

JOYFUL FOR YOU AND TENDER FOR US: THE INFLUENCE OF INDIVIDUAL CHARACTERISTICS AND LANGUAGE ON EMOTION LABELING AND CLASSIFICATION

Juan Sebastián Gómez-Cañón¹ Estefanía Cano²
Perfecto Herrera¹ Emilia Gómez^{3,1}

¹ Music Technology Group, Universitat Pompeu Fabra, Spain

² Songquito UG, Erlangen, Germany

³ European Commission, Joint Research Centre, Seville, Spain

juansebastian.gomez@upf.edu

ABSTRACT

Tagging a musical excerpt with an emotion label may result in a vague and ambivalent exercise. This subjectivity entangles several high-level music description tasks when the computational models built to address them produce predictions on the basis of a "ground truth". In this study, we investigate the relationship between emotions perceived in pop and rock music (mainly in Euro-American styles) and personal characteristics from the listener, using language as a key feature. Our goal is to understand the influence of lyrics comprehension on music emotion perception and use this knowledge to improve Music Emotion Recognition (MER) models. We systematically analyze over 30K annotations of 22 musical fragments to assess the impact of individual differences on agreement, as defined by Krippendorff's α coefficient. We employ personal characteristics to form group-based annotations by assembling ratings with respect to listeners' familiarity, preference, lyrics comprehension, and music sophistication. Finally, we study our group-based annotations in a two-fold approach: (1) assessing the similarity within annotations using manifold learning algorithms and unsupervised clustering, and (2) analyzing their performance by training classification models with diverse "ground truths". Our results suggest that a) applying a broader categorization of taxonomies and b) using multi-label, group-based annotations based on language, can be beneficial for MER models.


1. INTRODUCTION

Several studies suggest that the main reason people engage with music is its emotional effect [1–3]. This makes the idea of computational algorithms that can "predict" the

emotions in music particularly intriguing and provocative. These algorithms evaluate emotionally relevant acoustic features from the audio signals, and correlate them with certain emotions that the music could convey, express or induce. Recently, deep learning approaches have been used to further improve emotion recognition [4–6]. However, the performance of these algorithms may have reached a "glass ceiling" possibly due to the subjective nature of the perception of emotions [7, 8], the limited agreement in the annotations of datasets [9–11], the lack of an agreed methodology for annotation gathering [2], the generalized confusion between perceived and induced emotions [3, 12], amongst other reasons. In particular, the low agreement problem extends to other high-level description tasks in Music Information Retrieval (MIR) such as music auto-tagging [13], music genre recognition [14, 15], music similarity [9, 11], and even computationally well-defined tasks like automatic chord estimation [16] and beat tracking [17]. Given the importance of annotation collection and in an attempt to improve their quality, we address two research questions in this paper: *RQ1* - Do personal characteristics and mother tongue have an influence on the annotation of perceived emotions for listeners? *RQ2* - Can this information be used to improve MER algorithms? The rest of the paper is structured as follows: Section 2 reviews basic definitions and previous work, in Section 3 we detail the methodology of our study, including annotation gathering, clustering, and classification schemes. Section 4 provides results of our study which are later discussed in Section 5.

2. RELATED WORK

In this study, we focus on the *perception* of emotions as a key factor in the "musical communication" between music itself and a listener. *Perceived* emotions refer to those recognized by the listener through the interpretation of musical properties [3]. In contrast, *induced* emotions concern the arousal of psycho-physiological responses [18].¹ The relation between musical properties and emotion perception has been widely researched in the literature [3, 12]:

 © Juan Sebastián Gómez-Cañón, Estefanía Cano, Perfecto Herrera, Emilia Gómez. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Juan Sebastián Gómez-Cañón, Estefanía Cano, Perfecto Herrera, Emilia Gómez, "Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020.

¹ For a review of the emotivist-cognitivist argument, refer to [19, 20].

happiness is linked to fast mean tempo, bright timbre, sharp duration contrasts; *sadness* is related to slow mean tempo, low sound level, dull timbre, slow vibrato; *fear* is linked with very low sound level, large sound level variability, large timing variations.

From music psychology, two dominant views or taxonomies for emotion representation prevail [21]:

- *Categorical approach*: emotions are represented as categories, distinct from each other - such as happiness, sadness, anger, and fear [22]. The major drawbacks of this approach are: (1) the amount of categories results too small compared to the richness of human emotion, and (2) the ambiguity of using language during self-reports [7].
- *Dimensional approach*: emotions are conceptualized based on their positions on a few dimensions, mainly arousal and valence (AV). Russell popularized the two-dimensional circumplex model, where the valence dimension describes the pleasantness or positiveness of the emotion, and the arousal dimension describes the activation or energy (i.e., *happiness* would have positive arousal and valence) [23]. However, the major flaw is that categories are not mutually exclusive and tend to overlap (i.e., rage-anger), making the mapping of categories on the dimensional space vague and unreliable [7].

In the particular case of instrumental classical music, Schedl et al. explored the relationship between listeners' characteristics and nine categories of emotion on segments from the 3rd Symphony *Eroica* by L.V. Beethoven [10]. Their results suggest that: (1) the perception of transcendence and power correlates significantly with affinity to classical music; (2) participants trained on classical music tend to disagree more on perceived emotions of peacefulness, tension, sadness, anger, disgust, and fear; (3) the agreement among perceived emotions decreases with increasing familiarity with the piece. Our work is based on this study, however we focus on music with lyrics of pop and rock style. Very few studies have explored different styles of music and researchers report that 50% of music and emotion studies focus on classical music [2,24]. When it comes to annotation reliability, researchers have studied ways of increasing inter- and intra-rater agreement for music similarity. Flexer and Lallai found evidence that upper bounds for inter-rater agreement (i.e., measured between different subjects) cannot be increased for this task, while the intra-rater case can be improved (i.e., measured on ratings from the same subject at different time) [11]. We base our research on increasing inter-rater agreement by analyzing listeners with similar characteristics and assembling group-based annotations based on listeners characteristics. Group-based MER has been attempted by Yang et al. by assembling annotations according to cultural factors, music experience, and personality traits [7, 25]. Their results suggest insignificant improvement for the regression task as compared to using general averaged annotations. However, and to the extent of our knowledge, this is the first

work that uses language and self-reported lyrics comprehension to group annotations of perceived emotion.

3. METHODOLOGY

The main contribution of our work is to address open questions from previous studies by focusing on rock and pop music. We use these musical styles since they appear to be similar across different cultures and are musically homogeneous, even when sung in different languages. Contrary to most studies, we focus on a small set of songs with existing emotion annotations and gather large-scale, diverse annotations per song from participants of different mother tongues. The goal of our work is to study the relationship between listeners' demographics, preference, familiarity, musical knowledge, and native language with agreement of perceived emotions in music. In order to achieve this, we use these personal characteristics to form group-based annotations and analyze the agreement of participants belonging to these groups. We then perform manifold learning and K-means clustering to study if group-based annotations yield representations that are more similar amongst them. Finally, we compare the performance of well-known classifiers trained using these annotations, in order to analyze the impact of grouping variants on the MER task.

3.1 Emotion annotation gathering

Our pool of annotators were presented with surveys designed with PsyToolkit [26] in four languages: Spanish, English, German and Mandarin. The survey was structured as follows: (1) collection of general demographic information (age, gender, country of origin and formation, and language), (2) volume adjustment task, (3) explanation of the difference between *perceived* and *induced* emotions, (4) random presentation and annotation of excerpts with a 5-point Likert response format per emotion, and (5) the Music Sophistication Index self-report inventory [27]. In (2), each user was asked to set the volume w.r.t. a 1 KHz sinusoid making it barely audible. In (4), we used synonyms for each emotion for clarity, which were validated by native speakers from each language, following [10]. For each excerpt, we collected information about the listeners' preference, familiarity, and understanding of the lyrics.²

We selected a set of excerpts from the 4Q emotion dataset [28], which was previously annotated with categories in the four arousal-valence (AV) quadrants: Q1 (positive valence and arousal, A+V+), Q2 (positive arousal and negative valence, A+V-), Q3 (negative valence and arousal, A-V-), Q4 (negative arousal and positive valence, A-V+). In order to gather annotations on a larger scale, we asked participants to rate the excerpts with the following emotion categories: Q1 - *joyful activation, power, surprise*, Q2 - *anger, fear, tension*, Q3 - *bitterness, sadness*, Q4 - *peace, tenderness, transcendence*. These emotion adjectives were selected from the Geneva Emotion Music Scale (GEMS) [29] and a subset of basic emotions [30]. To select the target songs, we queried for these adjectives in

² Refer to Figure 1 - supp. mat. for the annotation interface.

English on the datasets’ metadata. Since some words were not found, synonyms were used for certain emotions (e.g., fear - anguished).³ From the resulting excerpts, we randomly selected two excerpts per emotion for a total of 22 fragments. All audio excerpts were normalized from -1 to 1 in amplitude, in order to balance the volume during playback. We collected additional audio features from Spotify API [31]: beats per minute, energy [0,1], and valence [0,1].

3.2 Emotion agreement analysis

Following [4, 10, 16], we used inter-rater reliability statistics to assess the agreement of the annotated data with respect to personal characteristics [32]. Researchers have found that Krippendorff’s α is used increasingly to assess reliability in content analysis methodologies [33]. We employed Krippendorff’s coefficient α defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (1)$$

where D_o is the measure of observed disagreement:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \cdot metric \delta_{ck}^2 \quad (2)$$

and D_e is a measure of the expected disagreement given chance:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_k \cdot metric \delta_{ck}^2 \quad (3)$$

The variables o_{ck} , n_c , n_k are the frequencies of values of observed coincidences of c and k values or ranks, and n is the total amount of paired $c - k$ values or ranks. Advantages of using α are: the suitability for any number of observers, handling of any type of metric (nominal, interval, ordinal), handling incomplete or missing data, and not requiring a minimum of sample size. When disagreement is absent ($D_o = 0$), there is perfect reliability ($\alpha = 1$). Conversely, when agreement and disagreement are a matter of chance ($D_e = D_o$), there is absence of reliability ($\alpha = 0$). Nevertheless, α could be smaller than zero if the sample size is too small or agreement below what would be expected by chance. According to [32], data with $\alpha \geq 0.8$ is considered to have good agreement and $0.4 \leq \alpha \leq 0.667$ shows fair agreement. Finally, $metric \delta_{ck}^2$ represents the difference function: the squared difference between any two values or ranks c and k , depending on the data gathering approach (in our case, we use an ordinal metric).

To obtain annotation groupings, we defined positive and negative filters to classify ratings based on different user responses and characteristics. Considering the 5-point Likert response format, we define a positive filter by keeping ratings higher than 3 (neither agree or disagree) and a negative filter by keeping those less than 3. These filters were used to form groups using the users’ response of preference, familiarity, and understanding for each excerpt. In the case of behavioral factors of music sophistication, we

³ Note that query synonyms in English differ to the description synonyms used and translated for each survey, refer to Table 1 (supp. mat.).

specified positive and negative filters by grouping the annotations of participants with higher and lower scores than the population mean, respectively. We collected six behavioral factors, yet used the ones in **bold** to group ratings: Active Engagement, Perceptual Abilities, **Musical Training, Emotion Perception**, Singing Abilities, and **General sophistication**.⁴

In order to evaluate the collected annotations, we clustered group-based ratings in 2D and 3D spaces. Our intuition is that group-based annotations are more similar amongst them, and that clusters obtained with these annotations show less variance than those obtained with the original "ground truth". We generated a low-dimensional representation of all annotations with manifold learning, used one of the proposed filters to keep a group, clustered the resulting embeddings, and compared the resulting clusters with the original "ground truth" (i.e., four quadrants of emotion). We use Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) for all filters as measures of quality of the clusters. We proceeded as follows: (1) standarization of the annotations, (2) categorical Principal Component Analysis (CATPCA), (3) manifold learning for dimensionality reduction, and (4) clustering of embeddings into the four quadrants of emotion using K-means (k-means++ initialization [34]). In (2), we use 10 components retaining 97.4% from variance.⁵ In (3), we used the following algorithms: Multi-dimensional Scaling (MDS) [35], t-distributed Stochastic Neighbor Embedding (t-SNE) [36], and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [37].⁶

3.3 Emotion classification approach

Despite the low number of data instances (22), we trained support vector machine (SVM) classifiers to study how group-based annotations compare to annotations from all participants. Following relevant literature [4, 38, 39], we extracted the IS13 ComParE feature set with the openSMILE toolbox [40]: 260 low-level features (mean and standard deviation of 65 emotionally-relevant acoustic descriptors, and their first order derivatives) with a frame size of 60 ms and a 10 ms hop size. These features are widely used as a benchmark for speech and music emotion recognition tasks [4]. Each feature vector was aggregated in segments of 5 seconds with a 75% overlap, resulting in 24 feature vectors per excerpt and 528 samples in total. We performed standarization over each feature and PCA for dimensionality reduction. After a Scree test, we selected 8 components that retained 65.7% of the variance. We performed a grid search for parameter optimization with the following settings (final parameters in **bold**): regularization parameter C [0.001, 0.01, 0.1, 1, **10**] and radial ba-

⁴ Refer to Table 2 - supp. mat. for music sophistication results.

⁵ Refer to Figure 2 - supp. mat. for an example of CATPCA.

⁶ We used different settings for each algorithm (final parameters in **bold**): (1) MDS - **metric** and non-metric, iterations [300, 1000, **3000**], epsilon [**1e-1**, 1e-3, 1e-9, 1e-12], (2) t-SNE - perplexity [3, 5, 10, 30, 50, 100], learning rate [**200**, 500, 1000], number of iterations [**1000**, 3000], (3) UMAP - minimum distance [0.1, 0.25, 0.5, **0.8**, 0.99], number of neighbors [**10**, 30, 100, 200], metric [**Euclidean**, Cosine and Chebyshev].

Configuration	Ratings	%	Q1			Q2			Q3			Q4	
			joy	surp.	pow.	ang.	fear	tens.	sad	bit.	pea.	tend.	tran.
All	27720	100.00%	0.393	0.064	0.273	0.335	0.169	0.253	0.340	0.215	0.357	0.396	0.062
By Preference (>3)	11275	40.67%	0.402	0.074	0.275	0.264	0.142	0.171	0.367	0.204	0.363	0.414	0.066
By Preference (<3)	8976	32.38%	0.325	0.036	0.267	0.368	0.158	0.309	0.308	0.191	0.352	0.384	0.049
By Familiarity (>3)	4477	16.15%	0.445	0.065	0.199	0.314	0.223	0.158	0.376	0.284	0.266	0.297	0.038
By Familiarity (<3)	21439	77.34%	0.319	0.047	0.275	0.329	0.145	0.270	0.304	0.173	0.385	0.414	0.070
By Understanding (>3)	13827	49.88%	0.428	0.074	0.276	0.269	0.160	0.195	0.361	0.250	0.324	0.365	0.046
By Understanding (<3)	9977	35.99%	0.328	0.044	0.263	0.361	0.156	0.299	0.308	0.159	0.361	0.391	0.070
By Music Training (> μ)	12320	44.44%	0.406	0.074	0.327	0.407	0.190	0.268	0.364	0.235	0.424	0.456	0.067
By Music Training (< μ)	15400	55.56%	0.381	0.052	0.231	0.283	0.151	0.236	0.318	0.194	0.306	0.348	0.054
By Emotion (> μ)	16500	59.52%	0.430	0.058	0.299	0.381	0.212	0.288	0.378	0.260	0.411	0.426	0.068
By Emotion (< μ)	11220	40.48%	0.343	0.067	0.230	0.273	0.114	0.206	0.287	0.149	0.289	0.351	0.046
By General Sophistication (> μ)	13640	49.21%	0.415	0.083	0.319	0.400	0.195	0.274	0.357	0.251	0.441	0.466	0.091
By General Sophistication (< μ)	14080	50.79%	0.371	0.043	0.225	0.275	0.147	0.226	0.327	0.180	0.283	0.331	0.038

Table 1. Krippendorff’s α for each emotion for all participants filtered by preference, familiarity, lyrics comprehension, and music sophistication (positive and negative) using only music with lyrics (17 in English and 3 in Spanish).

sis function kernel with coefficient gamma [0.001, 0.01, 0.1, 1]. We report precision, recall, and F1-score using 5-fold cross-validation to evaluate the models. We test three possible "ground truths" per excerpt: (1) single-label annotations from the original metadata (MD), (2) multi-label annotations from all participants (All), and (3) multi-label annotations from participants belonging to a group (Filtered). In cases (2) and (3), we summarized ratings by taking the statistical mode of each emotion rating across participants. We employ the mode as some ratings showed a bimodal distribution and our annotations were categorical. Since the mode may result in multiple maximum values, we created multi-label annotations (i.e., an excerpt can have a mode of 4 for both *anger* and *tension*). Anonymized data and evaluation code are available online.⁷

4. RESULTS

4.1 Emotion annotation and agreement analysis

The participation was unbalanced regarding languages: a total of 126 (65 Male, 61 Female, $M = 34.12$ years, $SD = 11.75$) participants completed all tasks in our experiment from English ($n = 26$), Spanish ($n = 56$), Mandarin ($n = 27$), and German ($n = 17$) surveys. Listeners that wanted to participate in the survey but were not native to any of the languages were asked to take the English

⁷ <https://github.com/juansgomez87/agreement-emotion>

Emotions	Eng. (26)	Spa. (56)	Man. (27)	Ger. (17)	All (126)
anger	0.429	0.311	0.367	0.482	0.364
bitter	0.278	0.209	0.155	0.278	0.202
fear	0.241	0.175	0.091	0.207	0.171
joy	0.304	0.437	0.311	0.476	0.372
peace	0.401	0.332	0.401	0.438	0.371
power	0.379	0.287	0.296	0.325	0.289
sad	0.330	0.343	0.279	0.378	0.326
surprise	0.041	0.055	0.068	0.218	0.075
tender	0.444	0.314	0.452	0.581	0.396
tension	0.264	0.324	0.282	0.323	0.296
transc.	0.080	0.049	0.083	-0.012	0.057

Table 2. Krippendorff’s α for each emotion across different languages.

version ($n = 15$). Since the surveys were made available through different channels, listeners were asked to state the country in which they spent the formative years of childhood and youth⁸. We evaluated outliers for every musical fragment, finding that the participants that systematically annotated outside the interquartile range (Q1-Q3) did so at most 10.33% of the time. Hence, we decided to keep all ratings for analysis. Median and Kruskal-Wallis H tests showed that there was a statistically significant difference in the ratings of emotions between raters from the surveys of each language ($p < 0.01$).⁹ Concretely, the ratings of *anger*, *bitterness*, *fear*, *sadness*, *surprise*, *tenderness*, *tension*, and *transcendence* have different distributions across the surveys. This suggests that emotion significance varies across cultures and languages as hypothesized in our study. Emotion adjectives might have varied meanings and associations across different languages and cultures as studied by [41], but results should be validated with a higher amount of participants and excerpts.

Results from the agreement analysis are presented in Tables 1 and 2, and appear **bold** when the agreement of the emotion is higher than 0.05 as to the agreement measured across all participants. Conversely, the text in *italic* indicates a difference less than -0.05, following [10]. Table 1 shows agreement over all participants that belong to a certain group (see Configuration column). Hence, it contains information about the number of ratings for the corresponding filters (groups). The music selection contained 20 songs with lyrics (17 in English and 3 in Spanish), thus agreement was analyzed only for this subset for lyrics comprehension to be meaningful. We find two tendencies for groups assembled with preference, familiarity, and lyrics comprehension: (1) positive filters will result in higher agreement with respect to all ratings for emotions such as *joy*, *surprise*, *power*, *sadness*, and *bitterness*; (2) positive filters will result in lower agreement for emotions such as *anger*, *fear*, *tension*, *peace*, *tenderness* and *transcendence*. Interestingly, emotions in (1) belong to

⁸ Subjects participated from the following countries: (1) *Spanish* - Bolivia, Colombia, Ecuador, Perú, Spain, Uruguay, (2) *Mandarin* - Mainland China, Taiwan, (3) *German* - Austria, Germany, Switzerland, and (4) *English* - Australia, Bulgaria, Belgium, Brazil, France, Greece, India, Italy, Portugal, Romania, United Kingdom, United States.

⁹ Refer to Table 3 - supp. mat. for multiple pairwise test results.

quadrants Q1 (A+V+) and Q3 (A-V-), while emotions in (2) belong to Q2 (A+V-) and Q4 (A-V+). Regarding musical sophistication, we find a consistent trend in groups assembled with musical training, emotion perception and general sophistication: positive filters will result in higher agreement in all filters with respect to all ratings. Table 2 shows agreement over all participants from each survey and exposes low agreement over certain emotions: *bitterness*, *fear*, *power*, *surprise*, and *transcendence*. On the other hand, a higher agreement is reached for emotions such as *anger*, *joy*, *peace*, *sadness*, and *tenderness*. The results confirm that participants of different surveys show significant differences in the emotions perceived from music. Additionally, we found positive linear correlations, as obtained by Pearson’s coefficient, between *anger*, *bitterness*, *fear*, and *tension*; *peace* and *tenderness*; *joy*, *power*, and *surprise*; *sadness* and *bitterness*.¹⁰

4.2 Implications for MER models

Table 3 shows the clustering evaluation when comparing clusters from manifold learning representations with the original "ground truth". Scores are reported in **bold** when they are higher than the scores obtained without filters (All). In every case, positive filters result in improved clusterability (particularly when selecting participants with high scores for music sophistication). MDS and UMAP appear to separate the data better than t-SNE before performing K-means. As a baseline, applying K-means on the raw data shows that manifold learning techniques can be useful to find similarities between group-based ratings. We argue that using manifold learning previous to clustering extracts possible similarities across annotations that belong to a given AV quadrant, yielding annotation embeddings that are easier to cluster.¹¹

	MDS + K-Means		t-SNE + K-Means		UMAP + K-Means		K-Means Raw data	
	ARI	AMI	ARI	AMI	ARI	AMI	ARI	AMI
All	0.248	0.255	0.191	0.214	0.244	0.257	0.221	0.229
Pref. (>3)	0.267	0.282	0.195	0.230	0.252	0.271	0.197	0.223
Pref. (<3)	0.234	0.234	0.184	0.206	0.244	0.248	0.215	0.234
Fam. (>3)	0.301	0.319	0.252	0.264	0.295	0.305	0.223	0.251
Fam. (<3)	0.231	0.241	0.183	0.202	0.236	0.245	0.218	0.220
Und. (>3)	0.253	0.260	0.225	0.239	0.250	0.263	0.213	0.230
Und. (<3)	0.235	0.250	0.187	0.210	0.244	0.243	0.209	0.219
MT (> μ)	0.283	0.290	0.210	0.242	0.302	0.309	0.286	0.289
MT (< μ)	0.214	0.227	0.180	0.203	0.207	0.220	0.172	0.185
Emo. (> μ)	0.303	0.305	0.216	0.242	0.297	0.309	0.262	0.271
Emo. (< μ)	0.186	0.199	0.159	0.176	0.179	0.196	0.151	0.162
GS (> μ)	0.303	0.303	0.221	0.245	0.291	0.303	0.307	0.308
GS (< μ)	0.202	0.217	0.169	0.193	0.202	0.219	0.155	0.165

Table 3. Clustering metrics for all filters and manifold learning algorithms. MT refers to Musical Training, Emo. to Emotion Perception, GS to General Sophistication. The last column shows clustering results on the raw data.

An example of the group-based, multi-label annotations produced by using the positive understanding filter can be seen in Figure 1. This plot compares different "ground

truths" for the data according to the collected ratings and a given filter that we used to train our classifiers. For example, excerpt 0 (originally labeled as *anger*) is also labeled with *bitterness*, *fear*, *power*, and *tension* when considering our annotations (top-right plot). In contrast, excerpt 1 is labeled as *power* when considering all ratings, but labeled as *anger*, *bitter*, and *power* when considering the filter (comparison of top- and bottom-right plots).

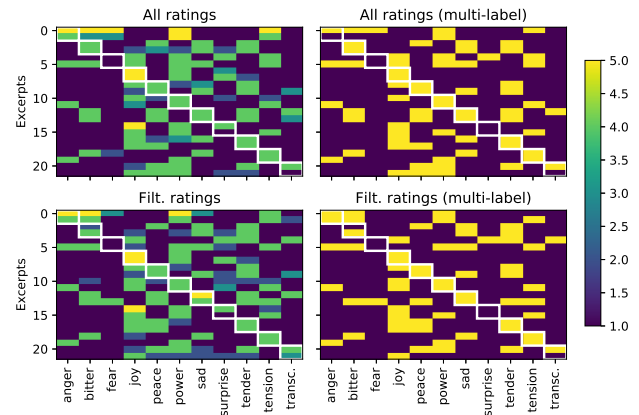


Figure 1. Example of annotations using *All* ratings (top row) and the *positive understanding* filter (bottom row). White rectangles highlight the original annotation from the metadata (MD - single-label). For every plot, rows represent one of the 22 excerpts and columns represent an emotion. The plots on the left show the mode over the participants for each emotion. The plots on the right show selected multi-labels for each excerpt used for classification. The color bar represents the 5-point Likert scale.

Classification experiments are reported in Table 4, including precision (P), recall (R) and F1-Score (F) of classifiers trained on annotations with different filters and 5-fold cross-validation. We compare three settings: original labels from the metadata (MD - single-label), annotations collected from all raters (All - multi-label), and annotations from selected raters with respect to the defined groups (Filter - multi-label). Comparisons are presented by subtracting the mean performance scores of classifiers trained on two annotation scenarios. For example, All - MD refers to the comparison between a classifier trained on all annotations and one trained on the original metadata. Scores are reported in **bold** when the difference is greater than 0.05 and in *italic* when it is less than -0.05. We also report the Jaccard coefficient (JC), bounded from [0,1], to estimate the similarity of the compared ratings in each case [42]. JC shows that the similarity from the original "ground truth" and the collected annotations is low (i.e., All - MD and Filt. - MD), which is expected since we compare single- and multi-label annotations. Interestingly in these two cases, classification results from the collected annotations (multi-label) provide consistent mean improvements of 15.01 percent points in precision and 11.8 in F1-scores with respect to classifiers trained on the original "ground truth" (single-label). We argue that improvements are due to the correlations between tags that are exploited in the multi-label

¹⁰ Refer to Table 4 - supp. mat. for full correlation analysis.

¹¹ Refer to Figures 3-5 - supp. mat. for visual sample embeddings.

case. In the case of comparing filtered and all collected annotations (i.e., Filt. - All), we find that a classifier trained on the group-based annotations generated from the positive understanding filter consistently results in better classification for this dataset. In this case, both classifiers were trained on multi-label annotations. In contrast, other filtered group-based annotations result in very similar performance as with all annotations, confirming previous findings from Yang [7].

Comparison	Filter	JC	ΔP	ΔR	ΔF
All - MD	-	0.287	0.171	0.056	0.124
Filt. - MD	Pref. (>3)	0.286	0.210	0.061	0.134
	Pref. (<3)	0.326	0.055	-0.086	-0.008
	Fam. (>3)	0.236	0.171	0.069	0.150
	Fam. (<3)	0.297	0.159	-0.026	0.038
	Und. (>3)	0.322	0.228	0.130	0.179
	Und. (<3)	0.284	0.051	-0.093	-0.017
	MT (> μ)	0.253	0.198	0.008	0.100
	MT (< μ)	0.314	0.084	-0.056	0.019
	Emo. (> μ)	0.314	0.144	0.015	0.091
	Emo. (< μ)	0.255	0.121	-0.025	0.041
	GS (> μ)	0.308	0.133	-0.037	0.037
	GS (< μ)	0.282	0.078	-0.002	0.053
Filt. - All	Pref. (>3)	0.718	0.039	0.005	0.010
	Pref. (<3)	0.639	-0.116	-0.142	-0.132
	Fam. (>3)	0.547	0.001	0.013	0.026
	Fam. (<3)	0.851	-0.011	-0.082	-0.086
	Und. (>3)	0.783	0.057	0.074	0.056
	Und. (<3)	0.697	-0.120	-0.149	-0.141
	MT (> μ)	0.861	0.027	-0.048	-0.024
	MT (< μ)	0.794	-0.087	-0.113	-0.105
	Emo. (> μ)	0.898	-0.026	-0.041	-0.033
	Emo. (< μ)	0.767	-0.050	-0.081	-0.083
	GS (> μ)	0.842	-0.038	-0.093	-0.087
	GS (< μ)	0.844	-0.093	-0.058	-0.070

Table 4. Performance comparison of models trained with different "ground truths". We report the difference of mean Precision (P), Recall (R), and F1-Score (F1). MD refers to metadata (single-label) and JC refers to Jaccard coefficient.

5. DISCUSSION AND CONCLUSION

In this paper, we systematically evaluated agreement of categorical annotations of emotions in 22 fragments of music. We characterized listeners by language, preference, familiarity, lyrics comprehension, and music sophistication.

With respect to *RQ1* - Do personal characteristics and mother tongue have an influence on the the annotation of perceived emotions for listeners? - our main finding is that there are substantial differences in the annotations of our surveys. In fact, the collected annotations show different distributions in the majority of emotions, and only the distributions of *joy* and *peace* appeared to be similar across languages. This relates to recent research on "colexification" of semantically related emotion concepts, where researchers found evidence that the relationship between emotion words varies significantly across languages [41]. Our results have also confirmed that certain basic emotions have higher agreement, while complex ones show the opposite. However, agreement in our experiment ap-

pears to be lower than values reported in [4] and similar to [10]. Our results advocate for taking into account diverse languages while gathering annotations and reducing the number of categories when dealing with cross-cultural MER models (i.e., four quadrants in AV space). Our findings suggest that preference, familiarity, and lyrics comprehension increase agreement for emotions corresponding to quadrants Q1 and Q3, and decreases it for quadrants Q2 and Q4. Regarding music sophistication, positive filters result in higher agreement for all emotions, conflicting with results from Schedl et al. [10]. We argue that in the case of classical music, music experts could tend to disagree more on subtle musical expression cues, while pop and rock music have stronger indicators for emotion (tempo, musical instruments, and meaning of lyrics). This has given us new understanding of the effect of language and lyrics comprehension: in the case of Q1 (A+V+) and Q3 (A-V-) higher agreement is found, contrasted to Q2 (A+V-) and Q4 (A-V+) where dimensions have opposite signs.

As to *RQ2* - Can this information be used to improve MER algorithms? - we find that models trained on multi-label annotations (all and filtered) will consistently show higher precision and F1-score than models trained with the original annotations from metadata (single-label) for this particular dataset. Our results show increments up to 18 percentage points in F1-Scores, when comparing single- and multi-labeled "ground truths". As to models trained with all collected annotations and our proposed filters (Filt. - All), we only find consistent gains for the case of positive lyrics comprehension. Nonetheless, further research is needed in order to confirm these findings. We propose four recommendations when creating datasets for MER algorithms with cross-cultural applications: (1) previous selection of listeners' population and music style have a deep impact on the agreement of annotations - good understanding of the population of annotators is required; (2) inter-rater reliability is crucial to define categories - agreement should be reported and analyzed; (3) group-based annotations can lead to improved agreement - models should be evaluated with both average ratings and group-based ratings; and (4) selecting annotators that are proficient in the language sung in the music may result advantageous - understanding the semantic content of lyrics could help increase the agreement in annotations and possibly lead to improving models. As future work, we consider balancing the styles with respect to different languages. It is also arguable that pop and rock are in fact musically homogeneous, since several variations across the world show different ways of conveying emotions (e.g., Hindi pop). Lastly, the experiment could have biased responses when asking for lyrics comprehension, forcing the participants' attention on lyrics and compromising ecological validity. Different studies regarding lyrics intelligibility should be taken into account in future research, such as [24, 43, 44]. Nevertheless, our study attempts to dispute Henry Wadsworth Longfellow's famous quote - is in fact music the universal language of mankind?

6. ACKNOWLEDGEMENTS

The research work conducted in the Music Technology Group at the Universitat Pompeu Fabra is partially supported by the European Commission under the TROMPA project (H2020 770376). We would like to thank participants to our surveys; Rafael Caro, Robert Graefe, Christopher Kenerski, and Tiange Zhu for help translating and distributing our surveys; and anonymous reviewers that helped us improve the paper with constructive feedback.

7. REFERENCES

- [1] P. Juslin, “Music and emotion: Seven questions, seven answers,” *Music and the mind: Essays in honour of John Sloboda*, pp. 113–35, 2011.
- [2] T. Eerola and J. K. Vuoskoski, “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.
- [3] P. N. Juslin, *Musical Emotions Explained*, 1st ed. Oxford: Oxford University Press, 2019.
- [4] A. Aljanaki, Y.-H. Yang, and M. Soleymani, “Developing a benchmark for emotional analysis of music,” *PLoS 1*, pp. 1–22, 2017.
- [5] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, “Towards Explainable Music Emotion Recognition: The Route via Mid-level Features,” in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 237–243.
- [6] K. W. Cheuk, Y.-J. Luo, G. Roig, and D. Herremans, “Regression-based Music Emotion Prediction using Triplet Neural Networks,” in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2020.
- [7] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*. CRC Press, 2011.
- [8] E. B. Lange and K. Frieler, “Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automated Features,” *Music Perception: An Interdisciplinary Journal*, vol. 36, no. 2, pp. 217–242, 2018.
- [9] A. Flexer and T. Grill, “The Problem of Limited Inter-rater Agreement in Modelling Music Similarity,” *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.
- [10] M. Schedl, E. Gómez, E. S. Trent, M. Tkalcic, 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, 2018.
- [11] A. Flexer and T. Lallai, “Can we increase inter- and intra-rater agreement in modeling general music similarity?” in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 494–500. [Online]. Available: <https://www.music-ir.org/mirex/wiki/2006>:
- [12] P. N. Juslin, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford: Oxford University Press, 2010.
- [13] E. Bigand and J.-J. Aucouturier, “Seven problems that keep MIR from attracting the interest of cognition and neuroscience,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 483–497, 2013.
- [14] B. L. Sturm, “Evaluating music emotion recognition: Lessons from music genre recognition?” in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, San Jose, USA, 2013, pp. 1–6.
- [15] —, “A Simple Method to Determine if a Music Information Retrieval System is a Horse,” *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 16, no. 6, 2014.
- [16] H. V. Koops, W. Bas De Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, “Annotator subjectivity in harmony annotations of popular music,” *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019. [Online]. Available: <https://www.tandfonline.com/action/journalInformation?journalCode=nnmr20>
- [17] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, “Selective sampling for beat tracking evaluation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012.
- [18] C. L. Krumhansl, “An exploratory study of musical emotions and psychophysiology,” *Canadian Journal of Experimental Psychology*, vol. 51, no. 4, pp. 336–353, 1997.
- [19] L. B. Meyer, *Emotion and meaning in music*. Chicago University Press, 1956.
- [20] P. Kivy, *Music alone: Reflections on a purely musical experience*. Cornell University Press, 1990.
- [21] 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.
- [22] K. Hevner, “Experimental studies of the elements of expression in music,” *American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.
- [23] J. A. Russell, “A Circumplex Model of Affect,” *Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

- [24] D. Vidas, R. Calligeros, N. L. Nelson, and G. A. Dingle, "Development of emotion recognition in popular music and vocal bursts," *Cognition and Emotion*, pp. 1–14, 2019.
- [25] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music Emotion Recognition: The Role of Individuality," National Taiwan University, Tech. Rep., 2007.
- [26] G. Stoet, "Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.
- [27] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population," *PLoS ONE*, vol. 9, no. 2, p. 89642, 2014.
- [28] R. Panda, R. M. Rui, and P. Paiva, "Musical texture and expressivity features for music emotion recognition," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.
- [29] K. R. Scherer, "Expression of Emotion in Voice and Music," *Journal of Voice*, vol. 9, no. 3, pp. 235–248, 1995.
- [30] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [31] Spotify, "Get audio features for a track," 2020, <https://developer.spotify.com/console/get-audio-features-track/>.
- [32] K. H. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. SAGE Publications, 2004.
- [33] J. Lovejoy, B. R. Watson, S. Lacy, and D. Riffe, "Three Decades of Reliability in Communication Content Analyses," *Journalism and Mass Communication Quarterly*, vol. 93, no. 4, pp. 1135–1159, 2016.
- [34] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.
- [35] I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer, 1997.
- [36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [37] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection." *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.
- [38] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, K. R. Scherer, and J. Krajewski, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, pp. 1–12, 2013.
- [39] E. Coutinho and B. Schuller, "Shared acoustic codes underlie emotional communication in music and speech - evidence from deep transfer learning," *PLoS ONE*, vol. 12, no. 6, 2017.
- [40] 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*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>
- [41] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 1522, no. December, pp. 1517–1522, 2019.
- [42] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [43] N. Condit-Schultz and D. Huron, "Catching the lyrics: intelligibility in twelve song genres," *Music Perception*, vol. 32, no. 5, pp. 470–483, 2015.
- [44] —, "Word Intelligibility in Multi-voice Singing: The Influence of Chorus Size," *Journal of Voice*, vol. 31, no. 1, pp. 121.e1–121.e8, 2017. [Online]. Available: <http://dx.doi.org/10.1016/j.jvoice.2016.02.011>