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

Sanga Chaki, Sourangshu Bhattacharya, Junmoni Borgohain,
Priyadarshi Patnaik, Raju Mullick, Gouri Karambelkar

Abstract—Data-driven analysis and modeling of music-perceived emotions have widespread applications in MIR, with representations of perceived musical emotions forming a crucial component. Though some emotion representations are popular in the literature, their relative merits and demerits in terms of expressiveness and broad applicability have been sparsely studied. The application-specific emotion representations used in multiple studies lead to incomparability of algorithms and performance metrics and non-reusability of representation-specific emotion data across studies. In this work, we study an intensity ratings-based, categorical emotion representation called Emotion-Word Intensity-Value (EWIV) representation, with emotion classes adapted from the aesthetic concept of Nava Rasa. We also introduce EmoRaga - a novel clip-set annotated with perceived emotions and emotion motifs for emotion analysis of Hindustani classical music. We explore the applicability of EWIV towards diverse MIR applications, e.g., dominant and secondary emotion identification, and temporal emotion pattern study. Last, we report a data-driven comparison of EWIV, categorical and dimensional representations, using statistical out-of-sample goodness of fit tests to measure and compare their representativeness over both benchmark datasets and collected emotion data. We conclude that EWIV is applicable to a range of MIR tasks, with higher representative and generalization potential compared to popular representations in certain cases.

Index Terms—Perceived emotions; Music emotion representation; Emotion motifs; Emotion classification; Statistical Emotion Modeling; Hindustani classical music

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 [1], [2], sentiment-based music generation [3], music emotion recognition (MER) [4], [5], [6], [7], etc. *Emotion representations* are an essential component of such studies, which determine the measurement (e.g., self-reports), storage (dataset creation), and processing (e.g., machine learning models) of perceived emotion data. The representation controls the information content extracted from the emotion annotations given by the subjects 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 categorical representation (e.g., [8]) and dimensional representation (e.g., [9]) are among the most popular representations of musical emotions. The task of MER can be formulated either as a classification problem using categorical representation [4] or a regression problem using a dimensional representation [7]. Datasets, algorithms, and performance metrics used in the studies vary depending on the emotion-representation format. Hence, the choice of an appropriate emotion representation for a given task is a widely debated and open research topic. In this paper, we analyze and compare various emotion representations for **broad applicability** in various MIR tasks.

While the categorical and dimensional representations are popular, they both have notable limitations like the number of emotion classes might be too small [10], fuzzy demarcations between emotions [11], [12], etc. To overcome these limitations, many researchers use features from multiple representation models [13], [1] or incorporate additional measurements in existing representations, e.g., rating scales for discrete emotions [14], or dynamic annotations [15], [7]. Many studies use the end task as a motivation for the selection of emotion representation, e.g., Lee et al. [1] used mood categories and arousal-valence for mood-based recommender systems, Shepstone et al. [2] used 12 categorical components and arousal-valence focus for granularity-adapted emotion classification of audio, Parada et al. [16] used 10 categorical components of emotion and intensity labels for MER under adverse conditions, Panda et al. [17] and Malheiro et al. [6] used mood tags and quadrant information to explore MER relevant feature. This approach of selecting "appropriate" emotion representations has two major drawbacks: (1) the differences in emotion representations result in different downstream algorithms and metrics, which cannot be compared directly, and (2) emotion information from existing high-quality annotated datasets cannot be borrowed and re-utilized for the design of different high-quality AI models and algorithms. These

- *S. Chaki, R. Mullick and G. Karambelkar are with the Advanced Technology Development Centre, Indian Institute of Technology, Kharagpur, India.*
  E-mail: sanga@iitkgp.ac.in
- *S. Bhattacharya is with the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India.*
- *J. Borgohain and P. Patnaik are with the Department of Humanities and Social Sciences, Indian Institute of Technology, Kharagpur, India.*

drawbacks can be mitigated by using a common general-purpose representation of musical emotions chosen based on the broad criterion like maximum information retained from several self-report datasets under a given statistical modeling assumption (e.g., normal distribution).

Creating high-quality annotated datasets for perceived emotions for different types of music clips is a costly endeavor. Some notable past efforts of such datasets include Soleymani et al. [15], which uses the Circumplex model of affect [9] for representing emotions on western-origin clips, and Eerola et al. [14], which uses both the discrete and dimensional representations of emotions on film music-clips. However, most high-quality freely-available datasets with annotations of perceived emotions do not include *Hindusthani Classical Music* (HCM) clips. Emotion representations in existing datasets also do not feature emotions-sets from the Indian Aesthetics literature, e.g., the *Nava Rasa* [18], [19] concept. In this paper, we study a dynamic (time-varying), intensity ratings-based, categorical emotion representation inspired by HCM literature (*Nava Rasa* [18], [19]), called the *Emotion-Word Intensity-Value* (*EWIV*) representation (section 3). We demonstrate the effectiveness of *EWIV* on existing benchmark clip sets [15], [20], and [14], as well as a newly introduced set of clips from Hindusthani classical music (HCM), called the *EmoRaga* clip set for perceived emotion analysis in HCM (section 4.1.2). Estimations of dominant emotions from self-reported emotions data obtained through crowd-sourced surveys (section 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 5.1). We also validate the quality of self-reported emotions through the typicality of a clip toward the estimated dominant emotion (section 5.2), as well as by measuring the inter-listener agreement through Cronbach's alpha (section 5.3). The EWIV representation is also used to detect clips with ambiguous perceived emotions (section 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* representation estimated from self-reported emotion data towards two MER applications in section 6, emotion classification (section 6.1), and temporal emotion variation detection (section 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 [21], [22] 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 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 [9]. In section 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 7.1). While this conversion can incur some loss, we observe (section 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.

2 Background

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 [10], [8], [23], [24]. The dimensional approach [10], [23], [25] identifies emotions from coordinates of dimensions like *valence*, *arousal* and *dominance*, e.g., Russell's Circumplex model of affect [9] uses arousal and valence. Though both are used extensively in different MER tasks [10], [26], their drawbacks are also much researched [27], [28], [12], [11]. While most studies use a single model, some prefer both models simultaneously or combinations of measures from both models [1], [16], [17], [6], [29]. The motivation is either task-specific or to overcome the disadvantages of any single approach. Eerola et al. [13] 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. [14] compared the intensity-based discrete emotion model with the dimensional model in their seminal work. Shepstone et al. [2] 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 interesting research question that might remain unanswered is: *Is the information content of discrete emotion categories with dynamic time-continuous 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.

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 emotion prediction and explanation. The Akaike information criterion (AIC) [30] 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 of the model. 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:

$$AIC_M = 2k_M - 2\ln(\hat{L}_M) \quad (1)$$

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

2.2 Perceived-Emotion Tasks in MIR

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. [14] collected perceived-emotion data in both discrete and dimensional representations for a set of 110 excerpts, and compared the two representations. Schubert et al. [20] 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. [15] 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 [32] and the IADS dataset [33]. To the best of our knowledge, no such dataset exists with Hindustani classical music (HCM) excerpts and perceived emotion annotations.

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) [34], [2], [35], or random forest classifiers to classify discrete emotions [36]. Recently, Han et al. [37] 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 recognition system. CNNs are used for audio tagging, music classification, speech emotion classification and sound event detection [38], [39]. Xie et al. [40] 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 [10], [41]. Most of the present literature on MEVD is based on the dimensional arousal-valence emotion annotations [26].

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 [42]. Each *raga* is characterized by a set of notes, the ascending-descending melodic progression and a specific set of melodic phrases termed *raga motifs* [43] . 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 [43], [44], analysis of melodic [45] and rhythmic components [46], and related corpus creation [47], [48], [49] 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 paper 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 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.

3.1 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 [19], a major part of Indian aesthetics [18], [50], which describes the nine primary aesthetic flavours and/or the emotions evoked by any 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. 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 intensity 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.

3.2 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* ($\boldsymbol{IIV}^{\boldsymbol{c,\lambda}}(\boldsymbol{n})$, $n \in \{1, \ldots, N^{c,\lambda}\}$) at time ($t$), such that all intensities have zero values, except the ε^{th} intensity, which has value I. The *cumulative intensity vector* ($\boldsymbol{CIV}^{\boldsymbol{c,\lambda}}$) is calculated from all the $\boldsymbol{IIV}^{\boldsymbol{c,\lambda}}$s so that each element ($\boldsymbol{CIV}^{\boldsymbol{c,\lambda}}_{\boldsymbol{\varepsilon}}$) is the summation of all intensities of the associated emotion (ε), independent of the other emotions (eq. 2). Each intensity in the *per listener-per excerpt* ($\boldsymbol{EWIV}^{c,\lambda}$) measure is estimated by normalizing $\boldsymbol{CIV}^{\boldsymbol{c,\lambda}}$ with respect to $N^{c,\lambda}$ (eq. 3). Each intensity in the *per excerpt* ($\boldsymbol{EWIV}^{c}$) measure is estimated by aggregating over the set of listeners Λ (eq. 4).

$$\boldsymbol{CIV}^{c,\lambda}_{\varepsilon} = \sum_{n=1}^{N^{c,\lambda}} \boldsymbol{IIV}^{c,\lambda}_{\varepsilon}(n) \tag{2}$$

$$\boldsymbol{EWIV}^{c,\lambda}_{\varepsilon} = \frac{\boldsymbol{CIV}^{c,\lambda}_{\varepsilon}}{N^{c,\lambda}} \tag{3}$$

$$\boldsymbol{EWIV}^{c}_{\varepsilon} = \sum_{\lambda \in \Lambda} \boldsymbol{EWIV}^{c,\lambda}_{\varepsilon} \tag{4}$$

Finally, each probability in the *probability vector* ($\boldsymbol{pEWIV}^{c}$) is calculated (eq. 5).

$$\boldsymbol{pEWIV}^{c}_{\varepsilon} = \frac{\boldsymbol{EWIV}^{c}_{\varepsilon}}{\sum_{\varepsilon \in \mathcal{E}} \boldsymbol{EWIV}^{c}_{\varepsilon}} \times 100\% \tag{5}$$

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. For illustration, we provide a demonstration of computation of $\boldsymbol{pEWIV}^{c}$ in Appendix D. The same procedure as above is followed to estimate the *per-segment per-excerpt* measure ($\boldsymbol{pEWIV}^{c,s}$), with the additional temporal constraint.

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.

4.1 Stimuli

4.1.1 Excerpts from Preexisting Datasets

Schubert_6: The six excerpt excerpt-set used by Schubert et al. [20] (section 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 [15] (section 2.2.1). The static arousal-valence annotations are mapped to emotion words [9] and are considered ground truth. We named both these excerpt sets for ease of discussion. For further details of all datasets, please refer to Appendix A.

4.1.2 EmoRaga: An HCM Excerpt-set for MER

The general guidelines for the design of research corpora for computational music studies [49], [47] 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 HCM practitioners and musicologists, with over ten years of training under reputed gurus. An overview of this dataset is reported in table 1. To ensure uniformity among the excerpts, the following criteria are maintained: a) The excerpts are 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 other 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 experts panel. The excerpts belong to 23 different *raga*s [51], [52], 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 [53], [54] to annotate each excerpt with relevant scores manually. The contextual

TABLE 1: EmoRaga Dataset Content Summary

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

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* [43] 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. These *emotion motifs* include, but may not be limited to: a) Presence of major or minor notes, b) Faster or slower tempo, c) *Raga* related significant multi-note structures or phrases called *mukhyang*s/*pakad*s [55]/*raga motif*s [43], in exact or broken forms, d) Presence of *raga*-dependent *Vadi* and *Samvadi* notes [56], 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.

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 members also volunteered for the surveys. The listeners do not have formal HCM training or significant prior exposure. 69.95% of the participants identified as male (μ_{age}=20.21, σ_{age}=4.89, $range_{age}$=[18,56]). 30.05% identified as female (μ_{age}=22.64, σ_{age}=6.99, $range_{age}$=[18,59]). 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 collected data were guaranteed.

4.3 Survey Procedure

The interface [21], [57] presented in Figure 1 of Appendix A 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.

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.

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 3.2). The static dataset consists of one probability vector, one dominant emotion, and one secondary emotion per excerpt. Table 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 rest is presented in Appendix A. 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 $OE\% \ldots E\%$. 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 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* [20]. Columns α and τ present Cronbach's alpha and typicality measures respectively, derived from analysis of the *per-excerpt* data, discussed in sections 5.2 to 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 6.

TABLE 2: EWIV survey results over two pre-exiting excerpt-sets - **Schubert_6** [20] and **Soleymani_5** [15], and the **EmoRaga** excerpt-set introduced in section 4.1.2. # = Original study's excerpt number, Excerpt Emotion = Ground Truth. Near Miss = Near miss emotion reported in Schubert et al [20]. #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 [20]															
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 [15]															
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 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

5.2 EWIV and Typicality

The *typicality* (τ) of an excerpt to a particular emotion [14] is described as the property by which that emotion is more easily perceived in that excerpt than other emotions. It is estimated as $\tau = \overline{E} - SE - \overline{NE}$ [14], 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 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.

5.3 Listener Consensus in EWIV

Cronbach's alpha (α) [14] 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 2, column (α). It is observed that most excerpts have a high α value, demonstrating 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 (Appendix A) 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.

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* (Appendix A) 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* (Appendix A). 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.

6 Applications

6.1 Dynamic Emotion Classification

In this section, the *dynamic EmoRaga* dataset (section 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} (3.1). The second one is joint dominant and secondary emotion labeling - a

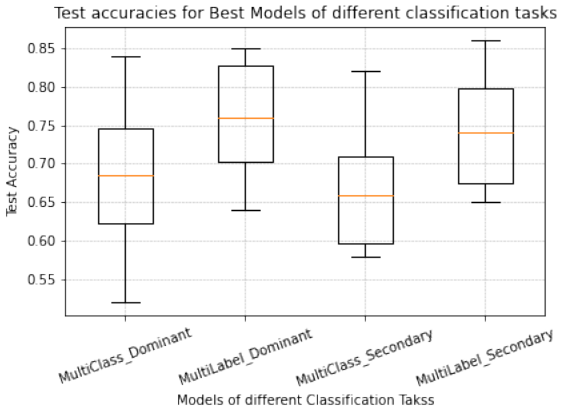


Fig. 1: Test accuracies across K=10 fold Cross Validation for *dominant* and *secondary* emotion classification using multi-class and multi-label approaches. The mean accuracies are represented as the colored lines. For multi-class dominant emotion prediction, the mean = 0.68, sd = 0.09. For multi-label dominant, the mean = 0.76, sd = 0.06. For multi-class secondary, mean = 0.67, sd = 0.07. For multi-label secondary, mean = 0.74, sd = 0.07.

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.

6.1.1 Experimental Setup

The dynamic dataset derived from the *per segment-per excerpt* probability vectors 5.1 is used for this task. The *spectral features* of the segments are extracted using the Librosa [58] tool. They denote the distributions of energy over a set of frequencies and have provided reasonably good emotion estimates previously [59]. 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 22050 Hz before feature extraction. The standard scalar normalization is used for preprocessing the data before MIR tasks.

The LSTM-RNN [60] 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 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.

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 1.

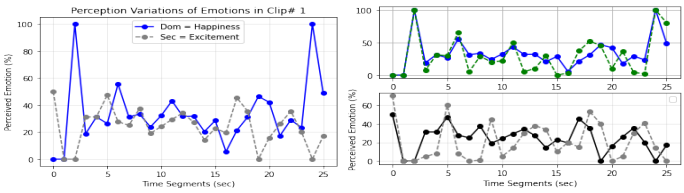
Results: The following are observed from this experiment: a) The mean test accuracies for the dominant and secondary emotion classifiers are calculated to be 0.68 and 0.67 respectively. b) EWIV representation can be used to classify *segments* into dominant or secondary emotions with good test accuracies. c) The best classification performance reported in this section is comparable to emotion classification results reported in literature [40] with a similar experimental setup. d) The single and double-layer models' performances are comparable.

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 ≥ 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 1. 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. For detailed results, please refer to Appendix B.

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). A random excerpt (Excerpt 1) is selected from the test clip-set of 4 clips - 2 happy and 2 sad for this purpose. 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 2(a) plots the ground truth variations of the *dominant*



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

Fig. 2: 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.

(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 2(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 6.1.3. It is observed that the *dominant* emotion prediction fares slightly better than the *secondary* emotion prediction.

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 significantly high ($\geq$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 4.1) - the ground truth. The Szymkiewicz–Simpson coefficient or Overlap Coefficient (OVL) is used for this comparison, which is given by $\frac{|A \cap B|}{min(|A|,|B|)}$, where A and B are two finite sets. This is reported in table 3 for the first 4 excerpts of the *EmoRaga* dataset, which consist of the test set. It is observed that the overlap coefficients are high ($\geq$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.

7 Comparison of EWIV with Circumplex

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 perform an information content-based comparison of the EWIV and the dimensional Circumplex [9] models over their *representativeness* of music perceived-emotions collected in relevant datasets. We use the Akaike information criterion (AIC) [30], which is a statistical measure of the suitability of a statistical model towards a given input data. AIC metric provides a tradeoff between 2 components: *goodness of fit* measured using log-likelihood of a dataset, and a measure of model complexity given by the number of parameters of the model. Due to this tradeoff, AIC provides a metric of *out-of-sample* prediction accuracy, or generalization ability of the statistical model to unseen data. We consider both EWIV and the Circumplex representations as probabilistic models for fitting the perceived emotions of the self-reported data points, and use the Akaike Information Criterion (AIC) [30] for the information content-based comparison.

To apply AIC for *goodness-of-fit* comparisons successfully, the primary requirement is the availability of perceived-emotion data using both formats, over a common set of excerpts, and listener participants. Such datasets are rare in existing literature, with the exception of the seminal work by Eerola et al [14], which studies both discrete and dimensional emotion representations over the same dataset. Moreover, collection of consistent self-reported emotions in two different representation format can cause cognitive overload on the subjects, and affect the reported emotions. To circumvent these problems and to ensure the broader applicability of this comparison in datasets where data of only one format is available, we propose to convert the available data from the existing format to the absent format. The conversion scheme could possibly be noisy, leading to some 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 7.1), and next (section 7.2), we consider different model estimations with respective AIC calculations and finally report the empirical results of the *representativeness* of four competing emotion representations over excerpts of three datasets.

7.1 Conversion of Representations

7.1.1 EWIV to Circumplex

In the Circumplex 2-D plane [9], each emotion term is associated with an angular value, indicating its location ([9], 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 3.1). Considering the angular value associated with emotion ε to be θ_ε [9], the corresponding valence v_ε and arousal a_ε values can be determined as: $v_\varepsilon = I_\varepsilon \cos\theta_\varepsilon$ and $a_\varepsilon = I_\varepsilon \sin\theta_\varepsilon$. 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:

$$V^c_{avg} = \frac{\sum_{\varepsilon \in \mathcal{E}} v_\varepsilon}{N^c} \quad \text{and} \quad A^c_{avg} = \frac{\sum_{\varepsilon \in \mathcal{E}} a_\varepsilon}{N^c} \tag{6}$$

TABLE 3: 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

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 (θ_ε^{min}) and a maximum (θ_ε^{max}) 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 [9]. Then, θ_ε^{min} and θ_ε^{max} are defined as: $\theta_\varepsilon^{min} = \theta_\varepsilon - \frac{(\theta_\varepsilon - \theta_{\varepsilon-1})}{2}$ and $\theta_\varepsilon^{max} = \theta_\varepsilon + \frac{(\theta_{\varepsilon+1} - \theta_\varepsilon)}{2}$. The *region* of emotion ε is thus demarcated by $[\theta_\varepsilon^{min}, \theta_\varepsilon^{max}]$, and can be further be sub-divided into 5 equal *sub-regions*, which we map to the five intensities of the present EWIV. This procedure is further detailed in Appendix E. Given an arousal-valence response (a_k, v_k), the angular coordinate of this point on the 2-D plane is given by $\theta_k = \tan^{-1}(\frac{a_k}{v_k})$. 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 unit norm.

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 7.2.1. Next, in section 7.2.2, we analyze and report the empirical results on three datasets.

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 3.2), can be considered a random sample from the following distribution:

$$EWIV_\varepsilon^{c,\lambda} \sim \mathcal{N}(\mu_{\varepsilon,c}, \sigma_{\varepsilon,c}) \tag{7}$$

Considering the 8 emotion words of EWIV (section 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: $\mu_{\varepsilon,c} = \frac{1}{|\mathcal{D}_{\varepsilon,c}|}\sum_{\lambda,n} EWIV_\varepsilon^{c,\lambda}(n)$ and $\sigma_{\varepsilon,c}^2 = \frac{1}{|\mathcal{D}_{\varepsilon,c}|}\sum_{\lambda,n}(EWIV_\varepsilon^{c,\lambda}(n) - \mu_{\varepsilon,c})^2$. The total log-likelihood for the model M is estimated as:

$$\ln(\hat{L}_c(M)) = \sum_\varepsilon \ln(\mathcal{L}(\mathcal{D}_{\varepsilon,c}|\mu_{\varepsilon,c}, \sigma_{\varepsilon,c})) \tag{8}$$

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 3.2). 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 8. For the Circumplex model [9], 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 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.

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.

a) *Eerola's Dataset* [14] (section 2.2.1): We would like to sincerely thank the authors 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 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 7.2.1). The results for the first five excerpts of *Eerola's Dataset* [14] are presented in table 4 (the rest in Appendix C). The columns *Excerpt#* and *Emotion Category 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

TABLE 4: AIC results for the first 5 excerpts from Eerola's Dataset [14]. 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)	$AIC_{EWIV+AV}$(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

TABLE 5: AIC results for the Soleymani_5 [15] 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)	$AIC_{EWIV+AV}$(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

TABLE 6: AIC results for the first 5 *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

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 (Appendix C). *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.

b) *Soleymani_5* Dataset (section 4.1): The original dataset provides data in the arousal-valence format, and EWIV data was collected for the purpose of this study (table 2). A similar procedure was followed for *Eerola's Dataset*. Data format conversions (section 7.1) were performed and AIC values were calculated (section 7.2.1) for each excerpt over the original, collected, converted, and integrated data The results are presented in table 5. *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 $AIC_{AV}(1)$). *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.

c) *EmoRaga* Dataset (section 4.1.2): The dataset provides EWIV format data, which was converted to dimensional format (section 7.1.1), and AIC values were calculated (section 7.2.1) for each excerpt over the collected and converted datasets. The results for the first five excerpts are presented in table 6 (rest in Appendix C). *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 $AIC_{AV}(3)$).

8 Discussion and Conclusion

In summary, 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-words* 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 the statistical goodness-of-fit measure of AIC.

Our future research will focus on (1) Extending the *EmoRaga* dataset to include more HCM excerpts, encompassing different prevalent *Raga, Tala, Laya* combinations while balancing the number of excerpts with respect to different *dominant* emotions; and (2) Building deep learning based emotion classification methods with more data and sophisticated set of features to improve the classification accuracy; (3) Exploring ambiguity in perceived emotion in music, and the factors contributing to it, and (4) Testing EWIV's applicability in other music styles.

REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *ICASSP*. IEEE, 2014, pp. 5412–5416.

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

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

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *ICASSP*. IEEE, 2015, pp. 693–697.

[32] 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.

[33] W. Yang, K. Makita *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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.
[39] 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.
[40] 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.
[41] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using kalman filtering," in *2010 Ninth International Conference on Machine Learning and Applications*. IEEE, 2010, pp. 655–660.
[42] S. Bagchee, *Nad*. BPI Publishing, 1998.
[43] S. Gulati, J. Serra, V. Ishwar, and X. Serra, "Discovering rāga motifs by characterizing communities in networks of melodic patterns," in *ICASSP*. IEEE, 2016, pp. 286–290.
[44] 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.
[45] K. K. Ganguli, S. Gulati, X. Serra, and P. Rao, "Data-driven exploration of melodic structure in hindustani music," in *Proceedings of the 17th ISMIR Conference; Aug 7-11; New York City (NY).p. 605-11.*, 2016.
[46] K. Narang and P. Rao, "Acoustic features for determining goodness of tabla strokes." in *ISMIR*, 2017, pp. 257–263.
[47] 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 ICMC/SMC*. Michigan Publishing, 2014.
[48] 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.
[49] 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*, 2014.
[50] BharataMuni, "Nātyasāstra of bharata. chapter six. rasādhyāyah on the sentiments: A commentary by abhinavagupta," Ph.D. dissertation, 1926.
[51] J. C. Ross, T. Vinutha, and P. Rao, "Detecting melodic motifs from audio for hindustani classical music." in *ISMIR*, 2012, pp. 193–198.
[52] P. Chordia and A. Rae, "Raag recognition using pitch-class and pitch-class dyad distributions." in *ISMIR*. Citeseer, 2007, pp. 431–436.
[53] V. Bhatkhande, "Hindustani sangeet paddhati: Kramik pustak maalika vol. i-vi," *Sangeet Karyalaya*, vol. 72, 1990.
[54] R. Jha, "Abhinav geetanjali vol. iv," *Sangeet Sadan*, vol. 72, 2001.
[55] 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.
[56] P. Dighe, H. Karnick, and B. Raj, "Swara histogram based structural analysis and identification of indian classical ragas." in *ISMIR*, 2013, pp. 35–40.
[57] 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.
[58] 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.
[59] S. Chaki, P. Doshi, S. Bhattacharya, and P. Patnaik, "Explaining perceived emotion predictions in music: An attentive approach," in *Proceedings of the 21st ISMIR*, 2020, pp. 150–156.
[60] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Dr. Sanga Chaki received the PhD degree with a specialty in Machine Learning for modelling and explaining perceived emotions in music from the Indian Institute of Technology, Kharagpur in 2023. Currently, she is an Assistant Professor of Computer Science and Engineering in ITER, SOA University, Bhubaneswar, India. Her research interests include explainable emotion recognition and affective computing. She is also a trained *Sarod* (HCM instrument) player and practicing musician.

Dr. Sourangshu Bhattacharya is an Associate Professor, Computer Science and Engineering, IIT Kharagpur. He is broadly interested in Machine Learning, with specific interests in Explainability and Data-centric AI, Multi-task Learning, Learning with Temporal Point Processes, Network Representation Learning, and Scalable Machine Learning.

Ms. Junmoni Borgohain is pursuing her PhD from IIT Kharagpur. She is also working as an Assistant Professor of Psychology at KIIT University, Bhubaneswar, India. Her research interests include music cognition, embodiment in spiritual music and happiness and wellbeing.

Dr. Priyadarshi Patnaik is a Professor of English and Communication at the Department of Humanities & Social Sciences and the coordinator of the Rekhi Centre of Excellence for the Science of Happiness, IIT Kharagpur. His areas of research include Visual communication, Music and emotion, Communication and culture, Media communication, Translation, and Digital humanities. He is a trained flute player.

Mr. Raju Mullick received MS degree from the Indian Institute of Technology, Kharagpur 2021. His research interests included musicology, exploration of perception of emotions, ornamentations, and other HCM features in listeners, along with music cognition, and happiness and wellbeing

Ms. Gouri Karambelkar is currently a PhD scholar at the Indian Institute of Technology, Kharagpur. She is also trained HCM vocalist and a practising musician. Her academic research interests include music psychology, music history and musical aesthetics in Hindustani classical music.