreciation. In the *Rasa* theory [22], this is termed the intensification of the *sthayee bhava* (dominant emotion) through the use of a *sanchaari bhava* (a glimmer of a contradicting emotion) as an impulse. Very often, hope and despair, romance and sadness, or excitement and calmness are paired in subtle ways, to intensify the perception of emotions. It is also a common practice in HCM renditions to improvise the same music phrase in different ways, evoking subtle nuances of perceived emotions. The study of such aesthetic stylistic components and their influence on listener-perceived emotions is possible through concepts of *immediate intrinsic musical context* and *intra-contextual influence*. Though many studies on the effects of Indian music on human psycho-physiological [156], [157] and emotional [153], [158] responses have been undertaken, to the best of our knowledge, no systematic study exists on the influence of *contexts* on perceived emotions in HCM, compared to its western counterpart. Non-availability of excerpt scores, the high cost of manual annotations of emotion and related metadata by both general listeners and experts, and inherent dissimilarities between form-fluid HCM and structured Western music (more popular in the music psychology field) might be some possible reasons. In this chapter, we attempt to bridge this gap.

4.1.3 Aims of the study

The main aim of the present study is to explore the possibility of *perceived* musical emotions being *influenced* by *intra-contextual* variations. Therefore, the research questions are:

RQ1: Is there any significant effect of *intra-contextual influence* on music emotion perception?

RQ2: Can the possible effects of *intra-contextual influence* on the perception of musical emotions be generalized across music excerpts grouped according to their dominant emotions?

The rest of the chapter describes various experiments and results related to *intra-contextual influence*, and concludes with a discussion on the contributions of this study, along with a comparison between emotion priming, emotion contagion,

Table 4.1: Details of the eight chosen excerpts. The excerpt numbers, Raga of origin, duration, and dominant and secondary emotions are presented.

Excerpt #	Raga/ Origin	Length (Secs)	Dominant Emotion	Secondary Emotion
1	Hamsdhwani	40	Happiness	Calmness
2	Adana	51	Happiness	Calmness
3	Desh	46	Happiness	Excitement
4	Desh	52	Happiness	Calmness
5	KR Asavari	45	Sadness	Calmness
6	Marwa	47	Sadness	Calmness
7	Marwa	40	Sadness	Calmness
8	Marwa	50	Sadness	Calmness

emotion induction, and the observed results of *intra-contextual influence*.

4.2 Method

A self-report-based study on perceived music emotion was conducted for the present work. Data were collected through online surveys and emotion opinions were approached through time-continuous ratings of music excerpts. The collected data was analyzed using various statistical approaches.

4.2.1 Expert Selection of Stimuli

In order to obtain unknown and emotionally unambiguous HCM excerpts, an expert panel was organized for choosing the stimuli. The expert panel first selected a set of *raga*s which are known to be predominantly perceived as happy (*Hamsdhwani*, *Desh*, *Adana*) or sad (*Marwa*, *Komal Rishabh Asavari*), both in Indian musical knowledge systems (Natyashastra by Bharatamuni, [22]), and in literature [153]. Further, each expert selected 3 short excerpts that they perceived as representative of happy/sad target emotions, from full-length recordings of these *raga*s by eminent Sitar (Hindustani classical instrument) maestros. To ensure uniformity among the excerpts, the following criteria were established: a) The excerpts were of duration between 40-60 seconds, depending on the natural phrasing of the excerpt. b) They were part of an instrument recital, to avoid the effect of lyrics

on perceived emotion. In the present study, only Sitar excerpts were used to avoid possible instrument-based bias. c) Only those excerpts were included that were dominantly perceived as happy or sad under no intentional influence. Next, the panel rated all chosen excerpts with their perceived emotion opinions on a scale of 1-5. Emotion categories included happiness, calmness, wonder, excitement, anger, fear, sadness, and romance. Finally, 8 excerpts were chosen with the highest mean ratings for happiness or sadness, for further use in this study. The emotion with the highest and second-highest ratings indicated the *dominant* and *secondary* perceived emotions of each excerpt. The details of these excerpts are presented in table 4.1.

4.2.2 Participants

Two types of listener-participants took part in this study - experts and general. The expert panel of participants identified the stimuli and consisted of five university faculty members and students, who are HCM practitioners and musicologists. After this, 160 general participants took part in the surveys using these excerpts. The majority were students belonging to different courses at the university. Some faculty, staff, and their family members also volunteered. 69.95% of the general participants identified as male ($\mu_{age} = 20.21, \sigma_{age} = 4.89, range_{age} = [13, 56]$), and 30.05% identified as female ($\mu_{age} = 22.64, \sigma_{age} = 6.99, range_{age} = [13, 59]$). The students were awarded grade points for their participation. All participants were Indian nationals.

4.2.3 Ethical considerations

Participants were informed of the nature and objective of the study prior to the surveys. Participation was voluntary and participants provided online consent before accessing the online survey. Response anonymity and pure academic use of collected data were guaranteed.

4.2.4 Survey Interface Description

All surveys started with an instruction page containing a short description of how to use the interface to report continuous-time perceived emotions during a music excerpt (figure 3.1). It also explained the meaning of *perceived* emotion versus *felt* emotion, and asked the participants to report *"emotions that you perceive or recognize from the music while listening to it and not that which you yourself feel"*. The novel online interface (figure 3.1) [182] consisted of the *central part*: with the play and pause buttons for the music, the concentric grey *neutral part*: the region where the cursor rested while a participant was not actively self-reporting perceived emotions, and the *10 spokes*: eight of which represented eight emotions, Happiness (H), Calmness (C), Wonder (W), Excitement (E), Anger (A), Fear (F), Sadness (S), Romance (R). In the wheel, Other emotion (OE) represented the possibility of absence of perceived emotion, and Don't know (DK) represented the participant's indecisiveness regarding the perceived emotion. Each spoke was divided into 5 sub-regions indicating the emotion intensity levels. The inner-most region, with the lightest shade of color, marked the lowest intensity of 1. The outermost region with the darkest color shade represented the highest intensity of 5. If an emotion is not reported at any instant, it is considered to be intensity 0, i.e. absence of that emotion.

4.2.5 Surveys

4.2.5.1 Baseline Surveys

were conducted to get the ground truth (baseline) perceived emotions of each excerpt. Participants were asked to report their perceived emotions from each excerpt presented to them in isolation, without any intentional musical influence. A time gap of 2 minutes was provided between each excerpt, where participants reported the prominent reasons behind their perceived emotions.

4.2.5.2 Influenced Surveys

were conducted to record the possible *intra-contextual influence* of one excerpt on the perceived emotion of the subsequent excerpt. Two excerpts were presented

Table 4.2: Two types of influenced Surveys, and the four possible excerpt sequences based on their dominant emotions.

Survey Type	Influencer Excerpt Type	Influenced Excerpt Type
Influenced_by_Happy	Happy →	Happy
	Happy →	Sad
Influenced_by_Sad	Sad →	Happy
	Sad →	Sad

in sequence without any time gap. The first one in the sequence was termed the *influencer* excerpt, and the subsequent one was termed the *influenced* excerpt. The dominant emotion of an *influenced* excerpt was considered its *target* emotion. All other emotions perceived were termed *non-target* emotions. The aim was to study if and how the presence of the *influencer* excerpt modified the target and non-target emotion perceptions of the *influenced* excerpt. Since only dominantly happy and sad excerpts were used in this study, four types of sequences were possible (table 4.2). Since eight excerpts were used, seven such sequences were surveyed for each *influenced* excerpt. Thus, for each type of excerpt (happy/sad), 28 such sequence combinations were generated, with a maximum duration of 103 seconds for each combination.

4.2.6 Procedure

At the individual level, each participant was directed to listen to the music and simultaneously respond with the perceived emotion intensity in the wheel as desired, and as many times as they felt necessary. The demographic questionnaire and other relevant survey questions were presented as web forms. These questions included name, gender, age, academic qualifications, interest in music, and musical training. Additionally, participants were also asked to identify prominent reasons that they thought governed their perception. The reasons were provided as multiple choice - tune, pace, rhythm, instruments, and participants' current mood. These options were so chosen that they were easily understandable by everyone. At the group level, the general participants were divided into two equal sets at random. Each group was assigned to rate only one type of excerpt (happy/sad)

under different conditions. First, each group rated their assigned excerpts through the *baseline* surveys. Each participant was presented with each excerpt at random. The total duration for the *baseline* survey for each participant was 15-20 minutes. Next, the *influenced* surveys were performed after a gap of 1 week from the *baseline* surveys. Each participant in each group was presented with 14 (out of 28) *influenced* sequence combinations of their assigned excerpts, selected at random, and perceived emotion ratings were collected. An interval of 2-3 minutes between each sequence was included. The participants were presented with interesting trivia about HCM, and simple mathematical puzzles during this time. The total duration for each participant was 45-60 minutes. After another gap of 1 week, the same procedure was followed with the rest of the 14 *influenced* sequence combinations.

4.2.7 Statistical Analyses

A mixed-method approach was taken to analyze the collected data. First, descriptive statistics on the ratings of the *baseline* and *influenced* surveys were reviewed. This provided an overview of the *target* and *non-target* emotions perceived for each music excerpt under uninfluenced and various influenced conditions. Next, Cronbach's alpha and ANOVA were employed to investigate the consensus between the participants and the clarity of perception of the *target* dominant emotion for each excerpt. IBM SPSS Statistics 22 was used for statistical analysis. Lastly, a normalized 10-dimensional probability distribution of perceived emotions for each excerpt was calculated to directly compare changes in perception of musical excerpts in *baseline* and *influenced* surveys. This was represented by the *perceived emotion probability vector*. Normalization allowed the creation of a uniform representation of various emotions with a normalized emotion vote for each excerpt, despite the freedom of multiple perceived emotion reports by each participant. The following normalization procedure [46] was used.

The *perceived emotion probability vector* of an excerpt denoted the probabilities of perceiving a predefined set of emotions during that excerpt, in terms of percentage. Here 10 dimensions were considered as there were only 10 emotion options available. For an excerpt e, a participant p could click on any emotion any num-

ber of times by the inherent design of the experiments. Assuming the participant clicked $n_x^{e,p}$ times on the emotion (x), where $x \in \{DK, OE, F, A, S, C, W, R, H, E\}$, the total number ($N^{e,p}$) of clicks by participant p for excerpt e was:

$$N^{e,p} = n_{DK}^{e,p} + n_{OE}^{e,p} + n_F^{e,p} + \cdots + n_E^{e,p} \tag{4.1}$$

Let each of these $n_x^{e,p}$ responses had intensities $I_x^{e,p}(1)$, $I_x^{e,p}(2)$, ..., $I_x^{e,p}(n_x^{e,p})$, where $1 \leq I_x^{e,p}(k) \leq 5$ and $1 \leq k \leq n_x^{e,p}$. The average intensity for each emotion ($\bar{I}_x^{e,p}$) was calculated as:

$$\bar{I}_x^{e,p} = \frac{\sum_{k=1}^{n_x^{e,p}} I_x^{e,p}(k)}{n_x^{e,p}}. \tag{4.2}$$

The normalization procedure considered two measures. First, the number of hits on a particular emotion, with respect to total hits for that excerpt by a participant was considered. Thus each participant got one normalized vote regarding an emotion in the excerpt, despite using the freedom to click that emotion as many times as the participant wanted. This measure normalized the votes of an over-active participant who reported many times with that of one who reported moderately. Thus, higher values of $\left(\frac{n_x^{e,p}}{N^{e,p}}\right)$ – where $0 \leq \left(\frac{n_x^{e,p}}{N^{e,p}}\right) \leq 1$ – indicated that the participant p perceived the emotion x in the excerpt e more than any other emotion. It also gave an indication of the frequency of the emotion perceived by the participant. Second, the intensity average of a particular emotion was considered. This provided a measure of how intense the perception of a particular emotion was. A high $\bar{I}_x^{e,p}$ value indicated elevated perceived intensity of emotion x over the excerpt e, of participant p. Therefore, we defined the weight of perceived emotion x, in excerpt e, for participant p as:

$$Wt_x^{e,p} = \left(\frac{n_x^{e,p}}{N^{e,p}}\right) * \bar{I}_x^{e,p} = \frac{\sum_{k=1}^{n_x^{e,p}} \bar{I}_x^{e,p}}{N^{e,p}} \tag{4.3}$$

The weight of each emotion x, for each excerpt e, rated by the participant p was calculated from equation 4.3. Lastly, for each excerpt, the averaged weights for each perceived emotion over all participants of a particular survey were used to derive the ultimate emotion distribution of that excerpt in terms of weights. This could then easily be converted into percentages to obtain a 10-dimensional vector

Table 4.3: Baseline (no influence) Survey Results: Means and standard deviations of target and non-target emotions, participant consistencies (Cronbach's α), repeated measures ANOVA results for each excerpt (η^2 for effect sizes, $p < 0.05$)

Excerpt #	Type	Target Emotion	Non-Target Emotion	α	η^2
1	Happy	3.26(1.06)	1.24(1.49)	0.87	0.72
2	Happy	3.46(1.26)	1.45(1.43)	0.85	0.81
3	Happy	3.69(1.14)	1.44(1.68)	0.84	0.86
4	Happy	3.00(1.41)	1.40(1.63)	0.85	0.79
5	Sad	3.15(1.39)	1.09(1.37)	0.93	0.70
6	Sad	3.95(1.18)	0.93(1.36)	0.94	0.83
7	Sad	3.34(1.37)	1.21(1.47)	0.88	0.75
8	Sad	3.74(1.33)	1.07(1.37)	0.92	0.78

containing the probability distribution of 10 emotion options that was perceived by the participants for a particular excerpt. For eg., let for an excerpt, the normalized probability distribution was calculated to be [OE%=0, DK%=0, F%=1.57, A%=1.98, S%=4.95, C%=21.15, W%=15.61, R%=10.87, H%=24.31, E%=19.56]. This meant the excerpt was dominantly perceived as *happy*, secondarily as *calm*, and so on. The probabilities of perceiving this particular excerpt as any of the 10 emotion options were thus expressed as percentage values. These perceived emotion probability vectors were used for comparing changes in emotion perception under *baseline* and *influenced* conditions.

4.3 Results

4.3.1 Perceived Emotions in Baseline Surveys

Qualitative analysis of the prominent reasons for the participant's choice of perceived emotions yielded that in most dominantly happy excerpts, 55%-70% participants reported *musical instrument*. 35%-50% of the participants reported *rhythm* and *tempo*and 25%-45% reported the *tune*. For dominantly sad excerpts, the most prominent reason was the *tune*, reported by 45%-70% of raters This was closely followed by the slow pace (*tempo*)(38%-45% of raters) of the excerpt. This gave an idea of which musical features triggered the perception of happiness or sad-

Table 4.4: Influenced survey results under happy and sad influences: Means and standard deviations of target and non-target emotions, Participant consistencies (Cronbach's α), and repeated measures ANOVA results (η^2 for effect sizes, $p < 0.05$)

Ex#	Target Emotion		Non-Target Emotion		Cronbach's α	
	Influenced by_Happy	Influenced by_Sad	Influenced by_Happy	Influenced by_Sad	Influenced by_Happy	Influenced by_Sad
1	1.92 (1.80)	2.43 (1.81)	0.56 (1.27)	0.22 (0.92)	0.61	0.63
2	1.67 (1.75)	3.75 (1.51)	1.19 (1.56)	1.50 (1.68)	0.65	0.69
3	1.33 (1.89)	2.24 (2.52)	0.52 (1.24)	0.42 (1.24)	0.77	0.81
4	1.53 (1.79)	2.59 (1.68)	0.58 (1.26)	0.32 (0.98)	0.80	0.75
5	2.28 (1.47)	3.13 (1.13)	0.83 (1.45)	0.22 (0.53)	0.61	0.68
6	1.25 (2.46)	3.28 (1.54)	0.93 (2.50)	1.80 (1.85)	0.60	0.64
7	1.49 (2.98)	3.77 (1.69)	0.31 (0.92)	0.41 (1.09)	0.82	0.80
8	1.74 (1.53)	4.16 (1.88)	0.25 (0.85)	0.53 (1.25)	0.79	0.93

ness in HCM for general listeners. Quantitative results of the *baseline* surveys are reported in table 4.3. The means and standard deviations of ratings for *target* and *non-target* perceived emotions of each excerpt were calculated (table 4.3, 3^{rd} and 4^{th} columns). It was observed that for most excerpts, the mean ratings for the *target* (dominant) emotions were high (≥ 3.00), and that of the *non-target* emotions were low (≤ 1.50). This indicated that the target dominant emotions were perceived well by the participants in the chosen excerpts. This was in accordance with the expert panel's decision on each excerpt's dominant emotion. The standard deviations were less (≤ 1.40) for target emotions, indicating consistently high ratings. But for non-target emotions, standard deviations fluctuated due to larger variations in the reported ratings. Some emotions congruent to the target emotion received higher ratings, and others received lower ratings.

Cronbach's alpha was employed to measure rating consistency between participants (table 4.3, 5^{th} column). Most excerpts scored high consistencies (≥ 0.85). This indicated high agreement between participants about the perceived emotions of individual excerpts in the uninfluenced *baseline* surveys.

Further, ANOVA was used to verify whether the *target* emotions were clearly evident in the *baseline* surveys for the dominantly happy and sad excerpts. *Emotion* was considered the independent variable and the perceived ratings were the

dependent variable. In this one-way repeated measures ANOVA, *Emotion* was considered as the within-subject factor, with ten levels (OE, DK, ... E). For individual excerpts, the ratings of each emotion were compared, yielding significant main effects for both happy and sad emotion targets with large effect sizes ($\geq$0.70, $p < 0.05, df = 9,711$), reported in table 4.3, 6^{th} column. Generally, the ANOVA results displayed strong discrimination between the target and non-target emotion categories, confirming pronounced perception of the *target* emotion.

Lastly, the data normalization procedure was used to obtain the probability distributions of perceived emotions along with the *dominant* and *secondary* emotions in each excerpt. For all the excerpts, the target dominant emotion was perceived with the highest probability. It was observed that the dominantly *happy* excerpts had calmness or excitement as the secondary emotion, denoting the *nuance* of the happiness, and generally the other associated sub-dominant emotions had positive valence (wonder/romantic/exciting). Thus a detailed idea of what kind of happiness (nuance) the excerpt evoked could be achieved. All the dominantly *sad* excerpts had *calmness* as the secondary emotion. The probabilities of perceiving excitement, happiness, or wonder were minimum, and generally, the associated sub-dominant emotions had negative valence (fear/anger). This detailed analysis was important as it made it easier to study the changes in perceived emotions in the *influenced* surveys using these. Comparisons between the *baseline* and *influenced* distributions were depicted in figure 4.2 and figure 4.3 and discussed later.

4.3.2 Perceived Emotions in Influenced Surveys

Results of the *influenced* surveys are reported in table 4.4. The *target* emotion mean ratings were found to be greater than the *non-target* emotions (3^{rd} - 6^{th} columns), indicating a clear perception of *target* emotions in the *influenced* surveys as well. These were then compared with the results of *baseline* surveys (figure 4.1). It was observed that in general, the mean ratings for the *target* emotions were higher in the *baseline* surveys, with the least standard deviation. These were least in the *influenced_by_happy* surveys, with high standard deviations, indicating larger variation in the reported ratings and notable impact on dominant emotion

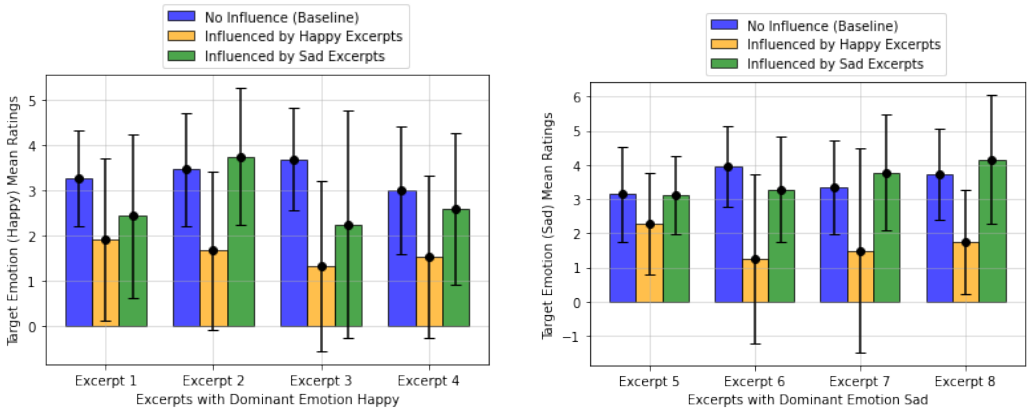
Table 4.5: Influenced Survey Results: Repeated measures ANOVA results for each excerpt (η^2 for effect sizes, $p < 0.001$). Two main effects of Emotion (df = 9,711) and Influence (df = 1,79) and the Interaction effect (df = 9,711) are reported.

Excerpt	Influence_by_Happy			Influenced_by_Sad		
#	Emotion	Influence	Interaction	Emotion	Influence	Interaction
1	0.68	0.214	0.433	0.71	0.217	0.377
2	0.65	0.207	0.418	0.671	0.212	0.361
3	0.66	0.261	0.493	0.683	0.238	0.431
4	0.69	0.237	0.411	0.726	0.219	0.265
5	0.61	0.22	0.318	0.632	0.21	0.32
6	0.59	0.281	0.406	0.67	0.27	0.41
7	0.76	0.249	0.291	0.752	0.268	0.33
8	0.62	0.219	0.412	0.638	0.21	0.40

perception. Interestingly, compared to *baseline* surveys, the mean ratings for the *target* emotions in *influenced_by_sad* surveys were greater in some cases (# 2,5,7,8) and lower in others (# 1,3,4,6). These results indicated the possibility of a variable impact on dominant emotion perception when influenced by sad excerpts.

Cronbach's alpha was employed to measure rating consistency (table 4.4, 7^{th}, 8^{th} columns). Among the two *influenced* surveys, a higher agreement was noted in the *influenced_by_sad* surveys. Compared to *baseline* surveys (table 4.3, 5^{th} column), most excerpts scored lower consistencies in the *influenced* surveys, indicating comparatively lower agreement between participants regarding the perceived emotions. Thus, the inter-participant agreement was observed to be affected both due to the presence and type of a preceding excerpt.

Further, a within-subjects two-way repeated measures ANOVA was used to explore possible effects of *intra-contextual influence* on the perceived emotion ratings of excerpts. *Emotion* was considered as one factor, with 10 levels (OE, DK, ..., E). *Influence* was second factor, with the two levels, *presence* of a particular type of influence or *absence* of it (baseline) . The ratings were the dependent variable. For each excerpt, two such repeated measures ANOVA were done, one for each type of influence (table 4.5). As observed from the results, these two factors taken together gave significant main effects for both target emotions and influence (table 4.5), under both *influenced_by_happy* and *influenced_by_sad* conditions. There were also significant interaction effects between these two factors,

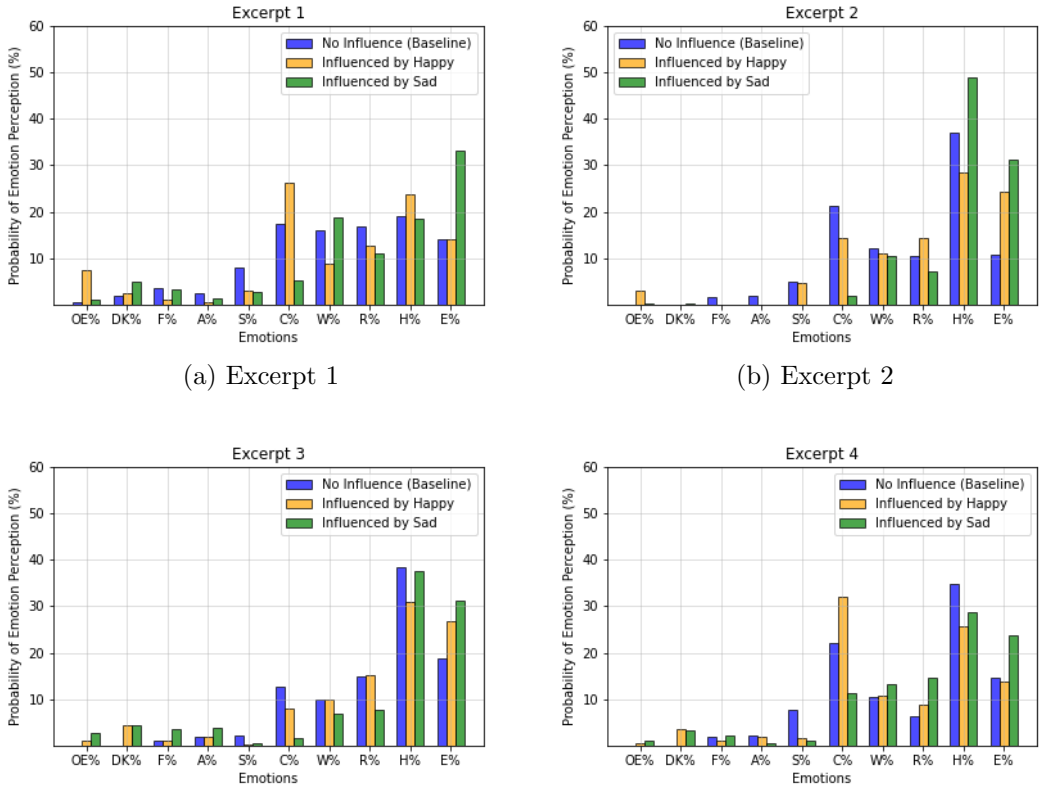


(a) Happy Excerpts (b) Sad Excerpts

Figure 4.1: Comparison of target dominant emotion mean ratings and standard deviations in (a) happy and (b) sad excerpts, under three different conditions - no influence, influenced by happy, and influenced by sad excerpts. The X-axis and Y-axis in each sub-figure represent the example excerpts and the mean ratings. The three bars in the histogram for each excerpt represent the three conditions.

indicating a notable difference in perceived emotion ratings resulting from *intra-contextual influence*.

Lastly, the data normalization procedure was used. For each excerpt, the probability distributions of perceived emotions in *baseline* and *influenced* conditions were compared in figure 4.2 and figure 4.3. The results indicated that notable variations in probabilities of perceived emotions occur due to *intra-contextual influences*. Three types of variations were observed: a) Change in both dominant and secondary perceived emotions - eg. excerpt #1, was perceived to be dominantly happy and secondarily calm in *baseline* survey. The perceptions changed to dominantly calm and secondarily happy when influenced by other happy excerpts. It was perceived as dominantly exciting and secondarily wondrous/happy when influenced by sad excerpts. b) Change in secondary emotion (*nuance*) only - eg. excerpt #7 was originally perceived as secondarily calm in *baseline* survey. This perception changed to *fear* under the influence of happy excerpts. The *dominant* emotion remained unchanged. c) Change in probability of perception of dominant and/or secondary emotions - eg. excerpt #3, which was perceived to be happy and exciting under all conditions, but the probabilities varied. Thus, in *baseline* and



(a) Excerpt 1

(b) Excerpt 2

(c) Excerpt 3

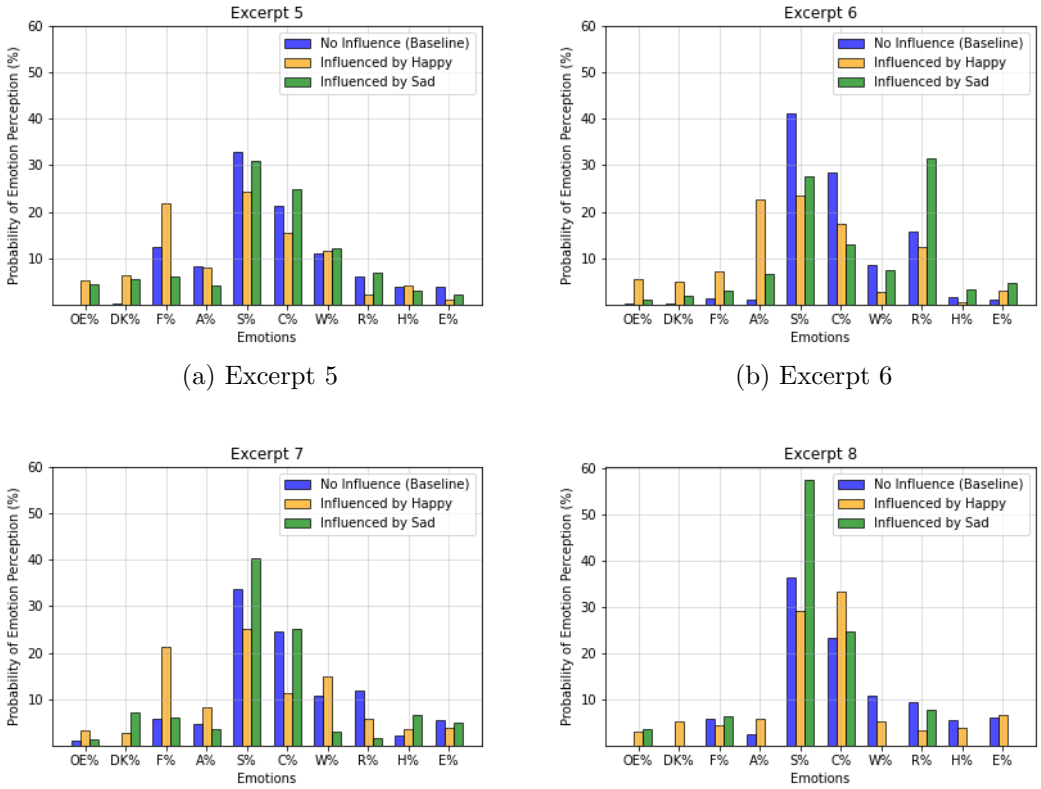
(d) Excerpt 4

Figure 4.2: Variations in probability distributions of perceived emotions in the happy excerpts, under three different conditions - no influence, influenced by happy, and influenced by sad excerpts. The X-axis and Y-axis of each sub-figure represent the emotions and the probability of perceiving them.

influenced_by_happy conditions, a *tertiary* presence of romance could be perceived. This became almost negligible in *influenced_by_sad* condition, as excitement became more prominent. The confusion indicators OE and DK were also observed to be reported more in the *infleunced* surveys.

4.4 Observations and Discussion

The present study aimed to systematically understand possible effects the change in antecedent *immediate intrinsic musical context* has on the perceived emotions



(a) Excerpt 5

(b) Excerpt 6

(c) Excerpt 7

(d) Excerpt 8

Figure 4.3: Variations in probability distributions of perceived emotions in the sad excerpts, under three different conditions - no influence, influenced by happy, and influenced by sad excerpts. The X-axis and Y-axis of each sub-figure represent the emotions and the probability of perceiving them.

of consequent excerpts, specifically in Hindustani classical music. *Baseline* and *influenced* surveys were set up and descriptive statistics, rater consensus, ANOVA, and probability distributions of perceived emotions were used to compare perceived emotions under different conditions. In the *baseline* surveys, the *dominant* emotions reported for individual excerpts by the general participants followed similar trends as in previous HCM-related studies. For instance, excerpts from *ragas* Hamsdhwani, Adana, and Desh were perceived as happy and calm, and Marwa and KR Asavari were perceived as sad and calm. This is in accordance with the findings of Mathur et al. [153], who stated that distinct *ragas* elicit dis-

tinct emotional responses [158]. The prominent reports of calmness as *secondary* emotion in both happy and sad excerpts were in consensus with the findings of [83], who stated that the highest-scoring music emotion factor for Indians was "Peaceful-Transcendence" (similar to calmness). From the qualitative analysis of the prominent reasons of perceived emotions, it was observed that for both dominantly happy and sad excerpts, the *tempo* played an important role in emotion perception, apart from the *tune* (raga). This is in accordance with the findings of Mathur et al. [153], who stated that the highest experienced emotions were calm and sad for the arrhythmic (slow tempo) phase of raga renditions and happy, tense, and longing for the rhythmic phases. This indicates that our surveys have run successfully and the expert-rated music stimuli selection has also been appropriate. Next, we explore the main research questions of this work.

RQ1: Does intra-contextual influence play any role in music emotion perception? The results of target emotion mean ratings (figure 4.1), inter-participant consensus, and ANOVA (table 4.5) clearly suggest the effect of *intra-contextual influence* on music-perceived emotions. We observe that *intra-contextual influence* is different from the following phenomena - *behavioral* [183] or *emotional contagion* [10], [184], *emotion induction* [24], and *emotion priming* [25], [99]. Behavioral contagion [183] refers to the inclination of a person to copy certain behaviors of others. It is essentially a human-human process. Emotional contagion [10], [51] is a process whereby an emotion is induced [24] by music, because the listener perceives the emotional expression of the music and *mimics* this expression internally. It is essentially a music-human process. Evidently, these are different than the proposed concept of *intra-contextual influence*, which is inherently a music-music process affecting the emotion perception stage. In affective priming [25], [99] the stimuli are used to evoke targeted emotions in the participants. The stimuli are also manipulated to evoke necessary affect. Also, in affective priming studies involving music, the prime, and the target are different media [99], [102], [103]. The music excerpt is either a prime or a target. On the other hand, the proposed phenomenon - *intra-contextual influence* - refers to *how* the listener's emotion perception of a music excerpt is *influenced* by the *context* provided by the preceding music excerpt. The *influencer* and the *influenced* media are both music excerpts. Listeners are also not subjected to intentional emotion induction

and are specifically informed that the study is interested only in their perceived emotions. In emotion perception, the induction (if we may use the term at all) is happening at the level of cognition only. Thus, in *intra-contextual influence* we are interested in the cognitive facet of emotions perceived.

Next, we discuss our observations with respect to the second research question. **RQ2**: Can the effects of *intra-contextual influence* on the perception of musical emotions be generalized across music excerpts grouped according to their dominant emotions?

The results of target emotion mean ratings (figure 4.1), and the probability distributions of perceived emotions in *baseline* and *influenced* conditions (figures 4.2, and 4.3) give an indication of the patterns of *intra-contextual influence* in happy and sad excerpts. Here as well, we can observe the differences in the effects of affective priming and intra-contextual influences. In priming, the evaluation of the target as pleasant or unpleasant is reduced when the emotional valence of the prime and target are congruent rather than incongruent [25]. In other words, the perception of happiness in a happy excerpt will decrease and increase, when it follows happy and sad excerpts respectively. The perception of sadness in a sad excerpt will decrease and increase, when it follows sad and happy excerpts respectively. This pattern is not followed in *intra-contextual influences*, as is evident from figures 4.2 and 4.3. Though the perception variations of target emotion (happiness) conform to that of affective priming for the case of *influencer* happy and *target* happy excerpts (happiness ratings diminished), no such specific trends are observed for any of the other three combinations of *influencer-influenced* excerpt types. In general, the probability of perception of happiness diminishes when happy excerpts are influenced by sad excerpts. The probability of perceiving excitement increases. The probability of perceiving sadness in sad excerpts diminishes slightly when influenced by happy excerpts. The probability of changing the nuance of sadness from calmness to other congruent emotions (fear/anger) is higher. The probability of perceiving sadness in sad excerpts is generally increased when influenced by other sad excerpts. None of these results are in agreement with the concept of affective priming. Irrespective of the type of the *influenced* excerpt, three types of variations are observed: a) Change in dominant and secondary perceived emotions, b) Change in secondary emotion (nuance) only, and c) Change

in probability of perception of dominant and/or secondary emotions. The second and third variations are observed to be more prevalent than the first. It stands to reason that changing the dominant emotion perception of any excerpt might be much harder through *intra-contextual* influence than changing the *nuance*. Since emotions are perceived as a composite, it might be much easier to subtly manipulate the nuance. Also, some excerpts might be more susceptible to be perceived differently under different conditions. In fact, both factors might interact to play roles in how a particular musical excerpt was perceived: a) intra-contextual influence, and b) excerpt's inherent properties. So, we conclude that *immediate intrinsic context* plays a key role in music-emotion perception, with observable trends for dominantly happy and sad excerpts. This might have deep relevance for music appreciation and aesthetics, as interpretation and appreciation of music may be strongly influenced by varying antecedents of recurrent phrases or themes, thus modifying their impact/experience without listeners even being aware of it.

To summarize the main contributions of this, the concepts of *immediate intrinsic musical context*, and *intra-contextual influence* are introduced and their effect on the perception of music emotions are explored. The current findings highlight that these have a significant effect on how subsequent music is perceived, with observable patterns for dominantly happy and sad music excerpts. We compare the observed results with the concepts of affective priming, emotion contagion, and induction, presenting the possible similarities and differences among them. We also initiate a structured investigation on the perception of Hindustani classical music excerpts, with Indian aesthetic philosophy in mind. The choice of emotions in our survey design conforms to the *Rasa* concept [22] of emotions in Indian aesthetics. The findings of the current study can be used to further explore how perceived emotions are modified when the dominant emotions of both *influencer* and *target* excerpts are the same. This is particularly interesting, as the use of emotionally congruent excerpts with different intensities might not necessarily lead to the emotional crescendo generally desired, with implications for music generation. The present work might also be used to extend the findings to excerpts of other dominant emotions, enhancing our understanding of why we perceive music as we do.

CHAPTER 5

Explaining Perceived Emotion Predictions in Music: An Attentive Approach

Dynamic prediction of perceived emotions in music is a challenging problem with interesting applications. Utilization of relevant context in an audio sequence is essential for effective prediction. Existing methods have used LSTMs with modest success. In this work, we describe three attentive LSTM-based approaches for dynamic emotion prediction from music clips. We validate our models through extensive experimentation on a standard dataset annotated with arousal-valence values in continuous time and choose the best performer. We find that the LSTM-based attention models perform better than the state-of-the-art transformers for the dynamic emotion prediction task, both in terms of R^2 and Kendall-τ metrics. We explore individual smaller feature sets in search of a more effective one and to understand how different features contribute to perceived emotion. The spectral features are found to perform at par with the generic ComPare feature set [120]. Through attention map analysis we visualize how attention is distributed over music clips' frames for emotion prediction. It is observed that the models attend to frames that contribute to changes in reported arousal-valence values and chroma to

produce better emotion predictions, effectively capturing long-term dependencies.

5.1 Introduction

Automatic determination of perceived emotion in music is an active and major area of focus for the music information retrieval (MIR) community. The aim of the dynamic perceived emotion prediction task is to output a sequence of time-synchronized arousal-valence labels when a music clip is given as input. It finds varied applications in the domains of personalized and/or generalized music recommendations, organizing music databases, automatic music creation, mood-based music search, etc. This task is challenging because: 1) perceived emotion might depend on the inherent relationship between different frames of music, distributed over time, and 2) emotion perception is inherently subjective in nature, highly contextual, and personal. Thus, it is understandable that the emotions related to music are a time-continuous process, where the context of the sequential music frames plays an immense role in the associated emotion. Relating this to the machine learning perspective, one can discern the need for context-sensitive models like recurrent neural networks (RNNs) for the task at hand. In this study, we use the attention mechanism with deep RNN-LSTMs (Long Short Term Memory) and the Transformer [185], to predict the perceived emotion in each defined time frame of music continuously. We compare our approach with recent works [112] using only LSTM. We also attempt to understand the importance of types of features contributing to dynamic perceived emotion. Lastly, attention is visualized with the help of attention map analysis. The following are the major contributions of this work:
1) The LSTM-based attention models are found to perform better than the state-of-the-art Transformers for the dynamic emotion prediction task. 2) Spectral features are found to perform at par with the generic ComPare feature set [120]. 3) Attention maps are interpreted to observe that the attention models are able to focus on relevant music frames for the dynamic emotion prediction task.

This chapter is organized as follows. In section 5.2, relevant literature regarding music emotion recognition and attention is reviewed. Section 5.3 provides details of the attention-based models and Transformer used in this work. All the

experiments carried out and the observations are reported in section 5.4. Finally, the conclusions drawn from the present study are detailed in section 5.5.

5.2 Related Work

5.2.1 Music Emotion Recognition

In the past, most music emotion prediction systems used features of timbre, pitch, MFCCs, and/or lyrics and applied them to classifiers like SVMs [127]. Current state-of-the-art methods for music emotion prediction are mostly based on deep neural networks like RNN-LSTMs. Coutinho et al. [136] proposed the use of this model for this task. Weninger et al. [112, 137, 53] used RNN-LSTM networks successfully to perform continuous time music emotion regression, using a modified cost function, on the *1000 Songs for Emotional Analysis of Music* dataset [3]. Giamusso et al. [138] used neural networks to predict playlist emotions based on lyrics. Fan et al. [139] performed ranking-based emotion recognition from experimental music. Delbouys et al. [55] used LSTM and ConvNet models on the Million Song Dataset [140] for audio and lyrics-based bimodal music emotion detection.

5.2.2 Emotion Representation

Over the years, Discrete and Dimensional models of emotion representation have been used in MIR. Studies using discrete models either tag their musical data with single [42] or cluster [43] of simple tags. In dimensional models like Russel's Circumplex model [2], emotion is mapped into a 2-D plane, spanned by two axes denoting *arousal* and *valence*. Using this well-known and satisfyingly exhaustive emotion representation, the problem of emotion recognition/prediction is turned into a two-dimensional regression problem [130].

5.2.3 Attention in MIR tasks

Recently, attention mechanism and Transformer models have found application in a wide range of MIR tasks, with success. Balke et al. [186] used a soft-attention

mechanism on the input of synthesized piano data for audio sheet music retrieval. Their results indicate that attention increases the robustness of the retrieval system by focusing on different parts of the input representation based on the tempo of the audio. The improved results led them to argue for the potential of attention models as a very general tool for many MIR tasks. Gururani et al. [187] explored an attention mechanism for handling weakly labeled data for multi-label instrument recognition. Their results show that incorporating attention leads to overall improvement in classification accuracy metrics and enables models to *attend to* specific time segments in the audio relevant to each instrument label leading to interpretable results. Donahue et al. [188] used the Transformer architecture to improve performance for the task of generating multi-instrumental music scores. Chen et al. [189] proposed the Harmony Transformer, a multi-task music harmony analysis model aiming to improve chord recognition. Park et al. [190] utilized a bi-directional Transformer for chord recognition (BTC) which showed competitive performance. Through attention map, they visualized how attention was performed, and it was observed that the model was able to divide segments of chords by utilizing the adaptive receptive field of the attention mechanism and capture long-term dependencies. These and other works have explored various feature sets like CQT (in [190]), Chroma (in [189]), along with other standard feature sets [120] (in [112]). These recent successes in varied MIR tasks in terms of model accuracy and interpretability, motivated us to apply the same in the music emotion regression task. To the best of our knowledge, neither attention models nor Transformers have been applied before to the task under examination.

5.3 Attention Based Models for Emotion Prediction in Music

5.3.1 Attention Model (AT)

In the past, the traditional LSTM-RNN approach has provided good results in music emotion regression [112]. In this work, we propose the use of the attention mechanism for dynamic emotion prediction in music. According to the *attention*

model [143], to compute each output of an encoder-decoder architecture, a distinct *context vector* is used, which is a function of all the hidden states at the encoder side and not just the last one. The encoder encodes the input into a set of hidden states and attention is applied to them to produce target arousal and valence values over fixed length segments or time frames of the music audio signal. The encoder reads the input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, which is a sequence of vectors, and produces the hidden states $(h_1, h_2, \ldots, h_T)$, using some RNN approach. In this work, LSTM is used. In traditional attention mechanism [143], the whole set of hidden states $(h_1, h_2, \ldots, h_T)$ are available to compute the context vectors. Each time, the context vector c_i is calculated as a weighted sum of all the hidden states. Let the output be $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. For the current problem, $\mathbf{y}$ can be defined as a set of arousal or valence values associated with each music time frame. The t^{th} output, y_t, will be a function $\mathbf{g}()$ of a) the present hidden state h_t, b) the previous output y_{t-1}, c) the unique context vector c_t, as given by equation 5.1.

$$p(y_t|y_1, y_2, \ldots, y_{t-1}, \mathbf{x}) = g(h_t, y_{t-1}, c_t) \tag{5.1}$$

The unique context vector c_t depends on the sequence of annotations $(h_1, h_2, \ldots, h_T)$, and is computed as a weighted sum of these annotations h_j, as given in equation 5.2.

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{5.2}$$

So, the model at time t, *attends* to each h_j corresponding to each of the inputs, with a weight of α_{tj}. To obtain each weight α_{tj} for each output y_t, the alignment between the corresponding h_t and each of h_j need to be calculated, where $1 \leq j \leq T$. So, the alignment model, when attending to h_j, is given by equation 5.3.

$$e_{tj} = a(h_{t-1}, h_j), 1 \leq j \leq (t-1) \tag{5.3}$$

This alignment is the measure of how well the inputs around position j and the output at position t match. Then, each of these scores e_{tj} is used to calculate the

attention weights for each h_j as given in equation 5.4.

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T} exp(e_{tk})} \tag{5.4}$$

So, for each output, the context vector will *attend* or focus on those parts of the entire input sequence, which are more relevant for that particular output, by assigning higher weights to the associated encoder-side hidden states, using an *alignment* model. These models are referred to as the AT models from hereon. The naming convention of the models is the acronym AT for attention, followed by the hidden layer dimensions.

5.3.2 Backward Attention Model (BAT)

A modified form of the traditional attention mechanism [143] is also used in the current work, called Backward Attention (BAT) models. In these models, for emotion prediction at each t^{th} time frame, attention is distributed only among h_k hidden states, where, $1 \leq k \leq (t-1)$.

5.3.3 Transformers

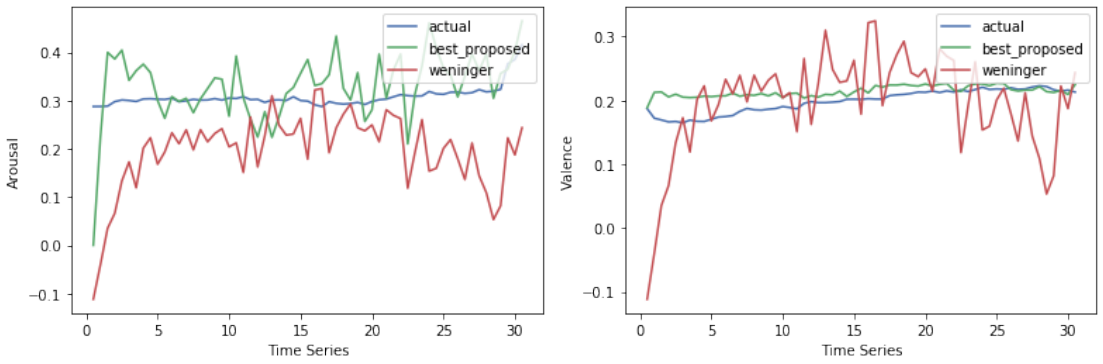
The transformer architecture as proposed by Vaswani et. al. [185] is used in this work, with changes in the number of encoder side layers, as appropriate for the experiments. Attention is calculated as in equation 5.5.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5.5}$$

where, Q, K, and V are matrices representing the set of queries, keys, and values respectively and d_k is the key dimension.

Table 5.1: Model Selection for Dynamic Arousal Prediction

Model	Parameter Search (Layer Size)	Best Model	R^2_A	$\overline{\tau}_A$	MAE_A
Baseline [112]	400	-	0.60	0.14	0.11
LSTM (Single Layer)	128, 300, 400, 512, 700, 1024, 2048	1024	0.73	0.12	0.12
LSTM (Multi Layer)	(700_128), (700_400), (2048_1024), (2048_1024_700)	(700_128)	0.69	0.20	0.12
AT (Single Layer)	32, 64, 128, 300, 400, 512, 700, 1024, 2048	300	0.75	0.15	0.13
AT (Multi Layer)	(300_128), (400_128), (1024_400), (2048_1024), (2048_1024_512)	(2048_1024)	**0.78**	0.24	0.11
BAT (Single Layer)	400, 1024, 2048	1024	0.55	0.04	0.12
BAT (Multi Layer)	(300_128), (400_128), (1024_400), (1024_512), (2048_1024), (2048_1024_512)	(2048_1024)	0.58	0.06	0.12
Transformer	1-Layer, 2-Layer, 4-Layer	2-Layer	0.64	0.61	0.27



(a) Arousal Comparison (b) Valence Comparison

Figure 5.1: Dynamic Emotion Predictions for Clip 584

5.4 Experiments

5.4.1 Data Description and Experimental Setup

We use the *1000 Songs for Emotional Analysis of Music* dataset [3] for all experiments. Of the thousand clips,the dataset provides arousal and valence annotations for only 744 clips, which are used as *ground truth* values. According to the dataset manual [3], arousal-valence continuous annotations for each song (second 15-45), with 2Hz sampling frequency are available in the dataset. We define each non-overlapping 500ms of the clips as *one music frame*. Thus, the last 30s or the last

Table 5.2: Model Selection for Dynamic Valence Prediction

Model	Parameter Search (Layer Size)	Best Model	R^2_V	$\overline{\tau}_V$	MAE_V
Baseline [112]	400	-	0.29	0.08	0.16
LSTM (Single Layer)	128, 300, 400, 512, 700, 1024, 2048	700	0.39	0.10	0.15
LSTM (Multi Layer)	(700_128), (700_400), (2048_1024), (2048_1024_700)	(2048_1024)	0.29	0.17	0.15
AT (Single Layer)	32, 64, 128, 300, 400, 2048 512, 700, 1024, 2048	400	**0.53**	0.08	0.16
AT (Multi Layer)	(300_128), (400_128), (1024_400), (2048_1024), (2048_1024_512)	(300_128)	0.51	0.04	0.16
BAT (Single Layer)	400, 1024, 2048	2048	0.16	0.13	0.15
BAT (Multi Layer)	(300_128), (400_128), (1024_400), (1024_512), (2048_1024), (2048_1024_512)	(400_128)	0.21	0.16	0.14
Transformer	1-Layer, 2-Layer, 4-Layer	1-Layer	0.12	0.11	0.10

61 frames of each clip are used for this work, since only those 61 emotion (arousal-valence) tags are available. 10-fold cross validation was used on the training and test sets. We used the Mean squared error (MSE) as the loss function. RMSProp, with the default learning rate of 0.001 was used for optimizing the loss with a batch size of 20, and a maximum of 50 epochs. An early stopping strategy is also used, if the validation error shows no improvement over 10^{-4} after 5 epochs, the processing is stopped. Sequences are presented in random order during training. All hyper-parameters not explicitly mentioned here are left to their default values as in Tensorflow 1.14. The feature sets used for different experiments are described below.

5.4.1.1 ComPare Feature Set

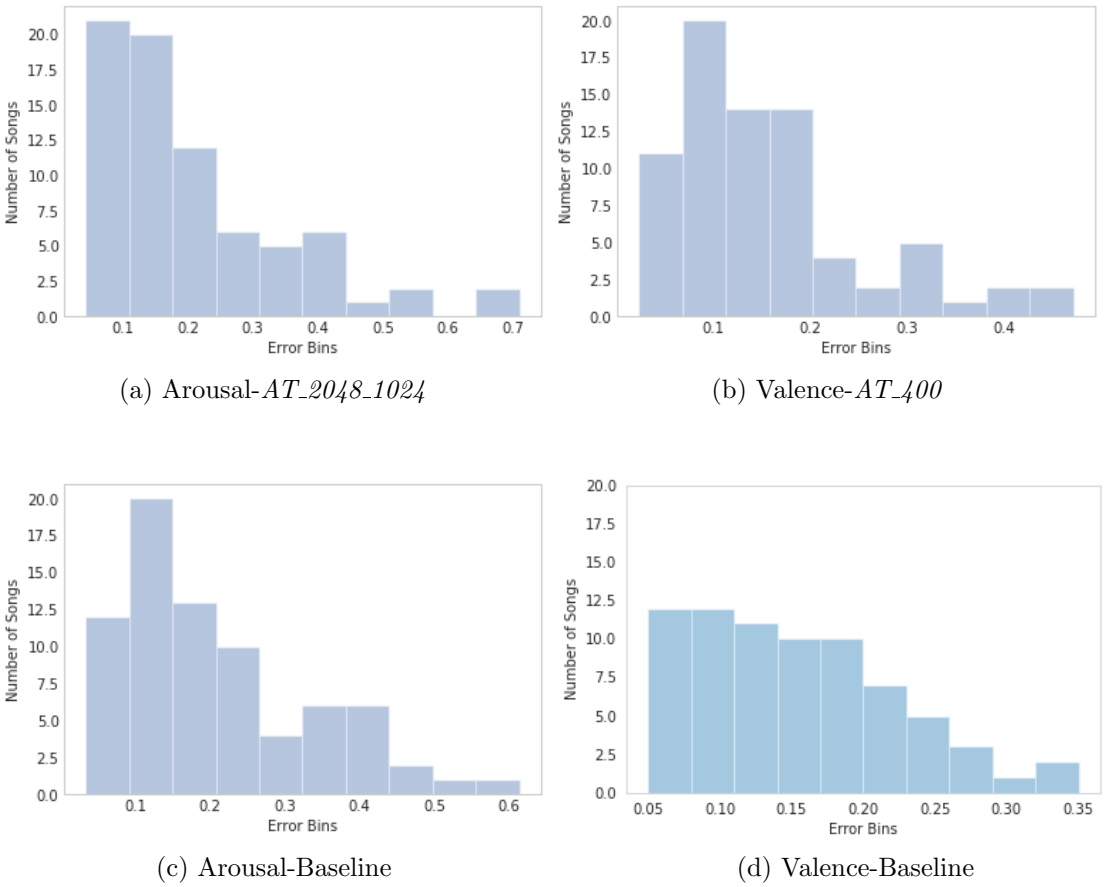
The 2013 Computational Paralinguistics Evaluation (ComParE) tasks feature set [120], containing 6670 features is used for all experiments in sections 5.4.2 and 5.4.4. TUM's open-source *openSMILE* feature extractor [125] is used to extract the ComParE feature set for each frame of each clip. Standard normalization was performed on the extracted feature values before the experiments. So, each clip is characterized by 61 feature vectors, each of size 6670.

5.4.1.2 Other Feature Sets

In experiments reported in section 5.4.3, subsets of the Compare feature set [120] and some other features are explored. These features extracted using Librosa [126] are detailed here. The *Chroma(STFT+CQT)* features [126] consists of chroma values derived using both STFT analysis and constant-Q transform (CQT) analysis implementations. The *CQT on Audio clip* features [126] are derived from the core Spectrogram operations of Librosa [126] suitable for pitch-based signal analysis. The *Spectral Features* [126] denote the distributions of energy over a set of frequencies and are very important in many MIR analysis techniques. These consist of Chroma(24), CENs (12) MFCC (20), RMS (1), Mel-scaled spectrogram (128), spectral centroid (1), spectral bandwidth (1), spectral contrast (7), spectral flatness (1), spectral roll-off (1), zero crossing rate (1). All clips were re-sampled to 44100 Hz before feature extraction. All features were extracted for non-overlapping frames of 500 ms each, corresponding to the available arousal-valence labels of the dataset.

5.4.1.3 Metrics

The metrics used for reporting the results are Coefficient of determination (R^2), average Kendall's τ per song ($\overline{\tau}$)and mean absolute error (MAE). The determination coefficient (R^2) is a key output of regression analysis, which provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. It can vary between 0 and 1. If a data set has n values marked $(y_1 \dots y_n)$, and each associated with a predicted value $(f_1 \dots f_n)$. So, R^2 is defined as $R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$ where, $SS_{res} = \sum_i (y_i - f_i)^2$ and $SS_{tot} = \sum_i (y_i - \overline{y})^2$, given $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Kendall's τ per song ($\overline{\tau}$) is a measure of how well the emotional profile of each song is captured by the regressor, as opposed to the overall correlation. It measures the correspondence between two rankings. Values close to 1 indicate strong agreement, and values close to -1 indicate strong disagreement. It is defined $\overline{\tau} = \frac{P-Q}{\sqrt{(P+Q+T)*(P+Q+U)}}$ where, P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in target set $(y_1 \dots y_n)$, and U the number of ties only in predicted set $(f_1 \dots f_n)$. The mean absolute error (MAE) is given for reference. In



(a) Arousal-*AT_2048_1024*

(b) Valence-*AT_400*

(c) Arousal-Baseline

(d) Valence-Baseline

Figure 5.2: Emotion Error Histograms over Validation Set

the next section, we report the results of applying the proposed model for dynamic music emotion regression.

Baseline: It has been shown by Weninger et. al. [112, 137] that LSTMs can be used to produce good performance in emotion prediction, using the ComParE feature set. We try to reproduce their results using single-layer LSTM-RNNs with a hidden layer size of 400 units. These results are considered as *Baseline* in this work and are reported in the "Baseline" annotated rows of tables 5.1 and 5.2 for arousal and valence respectively.

Table 5.3: Feature Sets for Arousal Prediction

Features Used	**# Features**	**Best Model**	$\boldsymbol{R_A^2}$	$\boldsymbol{\overline{\tau}_A}$	$\boldsymbol{MAE_A}$
Chroma(STFT+CQT)	24	AT_64	0.15	0.04	0.19
CQT on Audio clip	252	AT_64	0.45	0.06	0.17
Chroma+CQT	276	AT_64	0.57	0.07	0.14
Spectral Features	197	AT_64	**0.70**	0.03	0.12

5.4.2 Experiment 1: Model Selection

In the first set of experiments, we aim to find the best model for dynamic arousal and valence prediction, among the ones proposed in section 5.3. Accordingly, the models with attention (*AT*, *BAT*, *Transformers*) and without attention (*LSTM*) are executed with varying layer sizes and layer numbers. The findings for arousal and valence are reported in tables 5.1 and 5.2 respectively. For dynamic arousal prediction (table 5.1) using the ComPare feature set [120] (sec 5.4.1.1), the best result is obtained with the multi-layer attention model *AT_2048_1024*. A comparable result is obtained with the single-layer attention model *AT_300*. The best model for dynamic valence prediction (table 5.2) is found to be the single-layer attention model *AT_400*. A comparable result is also obtained with the multi-layer attention model *AT_300_128*.

The following are observed from this experiment: a) The best prediction performances reported in this section are better than that reported by the baseline methods (sec 5.4.1.3). b) Among all the experiments conducted, AT models fare best in dynamic arousal-valence prediction using the full ComPare feature set [120]. c) The best single and multi layer AT models' performances are comparable. d) Performance for arousal prediction (R_A^2 and $\overline{\tau}_A$) in general is much better than valence (R_V^2 and $\overline{\tau}_V$) - across all models tested. Though performance with respect to MAE are comparable.

In the following subsections, we demonstrate an illustrative example of dynamic emotion prediction using a clip chosen at random, followed by an error analysis of the predictions by the best-proposed models, over the validation set clips.

Table 5.4: Feature Sets for Valence Prediction

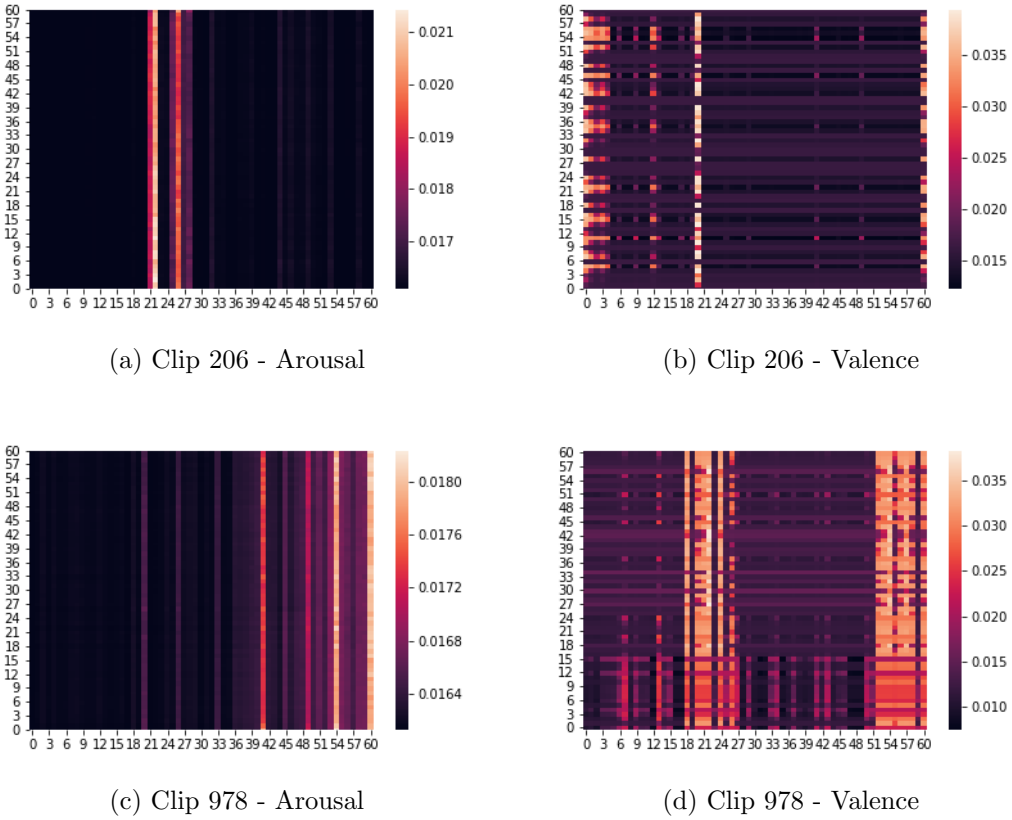
Features Used	**# Features**	**Best Model**	R_V^2	$\overline{\tau}_V$	MAE_V
Chroma(STFT+CQT)	24	AT_64	0.01	0.002	0.09
CQT on Audio clip	252	AT_64	0.07	0.01	0.17
Chroma+CQT	276	AT_64	0.17	0.06	0.14
Spectral Features	197	AT_128	**0.35**	0.07	0.16

5.4.2.1 Illustrative examples

In this section, we demonstrate an illustrative example of a dynamic emotion prediction pattern, with respect to ground truth (sec 5.4.1) and baseline (sec 5.4.1.3), using a clip chosen at random from the dataset [3]. The best models, *AT_2048_1024* for arousal and *AT_400* for valence, obtained in section 5.4.2 are used for dynamic (per 500 ms) arousal and valence prediction of music clip *584.mp3*. This is presented in 5.1. This clip is of the Folk genre, significantly upbeat, and has a positive valence. 5.1a and 5.1b denote the time varying arousal and valence predictions respectively. In the figures, X-axis denotes the time (in seconds), and the Y-axis denotes arousal and valence values respectively. It is seen that the proposed best models follow the pattern of reported emotions more closely than the baseline model.

5.4.2.2 Errors Analysis

In this section, we aim to observe patterns and biases in the best-proposed models' (sec 5.4.1) emotion predictions, with respect to the baseline (sec 5.4.1.3). The respective predictions are utilized to group the validation set clips into error bins for this study. These are shown as histograms in figure 5.2. The X-axis denotes the error bins of the models over the validation set clips. The Y-axis denotes the number of clips of the validation set, which fall into each error bin. Comparing 5.2a and 5.2c, it can be seen that, for the proposed model, the number of clips with higher values of errors is less, in the case of arousal. In the case of valence, for the proposed model, almost all the clips are grouped into the error bins ≤ 0.05 (5.2b). Whereas for the baseline model (5.2d), a significant number of clips across bins are present.



(a) Clip 206 - Arousal

(b) Clip 206 - Valence

(c) Clip 978 - Arousal

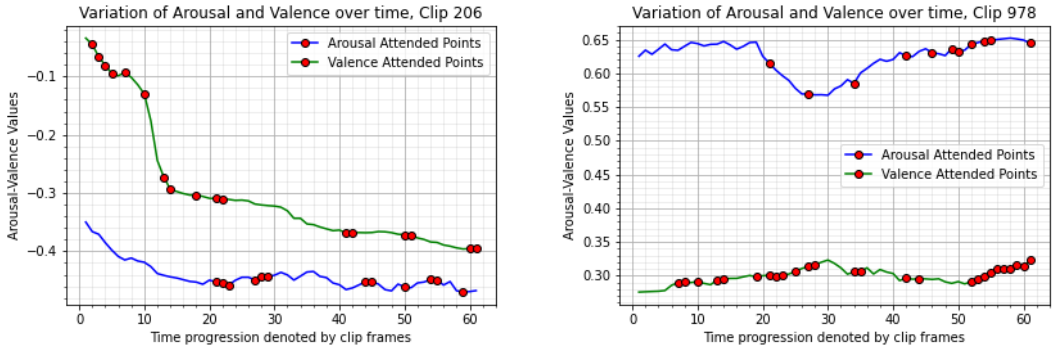
(d) Clip 978 - Valence

Figure 5.3: Attention Maps using AT models. X-axis = attention points (500ms clip frames), Y-axis = prediction points (clip's progression through time)

5.4.3 Experiment 2: Exploring Other Feature Sets

In section 5.4.2, all the experiments use the full ComPare feature set [120]. Though it performs well in dynamic emotion prediction in music, it might be noted that it is generic, not music-specific. It is large, which causes models to have a large number of parameters. Also, there might be other relevant features, which might be used for this task, eg. Constant Q Transform features. In this section, we explore some smaller feature sets detailed in section 5.4.1.2, which might possibly produce similar or better results, over the same dataset [3], with the additional benefit of being smaller in size.

Single layer *AT* models were used to train on these new feature sets, since, it was observed in section 5.4.2 that they perform best and at par with multi-layer



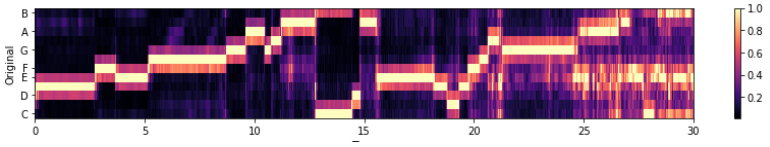
(a) Clip 206 (b) Clip 978

Figure 5.4: Comparing attended frames with ground truth Emotion ratings of dataset [3]

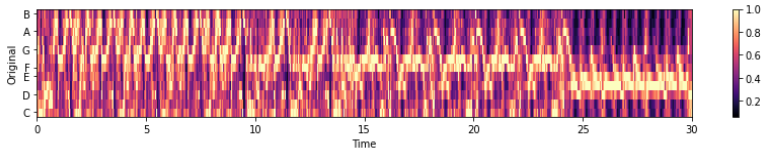
models for emotion prediction. The results are presented in tables 5.3 and 5.4 for arousal and valence respectively. For arousal (table 5.3), it is observed that *AT_64* performs well, when using the *Spectral Features* set, with a R_A^2 comparable to the best model *AT_2048_1024* using full ComPare [120] feature set. It is evident that Chroma features alone have a negligible contribution to arousal prediction. *CQT* set performs moderately. For valence prediction (table 5.4) also, *Spectral features* set performs best among all. *CQT* set does not contribute much to valence prediction. Thus, we conclude that there might be a possibility of a smaller feature set for emotion prediction.

5.4.4 Attention Maps for Emotion Prediction

Attention maps demonstrate the relative importance of layer activations at different 2D spatial locations with respect to arousal and valence predictions. In this section, the best *AT* and *BAT* models are used to generate the attention maps for both arousal and valence, for some clips chosen at random from the dataset [3], presented in 5.3 and 5.6. These maps provide information about those frames of the clip, which are attended to during emotion prediction. This in turn can yield valuable insights into specific audio features of those frames, conducive to certain emotion perception. For all the maps, X-axis signifies the attention points, which are the 500 ms frames of the clip the model attends to. The Y-axis signifies the



(a) Clip 206



(b) Clip 978

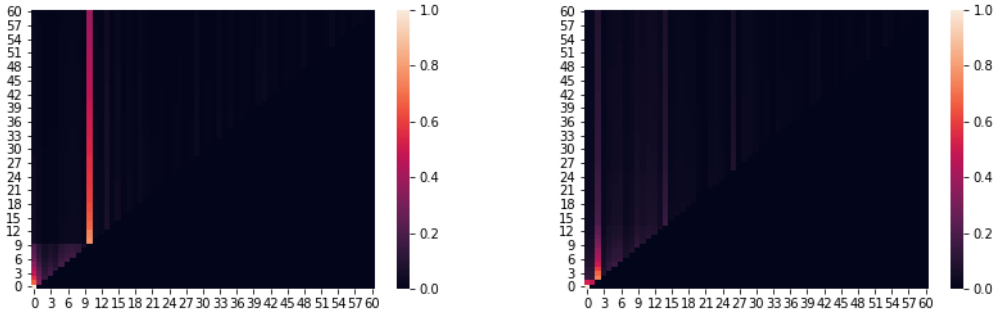
Figure 5.5: Chromagrams for Attention Map Analysis. X-axis = time (in seconds), Y-axis = Chroma. Vertical bars=Chroma intensities

prediction points and the clip's progression through time. It is to be noted that these 61 frames in the maps correspond to the last 30 seconds of each clip, as per the dataset [3]. So, the s^{th} frame of a clip, is actually the $(15 + \frac{x-1}{2})^{th}$ second of the entire 45 second clip. The vertical bars on the right of each attention map give the attention weight values present in each map. The observations are discussed in the following subsections.

5.4.4.1 Attention Maps Using AT models

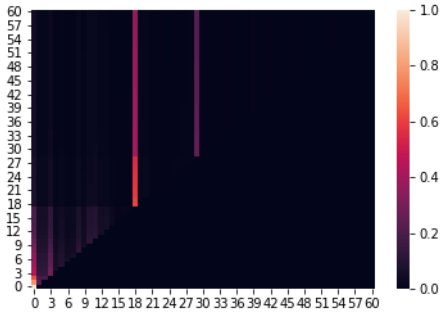
The attention maps for arousal and valence prediction, generated using *AT_2048_1024* and *AT_400*, for clips 128.mp3, 171.mp3, 206.mp3, and 978.mp3 from the dataset [3] are presented in 5.3. 5.3a and 5.3b demonstrate the attention maps for arousal and valence prediction in clip 206.mp3 . As evident from the figure 5.3a, the model attends mostly to the clip frames 20-22, 26-28, and then again frames between 43-44, 49, 53-54, and 58 to predict arousal. From figure 5.3b, it is observed that the model attends to the frames 1-4, 6, 9, 12-13, 17, 20-21, 40-41, 49-50 and 59-61 to predict valence. Similar observations can be made about the other clips as well from 5.3.

Observations: For arousal prediction, the model attends to comparatively fewer frames of the clip. These attended frames are observed to occur around 10 seconds (20 frames) after the clip has started. It can be concluded that the arousal gener-

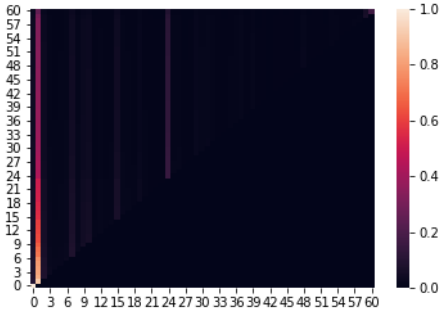


(a) Clip 60 - Arousal

(b) Clip 308 - Arousal



(c) Clip 60 - Valence



(d) Clip 308 - Valence

Figure 5.6: Attention Maps using BAT models. X-axis = attention points (500ms clip frames), Y-axis = prediction points (clip's progression through time)

ated in the later part of the music clip plays a significant role in determining the arousal perception of the entire clip. The attended frames have arousal ratings that are approximately average for all the arousal ratings for a particular clip. On the other hand, for valence prediction, attention is distributed across the clip, whenever there is a perceptible change in valence ratings. Thus it can be concluded that reports of valence depend on momentary perception. Even small changes are registered. The attended frames have quite varied valence rating values within a particular clip.

For further investigation, we juxtapose our findings with a) The dynamic arousal and valence ratings provided by the dataset [3] - ground truth, given in 5.4, and b) Chromagrams of the clips obtained using Librosa [126], presented in

5.5. In each line graph of 5.4, the X-axis denotes time frames, and Y-axis denotes the arousal and valence values.

It is to be noted here that clips 206 and 978 are so chosen that they have significantly different arousal and valence ground truth values. In clip 206, the arousal values are lesser than the valence values. In clip 978, the reported arousal values are greater than the valence values. The blue and green lines denote arousal and valence respectively, and the red dots highlight the time frames attended to by the AT models, as evident from 5.3. In each subplot of 5.5, the X-axis denotes time (in seconds), and the Y-axis denotes the Chroma. The vertical bars indicate the intensities of the Chroma. 5.5a demonstrates the chromagram for clip 206.mp3.
Observations: For arousal prediction, the model attends to those frames with the stable presence of higher notes (eg. A, B). For valence prediction, the model attends all over the chroma bins, especially when there is a change in notes in the chroma sequence of the clip. Similar observations might be made from the other chromagrams as well.

5.4.4.2 Attention Maps using BAT models

The attention maps generated using the BAT models, *BAT_2048_1024* for arousal are presented in 5.6. 5.6a gives the attention map for arousal prediction in clip 60.mp3 of the dataset [3]. As evident from the figure, the attention of the model shifts continuously throughout the clip, as it progresses in time, though Segments 11-12 receive maximum attention overall. Similar trends are observed in 5.6c as well, which represents the map for valence prediction for the same clip. Initially, the first few segments are attended to. As the clip progresses in time, the attention is shifted to later segments, with segments 18-19 and 29-30 being more prominent. As the clip progresses, the attention to initial segments reduces, rendering the lower right triangular region of the maps devoid of any attention traces.

5.5 Conclusion

We demonstrate that the state-of-the-art models for continuous-time emotion prediction perform modestly, thus emphasizing the need for further research in this

area. We have proposed an attentive LSTM-based model which improves the state-of-the-art performance significantly, on a standard benchmark dataset with standard metrics. Further, we observe that a reduced, music-specific feature set achieves similar performance to the new state-of-the-art model on arousal prediction, leading to much smaller models. Finally, we analyze attention maps for the full attention model to conclude that the model indeed attends to critical portions of the music in order to predict the dynamic emotions. We also observe that the nature of attention is different in the case of arousal and valence prediction tasks.

CHAPTER 6

Conclusions and Future Scope

In this concluding chapter, we summarize the contributions of this thesis and discuss the important directions of future work.

6.1 Summary of Contributions

The key contributions from this thesis are enumerated below:

1. The *Emotion-Word Intensity-Value* (EWIV) emotion representation: In chapter 3 we focused on arriving at a general representation of musical emotions, that maximizes the information retained from the self-report data, under a given modeling assumption. We found that the studied *EWIV* representation is able to capture perceived emotion data over both HCM excerpts and non-HCM excerpts equally well. The *reduced EWIV* emotion representation variant is statistically better than other emotion representations over various datasets.

2. The *EmoRaga* clip-set: Establishing a dedicated HCM dataset is an essential and primary step to developing MER solutions for HCM. To this end, we created the *EmoRaga* excerpt-set, containing metadata and annotated

emotion motifs in chapter 3. We also defined the concept of an *emotion motif* as any perceivable musical feature that provides strong cues to listeners to perceive particular emotions. We captured emotion opinion data using *EWIV* representation and used statistical analysis to investigate its effectiveness as an emotion representation for HCM excerpts of this dataset.

3. *Intra-contextual influence* and its effect on music emotion perception: In chapter 4, we considered a problem specific to *contexts* influencing perceived emotions in music. We used excerpts from *EmoRaga* excerpt-set to explore the effects of *immediate intrinsic musical context* on the perception of any musical piece. We termed this possible phenomenon *intra-contextual influence*. It was observed that the *immediate intrinsic musical context* plays a significant role in how we perceive emotions in music. Most often, it was able to change the *secondary* emotion or the *nuance* of the perceived emotion in a musical piece. We also compared it with similar phenomena of emotion induction [24], contagion [10], and affective priming [25].

4. Lastly, in chapter 5, we tackled the problem of explainable MEVD studies in MER and presented an attentive LSTM-based explainable dynamic emotion prediction model, using a benchmark MER dataset. We observed that the attentive-LSTM models predict dynamic perceived emotion in music better than the baseline study considered [112], and are able to *attend* to those *frames* of music which are important for emotion perception in both the benchmark dataset. We applied this model to the proposed *EmoRaga* excerpt-set (chapter 3), to predict dominant and secondary emotions in HCM excerpts. We also compared the model-predicted important *frames* with those annotated with *emotion motifs* by experts, to investigate whether these are indeed captured by the model, explaining the emotion predictions in HCM. The comparison between model-predicted frames, audience response, and expert-annotated ground truth for HCM excerpts yields significant overlap.

6.2 Future Scopes of the Work

Some future directions of each of the main contributions of this thesis are discussed in this section.

In chapter 3 of this thesis, we demonstrate the effectiveness, applicability, and *representativeness* of the *Emotion-Word Intensity Value* emotion representation (EWIV) with respect to other representations, eventually observing that the reduced variant *reduced EWIV* is consistently the best quality representation for perceived emotions in music among competing representations. The presence of a feedback loop [21] – a link between the final evaluation stage and the initial taxonomy definition might help refine the results of MER studies which are so dependent on such initial study choices. How to incorporate such feedback loops might be an interesting topic to explore further. In chapter 3 of this thesis, we also introduce a novel HCM clip-set *EmoRaga* created for MER studies in HCM. To enhance the MER results and increase the scope of possible explainable MER studies in HCM, the *EmoRaga* clip-set can be expanded to include excerpts with varied *dominant* and *secondary* emotions, other *ragas* rendered in diverse tempos and rhythmic cycles.

Chapter 4 of this thesis introduces the concepts of *intra-contextual* influence. The findings of the current study can be used to further explore how perceived emotions are modified when the dominant emotions of both *influencer* and *target* excerpts are the same, as the use of emotionally congruent excerpts might not necessarily lead to the emotional crescendo generally desired, with implications in music generation. The present work might also be used to extend the findings to excerpts of dominant emotions other than happiness and sadness.

In chapter 5 of this thesis, we present attentive-LSTM models that predict dynamic perceived emotion in music while attending to those *frames* of music which are important for emotion perception. The present work can be extended to identify musical feature sets which are responsible for perceived music emotion, by applying different methods like *LIME* [144] (Local Interpretable Model-Agnostic Explanations).

References

[1] K. Hevner, "Expression in music: a discussion of experimental studies and theories." *Psychological review*, vol. 42, no. 2, p. 186, 1935.

[2] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 1–6.

[4] E. Schubert, S. Ferguson, N. Farrar, D. Taylor, and G. E. McPherson, *The Six Emotion-Face Clock as a Tool for Continuously Rating Discrete Emotional Responses to Music*, 2013, pp. 1–18.

[5] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[6] P. J. Richerson, R. Boyd, and J. Henrich, "Gene-culture coevolution in the age of genomics," *Proceedings of the National Academy of Sciences*, vol. 107, no. supplement_2, pp. 8985–8992, 2010.

[7] D. Huron, "Is music an evolutionary adaptation?" *Annals of the New York Academy of sciences*, vol. 930, no. 1, pp. 43–61, 2001.

[8] P. N. Juslin, S. Liljeström, D. Västfjäll, G. Barradas, and A. Silva, "An experience sampling study of emotional reactions to music: listener, music, and situation." *Emotion*, vol. 8, no. 5, pp. 668–683, 2008.

[9] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, no. 3, pp. 217–238, 2004.

[10] P. N. Juslin, "From mimesis to catharsis: expression, perception, and induction of emotion in music," *Musical communication*, pp. 85–115, 2005.

[11] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences*, vol. 31, no. 05, pp. 559–575, 2008.

[12] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE transactions on consumer electronics*, vol. 64, no. 2, pp. 196–203, 2018.

[13] P. Álvarez, A. Guiu, J. R. Beltrán, J. G. de Quirós, and S. Baldassarri, "Dj-running: An emotion-based system for recommending spotify songs to runners." in *icSPORTS*, 2019, pp. 55–63.

[14] A. V. Iyer, V. Pasad, S. R. Sankhe, and K. Prajapati, "Emotion based mood enhancing music recommendation," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017, pp. 1573–1577.

[15] S. Lee, J. H. Kim, S. M. Kim, and W. Y. Yoo, "Smoodi: Mood-based music recommendation player," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–4.

[16] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Audio-based granularity-adapted emotion classification," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 176–190, 2016.

[17] L. Ferreira and J. Whitehead, "Learning to generate music with sentiment," 2019, pp. 384–390.

[18] G. Widmer, "Getting closer to the essence of music: The con espressione manifesto," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, pp. 1–13, 2016.

[19] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of music*, vol. 24, no. 1, pp. 68–91, 1996.

[20] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, p. 166335, 2022.

[21] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[22] M. Ghosh, *The Natyasastra: A Treatise on Hindu Dramaturgy and Histrionics Ascribed to Bharata-Muni: Vol. 1 (chapters 1.-27.)*. Royal Asiatic Society of Bengal, 1950.

[23] P. Patnaik, *Rasa in aesthetics: an application of rasa theory to modern Western literature*. DK Printworld, 1997.

[24] D. Västfjäll, "Emotion induction through music: A review of the musical mood induction procedure," *Musicae Scientiae*, vol. *5*, no. 1_suppl, pp. 173–211, 2001.

[25] R. H. Fazio, "On the automatic activation of associated evaluations: An overview," *Cognition & Emotion*, vol. *15*, no. 2, pp. 115–141, 2001.

[26] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[27] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–30, 2012.

[28] P. N. Juslin and J. A. Sloboda, *Music and emotion: Theory and research*. Oxford University Press, 2001.

[29] R. J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion." *Review of personality and social psychology: Emotion*, vol. 13, pp. 25–59, 1992.

[30] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press on Demand, 1991.

[31] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 693–697.

[32] W. Yang, K. Makita, T. Nakao, N. Kanayama, M. G. Machizawa, T. Sasaoka, A. Sugata, R. Kobayashi, R. Hiramoto, S. Yamawaki *et al.*, "Affective auditory stimulus database: An expanded version of the international affective digitized sounds (iads-e)," *Behavior Research Methods*, vol. 50, no. 4, pp. 1415–1429, 2018.

[33] D. J. Hargreaves, "Musical imagination: Perception and production, beauty and creativity," *Psychology of music*, vol. 40, no. 5, pp. 539–557, 2012.

[34] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 375–376.

[35] P. Laukka*, "Instrumental music teachers' views on expressivity: a report from music conservatoires," *Music Education Research*, vol. 6, no. 1, pp. 45–56, 2004.

[36] P. N. Juslin, L. Harmat, and T. Eerola, "What makes music emotionally significant? exploring the underlying mechanisms," *Psychology of Music*, vol. 42, no. 4, pp. 599–623, 2014.

[37] A. Pike, "A phenomenological analysis of emotional experience in music," *Journal of Research in Music Education*, vol. 20, no. 2, pp. 262–267, 1972.

[38] J. A. Sloboda, "Empirical studies of emotional response to music." *Cognitive bases of musical communication*, pp. 33–46, 1992.

[39] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[40] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[41] C. Mohn, H. Argstatter, and F.-W. Wilker, "Perception of six basic emotions in music," *Psychology of Music*, vol. 39, no. 4, pp. 503–517, 2011.

[42] Y. Song, S. Dixon, and M. T. Pearce, "Evaluation of musical features for emotion classification." in *ISMIR*. Citeseer, 2012, pp. 523–528.

[43] R. Panda, R. Malheiro, and R. P. Paiva, "Musical texture and expressivity features for music emotion recognition." in *ISMIR*, 2018, pp. 383–391.

[44] M. Schedl, E. Gómez, E. S. Trent, M. Tkalčič, H. Eghbal-Zadeh, and A. Martorell, "On the interrelation between listener characteristics and the perception of emotions in classical orchestra music," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 507–525, 2017.

[45] E. Schubert, S. Ferguson, N. Farrar, D. Taylor, and G. E. Mcpherson, "Continuous response to music using discrete emotion faces," *Proc. of CMMR*, 2012.

[46] S. Chaki, S. Bhattacharya, R. Mullick, and P. Patnaik, "Analyzing music to music perceptual contagion of emotion using a novel contagion interface: A case study of hindustani classical music," *Proc. of CMMR*, 2017.

[47] C. L. Krumhansl, "An exploratory study of musical emotions and psychophysiology." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 51, no. 4, p. 336, 1997.

[48] N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect." *Journal of personality and social psychology*, vol. 79, no. 2, pp. 286–300, 2000.

[49] P. E. Bestelmeyer, S. A. Kotz, and P. Belin, "Effects of emotional valence and arousal on the voice perception network," *Social cognitive and affective neuroscience*, vol. 12, no. 8, pp. 1351–1358, 2017.

[50] R. E. S. Panda, "Emotion-based analysis and classification of audio music," Ph.D. dissertation, 00500:: Universidade de Coimbra, 2019.

[51] H. Egermann and S. McAdams, "Empathy and emotional contagion as a link between recognized and felt emotions in music listening," *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 2, pp. 139–156, 2013.

[52] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 209–222, 2017.

[53] F. Weninger, F. Eyben, and B. Schuller, "The tum approach to the mediaeval music emotion task using generic affective audio features," in *Proceedings MediaEval 2013 Workshop, Barcelona, Spain*, 2013.

[54] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a gaussian mixture model," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 56–68, 2015.

[55] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *ISMIR*, pp. 370–375, 2018.

[56] D. Watson, D. Wiese, J. Vaidya, and A. Tellegen, "The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence." *Journal of personality and social psychology*, vol. 76, no. 5, pp. 820–838, 1999.

[57] G. L. Collier, "Beyond valence and activity in the emotional connotations of music," *Psychology of Music*, vol. 35, no. 1, pp. 110–131, 2007.

[58] T. Eerola and K. Vuoskoski, J., "A review of music and emotion studies: approaches, emotion models, and stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2013.

[59] E. Parada-Cabaleiro, M. Schmitt, A. Batliner, S. Hantke, G. Costantini, K. Scherer, and B. Schuller, "Identifying emotions in opera singing: Implications of adverse acoustic conditions," *ISMIR*, pp. 376–382, 2018.

[60] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-relevant features for classification and regression of music lyrics," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240–254, 2016.

[61] X. Hu, F. Li, and T.-D. J. Ng, "On the relationships between music-induced emotion and physiological signals." in *ISMIR*, 2018, pp. 362–369.

[62] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.

[63] Y. Song, S. Dixon, M. T. Pearce, and A. R. Halpern, "Perceived and induced emotion responses to popular music: Categorical and dimensional models," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 4, pp. 472–492, 2016.

[64] L. B. Meyer, *Emotion and meaning in music*. University of Chicago Press, 2008.

[65] P. N. Juslin, S. Liljestrom, D. Vastfjall, and L.-O. Lundqvist, "How does music evoke emotions? exploring the underlying mechanisms." *Handbook of music and emotion: theory, research, applications.*, pp. 605–642, 2010.

[66] D. J. Hargreaves, R. MacDonald, and D. Miell, "How do people communicate using music," *Musical communication*, vol. 1, pp. 1–26, 2005.

[67] T. Eerola, A. Friberg, and R. Bresin, "Emotional expression in music: contribution, linearity, and additivity of primary musical cues," *Frontiers in psychology*, vol. 4, p. 487, 2013.

[68] C. Labbé, W. Trost, and D. Grandjean, "Affective experiences to chords are modulated by mode, meter, tempo, and subjective entrainment," *Psychology of Music*, vol. *49*, no. 4, pp. 915–930, 2021.

[69] L. Quinto, W. F. Thompson, and A. Taylor, "The contributions of compositional structure and performance expression to the communication of emotion in music," *Psychology of Music*, vol. 42, no. 4, pp. 503–524, 2014.

[70] D. Tan, F. M. Diaz, and P. Miksza, "Expressing emotion through vocal performance: Acoustic cues and the effects of a mindfulness induction," *Psychology of Music*, vol. *48*, no. 4, pp. 495–512, 2020.

[71] J. K. Vuoskoski and T. Eerola, "Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences," *Musicae Scientiae*, vol. 15, no. 2, pp. 159–173, 2011.

[72] L. Xu, X. Wen, J. Shi, S. Li, Y. Xiao, Q. Wan, and X. Qian, "Effects of individual factors on perceived emotion and felt emotion of music: Based on machine learning methods," *Psychology of Music*, vol. *49*, no. 5, pp. 1069–1087, 2021.

[73] B. Kerkova, "Perception and experience of musical emotions in schizophrenia," *Psychology of Music*, vol. 48, no. 2, pp. 199–214, 2020.

[74] C. F. Lima, A. I. Correia, D. Müllensiefen, and S. L. Castro, "Goldsmiths musical sophistication index (gold-msi): Portuguese version and associations with socio-demographic factors, personality and music preferences," *Psychology of Music*, vol. *48*, no. 3, pp. 376–388, 2020.

[75] A. Gabrielsson and E. Lindström, "The role of structure in the musical expression of emotions," *Handbook of music and emotion: Theory, research, applications*, vol. 367400, 2010.

[76] K. L. Spreadborough and I. Anton-Mendez, "It's not what you sing, it's how you sing it: How the emotional valence of vocal timbre influences listeners' emotional perception of words," *Psychology of Music*, vol. 47, no. 3, pp. 407–419, 2019.

[77] S. Beveridge and D. Knox, "Popular music and the role of vocal melody in perceived emotion," *Psychology of Music*, vol. 46, no. 3, pp. 411–423, 2018.

[78] S. Liljeström, P. N. Juslin, and D. Västfjäll, "Experimental evidence of the roles of music choice, social context, and listener personality in emotional reactions to music," *Psychology of Music*, vol. *41*, no. 5, pp. 579–599, 2013.

[79] T. t. Bogt, N. Canale, M. Lenzi, A. Vieno, and R. v. d. Eijnden, "Sad music depresses sad adolescents: A listener's profile," *Psychology of Music*, vol. *49*, no. 2, pp. 257–272, 2021.

[80] J. L. Larwood and G. A. Dingle, "The effects of emotionally congruent sad music listening in young adults high in rumination," *Psychology of Music*, p. 0305735620988793, 2021.

[81] M. D. Rucsanda, A.-M. Cazan, and C. Truța, "Musical performance and emotions in children: The case of musical competitions," *Psychology of music*, vol. 48, no. 4, pp. 480–494, 2020.

[82] E. Commodari and J. Sole, "Music education in junior high school: Perception of emotions conveyed by music and mental imagery in students who attend the standard or musical curriculum," *Psychology of Music*, vol. 48, no. 6, pp. 824–835, 2020.

[83] S. Saarikallio, V. Alluri, J. Maksimainen, and P. Toiviainen, "Emotions of music listening in finland and in india: comparison of an individualistic and a collectivistic culture," *Psychology of Music*, vol. *49*, no. 4, pp. 989–1005, 2021.

[84] H. Argstatter, "Perception of basic emotions in music: Culture-specific or multicultural?" *Psychology of Music*, vol. 44, no. 4, pp. 674–690, 2016.

[85] H. Y. Park and H. J. Chong, "A comparative study of the perception of music emotion between adults with and without visual impairment," *Psychology of Music*, vol. 47, no. 2, pp. 225–240, 2019.

[86] N. Ruth, ""they don't really care...": Effects of music with prosocial content and corresponding media coverage on prosocial behavior," *Musicae Scientiae*, vol. 22, no. 3, pp. 415–433, 2018.

[87] N. Ruth and H. Schramm, "Effects of prosocial lyrics and musical production elements on emotions, thoughts and behavior," *Psychology of Music*, vol. 49, no. 4, pp. 759–776, 2021.

[88] K. E. Buckley and C. A. Anderson, "A theoretical model of the effects and consequences of playing video games," *Playing video games: Motives, responses, and consequences*, pp. 363–378, 2006.

[89] E. Coutinho and K. R. Scherer, "The effect of context and audio-visual modality on emotions elicited by a musical performance," *Psychology of Music*, vol. 45, no. 4, pp. 550–569, 2017.

[90] F. Greb, W. Schlotz, and J. Steffens, "Personal and situational influences on the functions of music listening," *Psychology of Music*, vol. *46*, no. 6, pp. 763–794, 2018.

[91] W. M. Randall and N. S. Rickard, "Personal music listening: A model of emotional outcomes developed through mobile experience sampling," *Music Perception: An Interdisciplinary Journal*, vol. *34*, no. 5, pp. 501–514, 2017.

[92] W. M. Randall, N. S. Rickard, and D. A. Vella-Brodrick, "Emotional outcomes of regulation strategies used during personal music listening: A mobile experience sampling study," *Musicae Scientiae*, vol. 18, no. 3, pp. 275–291, 2014.

[93] J. Steffens, "The influence of film music on moral judgments of movie scenes and felt emotions," *Psychology of Music*, vol. *48*, no. 1, pp. 3–17, 2020.

[94] A.-K. Herget, "On music's potential to convey meaning in film: A systematic review of empirical evidence," *Psychology of Music*, vol. *49*, no. 1, pp. 21–49, 2021.

[95] J. Cespedes-Guevara and T. Eerola, "Music communicates affects, not basic emotions–a constructionist account of attribution of emotional meanings to music," *Frontiers in psychology*, vol. *9*, p. 215, 2018.

[96] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[97] L. F. Barrett, "Are emotions natural kinds?" *Perspectives on psychological science*, vol. 1, no. 1, pp. 28–58, 2006.

[98] N. Cook, "Theorizing musical meaning," *Music theory spectrum*, vol. 23, no. 2, pp. 170–195, 2001.

[99] R. Timmers and H. Crook, "Affective priming in music listening: Emotions as a source of musical expectation," *Music Perception: An Interdisciplinary Journal*, vol. *31*, no. 5, pp. 470–484, 2012.

[100] R. Timmers, Y. Arthurs, and H. Crook, "Stream segregation revisited: Dynamic listening and influences of emotional context on stream perception and attention," *Consciousness and Cognition*, vol. *85*, p. 103027, 2020.

[101] E. M. Huziwara, Á. M. Cedro, R. M. Rodrigues, T. P. de Oliveira, R. Bortoloti, and A. Jaeger, "Affective priming caused by musical chords on human facial expressions," *Psychology of Music*, p. 03057356221097996, 2022.

[102] R. Y. L. Tay and B. C. Ng, "Effects of affective priming through music on the use of emotion words," *PloS one*, vol. *14*, no. 4, p. e0214482, 2019.

[103] N. Steinbeis and S. Koelsch, "Affective priming effects of musical sounds on the processing of word meaning," *Journal of cognitive neuroscience*, vol. *23*, no. 3, pp. 604–621, 2011.

[104] J. Armitage and T. Eerola, "Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced, and convenience web samples," *Frontiers in psychology*, vol. *10*, pp. 28–83, 2020.

[105] ——, "Cross-modal transfer of valence or arousal from music to word targets in affective priming?" *Auditory Perception & Cognition*, vol. *5*, no. 3-4, pp. 192–210, 2022.

[106] K. S. Goerlich, J. Witteman, N. O. Schiller, V. J. Van Heuven, A. Aleman, and S. Martens, "The nature of affective priming in music and speech," *Journal of cognitive neuroscience*, vol. *24*, no. 8, pp. 1725–1741, 2012.

[107] J. D. March, "Affective priming of music and words," Ph.D. dissertation, Memorial University of Newfoundland, 2010.

[108] D. Yang and T. Tsai, "Piano sheet music identification using dynamic n-gram fingerprinting," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, 2021.

[109] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[110] E. Zangerle, C.-M. Chen, M.-F. Tsai, and Y.-H. Yang, "Leveraging affective hashtags for ranking music recommendations," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 78–91, 2018.

[111] C. Laurier, P. Herrera, M. Mandel, and D. Ellis, "Audio music mood classification using support vector machine," *MIREX task on Audio Mood Classification*, pp. 2–4, 2007.

[112] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5412–5416.

[113] S. Shih and H. Chi, "Automatic, personalized, and flexible playlist generation using reinforcement learning," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 168–174.

[114] Y. Liu, Y. Liu, Y. Zhao, and K. A. Hua, "What strikes the strings of your heart?—feature mining for music emotion analysis," *IEEE TRANSACTIONS on Affective computing*, vol. 6, no. 3, pp. 247–260, 2015.

[115] E. Schubert, S. Ferguson, N. Farrar, and G. E. McPherson, "Sonification of emotion i: Film music." International Community for Auditory Display, 2011.

[116] D. Makris, I. Karydis, and S. Sioutas, "The greek music dataset," in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, 2015, pp. 1–7.

[117] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Towards time-varying music auto-tagging based on cal500 expansion," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[118] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[119] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A comparative study of collaborative vs. traditional musical mood annotation." in *ISMIR*, vol. 104, 2011, pp. 549–554.

[120] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[121] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[122] F. Eyben, *Real-time speech and music classification by large audio feature space extraction.* Springer, 2015.

[123] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software." in *ISMIR*, vol. 2010. Citeseer, 2010, pp. 441–446.

[124] C. McKay, I. Fujinaga, and P. Depalle, "jaudio: A feature extraction library," in *Proceedings of the international conference on music information retrieval*, vol. 11, no. 15, 2005, pp. 600–3.

[125] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[126] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[127] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR*, vol. 86, 2010, pp. 937–952.

[128] X. Hu, J. S. Downie, and A. F. Ehmann, "Lyric text mining in music mood classification," *American music*, vol. 183, no. 5,049, pp. 2–209, 2009.

[129] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *2008 seventh international conference on machine learning and applications*. IEEE, 2008, pp. 688–693.

[130] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[131] S. Fukayama and M. Goto, "Music emotion recognition with adaptive aggregation of gaussian process regressors," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 71–75.

[132] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. W. Schuller, R. Veltkamp, F. Weninger, and F. Wiering, "Emotional analysis of music: A comparison of methods," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1161–1164.

[133] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 267–274.

[134] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 5–18, 2005.

[135] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, pp. 765–783, 2020.

[136] E. Coutinho, F. Weninger, B. W. Schuller, and K. R. Scherer, "The munich lstm-rnn approach to the mediaeval 2014" emotion in music'" task." in *MediaEval*, 2014.

[137] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio." in *IJCAI*, vol. 2016, 2016, pp. 2196–2202.

[138] S. Giammusso, M. Guerriero, P. Lisena, E. Palumbo, and R. Troncy, "Predicting the emotion of playlists using track lyrics," *ISMIR, Late Breaking Session*, 2017.

[139] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based emotion recognition for experimental music." in *ISMIR*, 2017, pp. 368–375.

[140] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," *ISMIR*, 2011.

[141] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[142] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[143] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[144] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[145] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[146] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *The Journal of Machine Learning Research*, vol. 11, pp. 1–18, 2010.

[147] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards explainable music emotion recognition: The route via mid-level features," *arXiv preprint arXiv:1907.03572*, 2019.

[148] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2014." in *MediaEval*, 2014.

[149] V. Haunschmid, S. Chowdhury, and G. Widmer, "Two-level explanations in music emotion recognition," *arXiv preprint arXiv:1905.11760*, 2019.

[150] J. de Berardinis, A. Cangelosi, and E. Coutinho, "The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability," in *ISMIR*, 2020, pp. 310–317.

[151] S. Bagchee, *Nad*. BPI Publishing, 1998.

[152] S. Gulati, J. Serra, V. Ishwar, and X. Serra, "Discovering rāga motifs by characterizing communities in networks of melodic patterns," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 286–290.

[153] A. Mathur, S. H. Vijayakumar, B. Chakrabarti, and N. C. Singh, "Emotional responses to hindustani raga music: the role of musical structure," *Frontiers in psychology*, vol. 6, p. 513, 2015.

[154] M. A. Castellano, J. J. Bharucha, and C. L. Krumhansl, "Tonal hierarchies in the music of north india." *Journal of Experimental Psychology: General*, vol. 113, no. 3, p. 394, 1984.

[155] BharataMuni, "Nātyasāstra of bharata. chapter six. rasādhyāyah on the sentiments: A commentary by abhinavagupta," Ph.D. dissertation, 1926.

[156] U. Gupta and B. Gupta, "Psychophysiological responsivity to indian instrumental music," *Psychology of music*, vol. *33*, no. 4, pp. 363–372, 2005.

[157] K. Kunikullaya Ubrangala, R. Kunnavil, J. Goturu, V. S. Prakash, and N. S. Murthy, "Effect of specific melodic scales of indian music in reducing state and trait anxiety: A randomized clinical trial," *Psychology of Music*, vol. *50*, no. 5, pp. 1390–1407, 2022.

[158] J. M. Valla, J. A. Alappatt, A. Mathur, and N. C. Singh, "Music and emotion—a case for north indian classical music," *Frontiers in psychology*, vol. 8, p. 2115, 2017.

[159] G. K. Koduri, S. Gulati, and P. Rao, "A survey of raaga recognition techniques and improvements to the state-of-the-art," *Sound and Music Computing*, 2011.

[160] K. K. Ganguli, S. Gulati, X. Serra, and P. Rao, "Data-driven exploration of melodic structure in hindustani music," in *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. ISMIR 2016. Proceedings of the 17th International Society for Music Information Retrieval Conference; 2016 Aug 7-11; New York City (NY).[Canada]: ISMIR; 2016. p. 605-11.* International Society for Music Information Retrieval (ISMIR), 2016.

[161] K. Narang and P. Rao, "Acoustic features for determining goodness of tabla strokes." in *ISMIR*, 2017, pp. 257–263.

[162] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Proceedings of the 2014 ICMC/SMC; 2014 Michigan Publishing; 2014.* Michigan Publishing, 2014.

[163] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, "Saraga: Open datasets for research on indian art music," *Empirical Musicology Review*, vol. 16, no. 1, pp. 85–98, 2021.

[164] X. Serra, "Creating research corpora for the computational study of music: the case of the compmusic project," in *Audio engineering society conference: 53rd international conference: Semantic audio.* Audio Engineering Society, 2014.

[165] S. Chaki, P. Doshi, P. Patnaik, and S. Bhattacharya, "Attentive rnns for continuous-time emotion prediction in music clips," in *Proceedings of the 3rd Workshop on Affective Content Analysis (AffCon 2020) co-located AAAI*, ser. CEUR Workshop Proceedings, vol. 2614. CEUR-WS.org, 2020, pp. 36–46.

[166] N. H. Frijda and L. Sundararajan, "Emotion refinement: A theory inspired by chinese poetics," *Perspectives on Psychological Science*, vol. 2, no. 3, pp. 227–241, 2007.

[167] H. Akaike, "Akaike's information criterion," *International encyclopedia of statistical science*, pp. 25–25, 2011.

[168] S. Mo and J. Niu, "A novel method based on ompgw method for feature extraction in automatic music mood classification," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 313–324, 2017.

[169] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, 2017.

[170] J. Han, Z. Zhang, Z. Ren, and B. W. Schuller, "Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Transactions on Affective Computing*, 2019.

[171] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[172] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[173] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[174] J. C. Ross, T. Vinutha, and P. Rao, "Detecting melodic motifs from audio for hindustani classical music." in *ISMIR*, 2012, pp. 193–198.

[175] P. Chordia and A. Rae, "Raag recognition using pitch-class and pitch-class dyad distributions." in *ISMIR*. Citeseer, 2007, pp. 431–436.

[176] V. Bhatkhande, "Hindustani sangeet paddhati: Kramik pustak maalika vol. i-vi," *Sangeet Karyalaya*, vol. 72, 1990.

[177] R. Jha, "Abhinav geetanjali vol. iv," *Sangeet Sadan*, vol. 72, 2001.

[178] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. A. Murthy, "Classification of melodic motifs in raga music with time-series matching," *Journal of New Music Research*, vol. 43, no. 1, pp. 115–131, 2014.

[179] P. Dighe, H. Karnick, and B. Raj, "Swara histogram based structural analysis and identification of indian classical ragas." in *ISMIR*, 2013, pp. 35–40.

[180] S. Chaki, P. Doshi, S. Bhattacharya, and P. Patnaik, "Explaining perceived emotion predictions in music: An attentive approach," in *Proceedings of the 21st Conference of the International Society for Music Information Retrieval (ISMIR 2020)*. International Society for Music Information Retrieval (ISMIR), 2020.

[181] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[182] J. Borgohain, D. Suar, and P. Patnaik, *Continuous Music Rating of Emotions in Hindustani Classical Music: An Exploration Using Web Interfaces.* Thesis for the Degree of Master of Science, Department of Humanities and Social Sciences, Indian Institute of Technology, Kharagpur, 2017.

[183] L. Wheeler, "Toward a theory of behavioral contagion." *Psychological review*, vol. *73*, no. 2, p. 179, 1966. [Online]. Available: http://dx.doi.org/10.1037/h0023023

[184] P. N. Juslin, "From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions," *Physics of life reviews*, vol. 10, no. 3, pp. 235–266, 2013.

[185] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[186] S. Balke, M. Dorfer, L. Carvalho, A. Arzt, and G. Widmer, "Learning soft-attention models for tempo-invariant audio-sheet music retrieval," *ISMIR*, pp. 216–222, 2019.

[187] S. Gururani, M. Sharma, and A. Lerch, "An attention mechanism for musical instrument recognition," *ISMIR*, pp. 83–90, 2019.

[188] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training," *ISMIR*, pp. 685–692, 2019.

[189] T.-P. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," *ISMIR*, pp. 259–267, 2019.

[190] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, "A bi-directional transformer for musical chord recognition," *ISMIR*, pp. 620–627, 2019.