Music Emotions in Solo Piano: Bridging the Gap Between Human Perception and Machine Learning

Emilia Parada-Cabaleiro[1,2,3], Anton Batliner[4], Maximilian Schmitt[4], Björn Schuller[4,5], and Markus Schedl[1,2]

[1] Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2] Human-centered AI Group, Linz Institute of Technology (LIT), Austria
[3] Department of Music Pedagogy, Nuremberg University of Music, Germany
[4] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[5] GLAM – Group on Language, Audio & Music, Imperial College London, UK
emiliaparada.cabaleiro@hfm-nuernberg.de

Abstract. Emotion is an important component of music investigated in music psychology. In recent years, the use of computational methods to assess the link between music and emotions has been promoted by advances in music emotion recognition. However, one of the main limitations of applying data-driven approaches to understand such a link is the scarce knowledge of how perceived music emotions might be inferred from automatically retrieved features. Through statistical analysis we investigate the relationship between perceived music emotions (rated by 41 listeners in terms of categories and dimensions) and multimodal acoustic and symbolic features (automatically extracted from the audio and MIDI files of 24 pieces) in piano repertoire. We also assess the suitability of the identified features for music emotion recognition. Our results highlight the potential of assessing perception and data-driven methods in a unified framework.

Keywords: Music emotion recognition, multi-modal features, perception

1 Introduction

Following decades of research about music emotions in psychology [1], an increasing interest in investigating music emotions through computational methods has been driven by advances in music emotion recognition (MER) [2]. However, despite music being a multifaceted channel characterised by a variety of communication modalities, such as acoustic cues, music syntax, or lyrics, multi-modal MER is still under-investigated, in part due to the scarcity of corpora [3, 4]. In addition, since emotions are subjective concepts for which a *ground truth* does not exist, emotion recognition systems rely on a *gold standard*, i. e., labels based on some consensus annotation [5]. Still, the validity of MER labels is often questioned due to the limited number of annotators [6]. Note that, throughout the article, we will refer to *gold standard*, a standardised term in *affective computing* [7], which is more appropriate than *ground truth* [8].

To assess how perceived music emotions can be mapped onto machine-readable features, we present a perceptual and data-driven study based on 24 classical piano pieces. Through statistical analysis, we identify the acoustic and symbolic features most suited to infer a categorical and dimensional gold standard, based on ratings by 41 listeners. Finally, to evaluate the generalisability of our results, we assess the machine learning (ML) performance obtained with different feature sets on EMOPIA [4], a multi-modal pop piano music corpus for MER. In sum, we assess two research questions (RQs):
RQ1: Which are the most appropriate multi-modal features to automatically identify emotions perceived in piano music?
RQ2: Can the suitability of these features be generalised to other dataset?

2 Materials and Methods

2.1 Musical data and emotion models

We concentrate on classical western compositions for piano solo, by that minimising the influence of genre and scoring diversity. As we aim to assess both acoustic and symbolic features, the dataset introduced by Poliner and Ellis [9], containing both recordings and MIDI files, was considered for the perception study and the feature assessment. Although developed for automatic music transcription, this dataset was chosen due to its suitable repertoire and considering the limited multi-modal corpora for MER. From the 29 files available, 24 with a homogeneous musical discourse, i. e., without contrasting sections that may lead to several perceived emotions, were selected. Although we perform the feature evaluation on a reduced data-set of classical piano compositions—which was needed in order to perform a reliable user study, the generalisability of our results will be assessed in RQ2 on EMOPIA, a well-stablished piano dataset for MER. EMOPIA contains 1 087 clips from 387 songs and is annotated at clip-level according to the 4 quadrants derived from the circumplex model of emotions [10].

We employ the two models predominantly used in research on music and emotion [6]: the dimensional and the categorical one. For dimensions, we employ the circumplex model [10] representing emotions in a 2-dimensional space delimited by arousal (intensity) and valence (hedonic value), generally used in MER [4, 3]. Although research on MER often refers to basic categories, such as those described by Ekman [11], arguments in favour of moving beyond the *Basic Emotion paradigm* when working with musical emotions have been presented [12]. Thus, for categories, we use the *Geneva Emotion Music Scale* (GEMS) [13], a domain-specific categorical model specially developed to investigate music emotions, already used for MER in western classical music [14]. As we investigate perceived emotions, the 10-factorial version of GEMS[6], used in Study 2 in [13] to assess perceived emotions, was preferred to the original GEMS (developed to assess felt emotions). Note that GEMS has proven to be as suitable to evaluate perception as felt emotions (see Study 2 [13] as well as [14]). In addition, as typical in MER [4, 3] and in order to assess RQ2, the four quadrants derived from the intersection of the two emotional dimensions will be considered as target categories for the ML experiments. The quadrants are defined as in [15]: Q1 (high arousal, positive valence); Q2

[6] The 10-factors (i. e., emotional categories) are: Activation, Amazement, Dysphoria, Joy, Power, Tenderness, Tranquility, Transcendence, Sadness, and Sensuality.

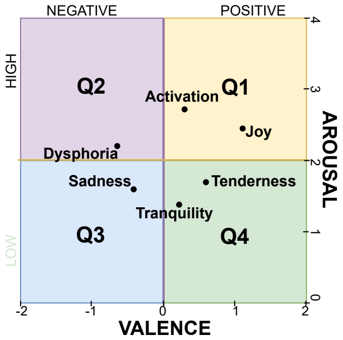


Fig. 1: Emotional categories distributed according to the 4 quadrants. The dots indicate the gold standard, i. e., the mean valence/arousal coordinate across samples per emotion.

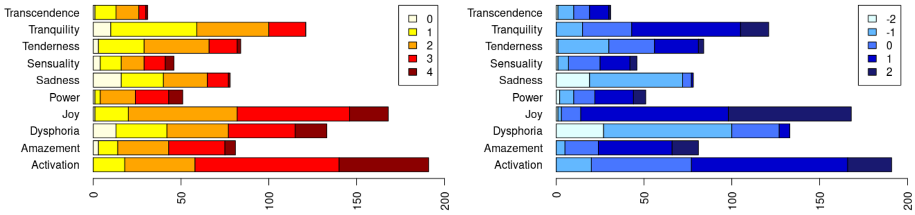
(high arousal, negative valence); Q3 (low arousal, negative valence); Q4 (low arousal, positive valence); cf. Figure 1 (positions of categories are explained in Section 2.2).

2.2 Annotation process

41 male students participated in the listening experiment as a requirement of a course.[7] The musical samples, each with a duration of 59 seconds, were presented in randomised order over headphones; the responses were given in a forced-choice format through a web-based interface. For each musical sample, the participants had to choose one of the 10 emotional categories, a level of arousal (from 0 to 4), and a level of valence (from -2 to 2). Note that valence (unlike arousal) can have negative values; thus the scale is not the same but more adequate. We used static annotations instead of continuous, i. e., each annotation was given at sample level. Despite the length of the samples, this was considered the best choice in order to be consistent with the annotations from EMOPIA, the dataset used to validate our results. As already mentioned, to prevent annotation ambiguity due to samples' length, those with a homogeneous musical discourse were selected. Finally, since liking and familiarity have played a role in previous works [16, 17], participants were also requested to indicate in binary form (yes/no) whether they were familiar with the evaluated repertoire and whether they liked it.

To create a gold standard for valence and arousal, we computed the mean across ratings per sample and dimension, as typical in MER [6]. In addition, we also computed the Evaluator Weighted Estimator (EWE), an standard method to compute a gold standard in affective computing [18] that takes into account an individual evaluator-dependent weight for each annotator. The evaluator-dependent weights are the normalised correlation coefficients obtained between each listener's responses and the average ratings across all listeners [18]. As both Spearman and Pearson correlations between mean and EWE are at 99 %, we use the mean in the following. To create the categorical gold standard, the emotional factor showing the highest agreement was considered as target category, as typical in MER [6]. In Figure 1, the categories chosen most frequently

[7] Although considering only males' ratings might affect the results, responses by the only three females who took part in the experiment had to be discarded to preserve a coherent cohort.



(a) Arousal ratings

(b) Valence ratings

Fig. 2: Distribution of the 984 ratings (41 listeners × 24 samples) for each dimension.

across samples are shown within the quadrants. The mean arousal and valence ratings across all samples identified with the given categories are shown. For the distribution of all the listeners' ratings across factors and dimensions, see Figure 2.

2.3 Feature extraction and processing

Symbolic and acoustic features were extracted from the MIDI and audio files and subsequently concatenated in a feature vector. Concerning the symbolic data, we extracted the features of `jSymbolic 2.2` [19], which include a variety of statistical descriptors related to pitch, rhythm, melody, chords, texture, and dynamics (related to MIDI velocity), i. e., musical properties suitable to automatically capture emotional content from MIDI [15]. Since we aim to evaluate the features in relationship to the perceptual results, we choose jSymbolic, whose features are highly interpretable in musical terms. As acoustic representation, we considered the *openEAR emobase* feature set extracted with the default parameters of `openSMILE` [20], which is tailored to model emotions in audio and has been used in the context of MIR as well [21]. *OpenEAR emobase* contains statistical descriptors related to intensity, loudness, pitch, envelope, and spectrum.

After excluding irrelevant features, e. g., those related to the Music Encoding Initiative format for the symbolic and the delta coefficients for the acoustic modelling, 188 symbolic and 494 acoustic features were retained for analysis and subsequently z-score normalised. In order to prevent collinearity [22], redundant features, i. e., those showing a pair-wise correlation of $r \geq 0.7$, were automatically identified; the one showing the largest mean absolute correlation was subsequently removed. For this, the correlations were recomputed at each step with the R function *findCorrelation*. This yielded a total of 91 features—68 symbolic and 23 acoustic. From now on, these constitute the 91-dimensional feature vector representing each sample.

2.4 Statistical methods

To explore which features might be suitable to predict perceived arousal and valence, Pearson correlation was computed between each feature and the gold standard for each dimension. Since features might also be suitable in combination, two multiple regression models were fitted separately for each dimension. In addition, to assess individual ratings instead of the gold standard, all *raw* responses were directly taken as outcome variable for these models. Note that, as every listener co-occurs in the design with every song, the variables user-ID and song-ID were considered crossed random effects. The need of applying a multi-level analysis was confirmed by the decreased *Akaike's information criterion* (AIC) of the intercept model with crossed random effects w. r. t. those with only one random effect: for both dimensions, $p < .001$. Suitable predictors were automatically recognised through a *Genetic Algorithm* (GA), implemented in R with default parameters and 100 iterations. Subsequently, forward selection was applied in order to evaluate if additional predictors might yield a lower AIC. Given the inherent problems of *p*-values [23], in particular for linear mixed models [24], we will interpret the role of the fixed effects according to the regression coefficients.

After identifying suitable features through correlation and multiple regression, in order to visually interpret the suitability of such a features in mirroring the listeners' ratings, we compare perception and classification results. For this, we used *Non-Metrical Multi-Dimensional Scaling* (NMDS) solutions [25], which aim at representing the optimal distances between items. To find the optimally scaled data, NMDS is initialised with a random configuration of data points and subsequently finds the optimal monotonic transformation of the proximities. This search for a new configuration is performed iteratively until Kruskal's normalised stress1 criterion or its gradient is below a threshold of 10^{-4}. Since our goal is not to achieve the best possible result through fine-tuning, but to compare classification performance across feature sets while keeping hyperparameters constant, for this experiment, the classification framework (described in Section 2.5) was implemented with default parameters and without optimisation.

2.5 Machine learning models and optimisation

Four classifiers, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and k-Nearest Neighbour (k-NN), were implemented. To leverage the advantages of all models, we created a hybrid classifier using late-fusion of results via majority voting, i. e., the class most frequently chosen by the four models was taken as final prediction. We do not concentrate on pushing the approaches towards their limits, but aim at baseline results with 'standard' settings, by this encouraging generalisation of the outcomes. As evaluation metric, we use Unweighted Average Recall (UAR) [7].

The data were randomly split into train, validation, and test. We targeted a similar distribution between classes across quadrants; samples from the same song did not occur in different sets. To increase validity, five different splittings were generated; we report the average results across experiments. The models were built on the scikit-learn python library [26] with the default hyperparameters, except for the following set-up: For the SVM, we use linear kernel and evaluate five different complexities [0.0001, 0.001, 0.01, 0.1, 1.0]. For the MLP, we use batch size 8, two hidden layers, and evaluated the same number (N) of neurons per layer from the following five N [25, 50, 100, 175, 300]. For the RF, we evaluate five different N of estimators [10, 50, 100, 150, 200]. For the k-NN, we evaluate five different N of neighbours [3, 5, 7, 9, 11]. All hyperparameters were optimised independently for each of the five splits via grid search.

3 Gold Standard Assessment

As first step to create the gold standard, we evaluated the role of familiarity and preference. For this, multiple regression was performed considering both variables as categorical predictors and the perceived valence and arousal individually as dependent variables. Our results show that neither preference nor familiarity play a role in the model, neither for arousal, nor for valence ($p \geq .084$). This is also confirmed for within song evaluation: the models yielded $p \geq .286$ for arousal, $p \geq .353$ for valence.[8] Thus, in the following, all listeners' responses will be taken into account for our experiments.

The gold standard computed from listeners' responses shows that joy is mainly associated with Q1 (5 songs) and to some extent with Q4 (1 song); activation with Q1 (5 songs) and to some extent with Q2 (2 song); dysphoria with Q2 and Q3 (2 songs each); sadness is clearly associated with Q3 (2 songs); tenderness with Q4 (1 song); tranquility with Q3 and Q4 (2 songs each). This distribution of emotional categories across the bi-dimensional space (cf. Figure 1) is consistent with the one described in previous works (cf. [10] and [1, p. 113]), where joy/dysphoria are associated with positive/negative valence; activation/tranquility are associated with high/low arousal; tenderness/sadness are related to low arousal and to positive/negative valence. This is displayed by the distribution of the dimensional ratings. For sadness, in particular, the ratings are mostly distributed across the lowest and intermediate arousal (cf. 0 to 2 in Figure 2a), and almost all display negative valence (cf. -2 and -1 in Figure 2b).

To gain more insights on the perceptual results, we investigated the relationship between both dimensions. For this, each of them was considered as outcome and predictor, respectively, in a linear model, disregarding the categorical ratings. The positive slope indicates that there is a direct relationship between both variables: $F = 83.56$, $\beta = 0.30$, $r = 0.28$, $p < .001$. In other words, as perceived ratings increase in one unit for a given dimension, the model predicts that the perception for the other one will also increase in 0.30 units. Still, the correlation of $r = 0.28$ indicates only a weak tendency.

Subsequently, to evaluate if the relationship between valence and arousal might be associated with categorical perception, for each emotion, an individual linear model was fitted with the corresponding dimensional ratings. The results show that the positive relationship between both dimensions is only marked for some emotions: the linear regression yields $p \leq .046$ for amazement, joy, sensuality, and tranquility, i. e., those generally associated with a more positive valence, cf. Figure 2b; for the others, $p \geq .346$. Indeed, fitting again the model with the dimensional ratings of only these emotions increased the correlation coefficient ($r = 0.48$), which confirms the positive association between valence and arousal but only within the positive half of the dimensional space, i. e., Q1 and Q4. To reproduce the gold standard and results, please visit our repository.[9]

4 Results

RQ1: Which are the most appropriate multi-modal features to automatically identify emotions perceived in piano music?

CORRELATION ANALYSIS: To investigate the relationship between the automatically extracted features and the perceived emotional dimensions, correlation analysis was performed. In Table 1, only the top ranked features ($|r \geq 0.4|$ in at least one dimension), i. e., those showing a moderate correlation, are displayed. Since a relationship between

[8] Bonferroni correction was applied for multiple testing throughout the results.

[9] `https://github.com/SEILSdataset/FeatureEval_MER/`

Table 1: Top ranked correlation with the mean (μ) perceived arousal and valence.

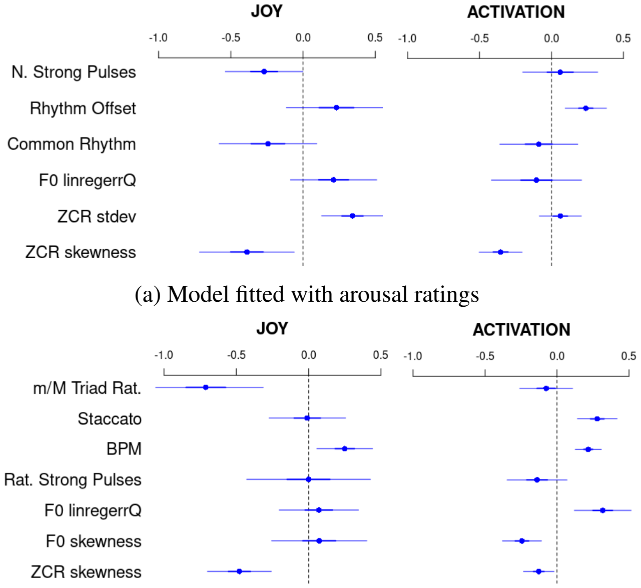
Arousal		**Valence**	
Feature	μ	Feature	μ
Common Rhythm	$-.65$	*m/M Triad Rat.*	$-.54$
ZCR Skewness	$-.57$	*F0 Quartile3*	$.53$
Note Density	$.54$	*Intensity abs. min.*	$-.48$
Mel. Large Int.	$-.49$	*m/M Mel. 3rd Rat.*	$-.46$
N. Strong Pulses	$-.48$	*Arousal*	$.46$
Standard Triads	$-.46$	*F0 Skewness*	$-.44$
Valence	$.46$	*Similar Motion*	$-.43$
Rat. Strong Pulses	$-.42$	*Rat. Strong Pulses*	$-.41$
BPM	$.42$	*Dynamic Range*	$-.40$
Prev. Dotted Notes	$-.41$	*Dim. Aug. Triads*	$-.40$

both dimensions was shown in the gold standard assessment, these are also included in the correlation analysis. In the following, the correlation results will be interpreted according to [1, p. 113], which summarises the outcomes from music psychology.

Arousal. The experimental results are consistent with the general believe that slow and fast mean tempo correspond to music expressing low and high arousal, respectively. This is shown by the positive correlation of arousal with Beat Per Minute (BPM, $r = .42$) as well as by the negative one with common rhythm and prevalence of dotted notes ($-.41 \leq r \leq -.65$), indicating that music characterised by a fast tempo and a prominent use of short (not dotted) notes is associated with higher arousal. Similarly, the use of accents on unstable notes (typically used to express highly aroused music) is shown by the negative correlation of arousal with number and ratio of strong pulses ($-42 \leq r \leq -48$): As perceived arousal increases, the amount of strong beat peaks decreases and is diversified towards non-beat ones.

High arousal is also associated with a high sound level, which is confirmed by the positive correlation of arousal with note density ($r = .54$) and the negative one with Zero-Crossing Rate (ZCR) skewness ($r = -.57$). While note density is implicitly related to sound level, a low ZCR skewness can be interpreted as a 'constant' (not skewed) distribution of frequency density over time: ZCR $= 0$ indicates no sound. Besides being consistent with outcomes from music psychology [1, p. 113], our experimental results for arousal also show that an increase in this dimension goes along with a decrease in the use of standard triads w. r. t. other vertical intervals ($r = -.46$). This can be interpreted as an association of high arousal with a more 'empty' (without third) sonority.

Valence. The small sound level variability typically associated with positive valence is shown by the negative correlation of this dimension with dynamic range ($r = -.40$). Our results are also consistent with the believe that minor/Major music expresses negative/positive emotions [27], as shown by the negative correlation of valence with m/M triad and melodic third ratio ($-.46 \leq r \leq -.54$). Similarly, positive valence goes along with a detriment in augmented and diminished triads ($r = -.40$), which indicates that negative valence is associated with a higher use of dissonant chords. Our results suggest that positive valence is linked to the use of a lower variety of pitches concentrated around high pitch, something that can be related to the common association of joy with bright timbre. This is shown by the positive correlation of valence with the Fundamen-



(a) Model fitted with arousal ratings

(b) Model fitted with valence ratings

Fig. 3: Fixed effects' regression coefficients (blue dot) and confidence intervals (blue line) for the two models: one for arousal, the other for valence.

tal frequency (F0) quartile 3 ($r = .53$) and by the negative one with the F0 skewness ($r = -.44$): Low F0 skewness indicates a similar distribution of frequencies over time.

Arousal and valence are positively correlated ($r = .46$). Still, the low sound level typically used to express emotions with positive valence and low arousal is also shown by the negative correlation of valence with absolute minimum intensity and dynamic range ($-.40 \leq r \leq -.48$). This indicates that, despite the positive correlation between both dimensions in the investigated samples, the extracted features are also suitable to identify emotions with a positive valence and low arousal.

MULTIPLE REGRESSION: To investigate the interplay between the automatically extracted features and the categorical as well as dimensional ratings, the best fitting models, separately identified for each dimension, were also fitted with the subset of dimensional ratings corresponding to each emotional category (cf. Section 2.4). Using the general models tailored to each dimension was preferred to retrieving an individual model per category, to enable comparability. Due to space limitations, in Figure 3, only results for joy and activation, i. e., the two categories with the highest number of observations—joy 168, activation 191, thus showing most robust results—are shown.

The features of the model tailored to recognise arousal include three symbolic, related to rhythm, and three acoustic ones, related to F0 and ZCR. Indeed, both note duration, related to rhythm, as well as intonation and spectral noise, related to F0 and ZCR, are relevant properties for the expression of arousal in music [1, p. 113]. In particular, the higher positive slope of ZCR standard deviation for joy indicates that unlike for activation, an increase in arousal goes along with a higher variability of silent and

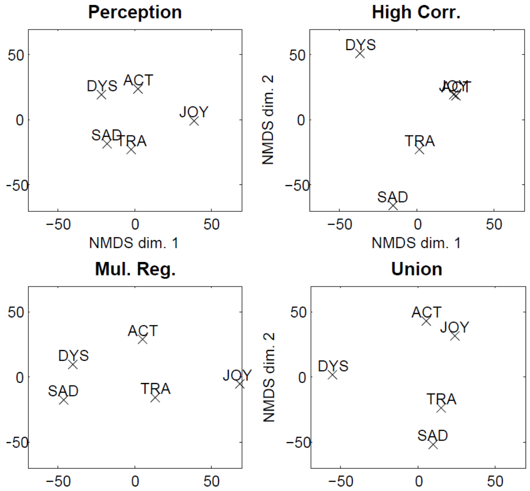


Fig. 4: NMDS for the perception and classification (High Correlation, Multiple Regression, and Union features) of JOY, ACTivation, DYSphoria, SADness, and TRAnquility. Kruskal's stress: Perception (.097); High Cor. (.093); Mul. Reg. (.024), Union (.006).

dense frames over time. Again, as shown in the correlation analysis, the m/M triad ratio is relevant to predict valence, as clearly displayed for joy. Interestingly, BPM and staccato are meaningful features for the valence model but not for the arousal one. The fact that these features show a relatively marked positive slope—for activation both, for joy only BPM—might again be an indicator of the positive relationship between both dimensions, as shown by the listeners' association of these two factors with high arousal and positive valence (cf. Q1 in Figure 1).

PERCEPTION VS CLASSIFICATION: To further explore the suitability of the identified features for discriminating between the perceived emotions, we compare classification performance with the perceptual results (cf. Figure 4). As there is a relationship between the emotional factors and specific regions of the bi-dimensional space (cf. Figure 1), the features tailored to arousal and valence are both considered for the classification of emotional categories. Three feature sets are assessed: the features with top correlation (High Corr., 17 features), shown in Table 1; the ones used for the Multiple Regression (Mult. Reg., 11), shown in Figure 3; and the union of both (Union, 21). As some features are part of both High Corr. and Mult. Reg., Union contains less features than the sum of these sets. For a description of the features see Table 2. More details are given in the official documentation of `jSymbolic 2.2` and `openSMILE`.Tenderness (cf. Figure 1) is not considered, as attributed to only one sample.

The Union feature set, showing the best fit (Kruskal's stress .006), is the one best mirroring the Perception NMDS: Joy and activation are shown towards Q1; dysphoria towards Q2; sadness and tranquility are close to each other. Although for perception, sadness is more clearly displayed in Q3 than for the Union feature set, this set, combining High Corr. and Mult. Reg., is a less condensed version of the Perception results; cf. Union in Figure 4. Thus, from now on, the Union feature set will be used.

Table 2: Description of the symbolic and acoustic features of the Union set.

Symbolic Features			
Common Rhythm	Most common rhythm in quarter note units	*Similar Motion*	Fraction of similar movements, e. g., parallel
N. Strong Pulses	N. of beat peaks with magnitudes over 0.1	*Staccato*	Fraction of notes shorter than 0.1 seconds
Rat. Strong Pulses	Ratio of the two highest beat magnitudes	*Note Density*	Average number of notes per second
Rhythm Offset	Median absolute duration offset	**Acoustic Features**	
m/M Mel. 3rd Rat.	Ratio of the minor/Major melodic thirds	*Intensity abs. min.*	Frame-based absolute minimum intensity
m/M Triad Rat.	Ratio of the minor/Major vertical triads	*BPM*	Beat per minute
Standard Triads	Fraction of minor or Major triads	*ZCR stdev*	Standard deviation of the zero-crossing rate
Mel. Large Int.	Fraction of melodic intervals $>$ octave	*ZCR Skewness*	Skewness of the zero-crossing rate
Dynamic Range	Highest loudness value minus the lowest	*F0 Skewness*	Fundamental freq. (F0) contour's skewness
Prev. Dotted Notes	Fraction of dotted notes	*F0 linregerrQ*	Quadratic error of the F0 contour
Dim. Aug. Triads	Fraction of diminished or augmented triads	*F0 Quartile3*	Third quartile of the F0 contour

RQ2: Can the suitability of the identified features be generalised?

To assess the generalisability of the identified features, we performed the classification experiments (optimising the models as described in Section 2.5) on the EMOPIA dataset. To interpret confusion patterns across the dimensional quadrants, i. e., the target categories in EMOPIA, besides the Union dataset (used to assess the RQ1), we now investigate the performance of the Union features tailored to recognise each dimension individually as well. In addition, since the size of EMOPIA enables to carry out a real evaluation of the results beyond NMDS interpretation, the ML models were also trained with all the features (i. e., the 91 described in Section 2.3). Thus, the experiments on EMOPIA were performed with four feature sets: all features (91), Union features tailored to arousal and valence (12 each), and the Union feature set (21).

The results on EMOPIA indicate that training the models with all the features shows a clear differentiation of the arousal dimension: Q1 and Q2 (both with high arousal) are clearly distinct from Q3 and Q4 (both with low arousal) while confused with each other (Q1 with Q2, Q3 with Q4); cf. *All features* in Table 3. As expected, this pattern is enhanced for the features tailored to arousal, which do not contain features tailored to recognise valence information and display a much more pronounced confusion between quadrants of the same arousal level (cf. dark cells of *Arousal selection* in Table 3). In contrast, besides a relatively high recall for Q4 and its confusion towards Q1 (both with positive valence), no clear distinction/confusion pattern is shown for the features tailored to recognise valence; cf. *Valence selection* in Table 3. This feature set yields the worst UAR ($39.2\,\%$), and the recall for Q1 and Q4 does not outperform the one achieved by the other feature sets either, which suggests its low capability in capturing information relevant to the target dimension. Finally, the *Union* features (without dimension selection, i. e., A + V) slightly outperform the *Arousal selection* (UAR $= 52.5\,\%$ vs UAR $= 50.7\,\%$), but without reaching the performance of *All features* (UAR $= 64.1\,\%$). Again, a differentiation in terms of arousal is displayed.

The experimental results suggest that the arousal dimension is more prominent in the evaluated data, something also observed in emotional speech, where arousal is better represented by acoustic cues than by linguistic ones [28]. The lower efficiency of the features tailored to model valence might be interpreted, to some extent, according to previous works which had shown the difficulties, from a listeners' point of view, of assessing valence, even in music expressing sadness [29], a basic emotion which is, however, clearly associated to negative valence. The classification results achieved with

Table 3: EMOPIA: confusion matrices averaged across splits. Columns show 'classified as'. UAR for each feature set: All (64.1 %); Arousal (50.7 %); Valence (39.2 %); Union (52.5 %), i. e., Arousal and Valence (A + V).

%	**All features**				**Arousal selection**				**Valence selection**				**Union (A + V)**			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Q1	81.2	14.1	1.7	3.0	67.1	23.5	3.8	5.6	51.7	27.8	6.8	13.7	65.4	24.8	3.0	6.8
Q2	34.2	55.8	3.8	6.2	33.8	51.2	7.7	7.3	39.2	37.7	8.8	14.2	37.7	50.0	3.5	8.8
Q3	7.6	8.6	61.1	22.7	14.1	8.6	43.4	33.8	29.3	26.8	20.7	23.2	13.1	7.1	48.0	31.8
Q4	12.8	7.8	21.0	58.4	15.5	11.4	32.0	41.1	25.1	19.6	13.2	42.0	14.2	10.5	30.1	42.2

all the features yielded the highest UAR, suggesting that the usability of the *Union* set for MER might be limited. Still, the identified features show reasonable results with a much lower dimensionality, something that might be beneficial for some MER systems.

5 Conclusion and Future Work

Besides confirming some of the outcomes presented in music psychology literature, our data-driven approach shows that automatically extracted multi-modal features might be suitable to infer perceived musical emotions. For instance, the statistical analysis suggests that in the evaluated repertoire, empty sonorities might be an indicator of perceived high arousal, while high pitch is related to positive valence. The machine learning experiments show that the features identified to model arousal lead to competitive classification results concerning the quadrants related to the target dimension. In contrast, those identified to model valence are considerably less efficient, which might be explained by the lower characterisation of this emotional dimension in music. Finally, the importance of a multi-modal approach becomes clear when evaluating the feature sets, which despite being selected in a fully automatic manner, encompass both symbolic and acoustic features. In future work, besides investigating a larger dataset from a more varied repertoire, we also plan to assess music with lyrics, by this assessing the suitability of linguistics in the identification of the valence dimension.

Acknowledgements

This work received support from the Austrian Science Fund (FWF): P33526 and DFH-23.

References

1. Juslin, P.: Musical Emotions Explained. Oxford University Press., Oxford, UK (2019)
2. Han, D., et al.: A survey of music emotion recognition. Frontiers of Computer Science **16** (2022) 1–11
3. Panda, R., et al.: Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In: Proc. of CMMR, Marseille, France (2013) 1–13
4. Hung, H.T., et al.: EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: Proc. of ISMIR, Virtual (2021) 318–325
5. Cardoso, R., et al.: What is gold standard and what is ground truth? Dental Press Journal of Orthodontics **19**(5) (2014) 27–30

6. Gómez-Cañón, J.S., et al.: Music emotion recognition: Towards new robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine **38** (2021) 106–114
7. Schuller, B., Batliner, A.: Computational paralinguistics: Emotion, affect and personality in speech and language processing. John Wiley & Sons, Sussex, UK (2014)
8. Parada-Cabaleiro, E., et al.: Perception and classification of emotions in nonsense speech: Humans versus machines. PLoS ONE **18**(1) (2023) e0281079
9. Poliner, G., Ellis, D.: A discriminative model for polyphonic piano transcription. EURASIP Journal on Advances in Signal Processing (2006) 1–9
10. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6) (1980) 1161–1178
11. Ekman, P.: Basic emotions. In: Handbook of emotion. John Wiley & Sons (1999) 226–232
12. Cespedes-Guevara, J., Eerola, T.: Music communicates affects, not basic emotions–A constructionist account of attribution of emotional meanings to music. Frontiers in Psychology **9** (2018) 1–19
13. Zentner, M., Grandjean, D., Scherer, K.: Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion **8** (2008) 494–521
14. Schedl, M., et al.: On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. IEEE Transactions on Affective Computing **9**(4) (2017) 507–525
15. Panda, R., et al.: Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing **11**(4) (2018) 614–626
16. Schubert, E.: The influence of emotion, locus of emotion and familiarity upon preference in music. Psychology of Music **35**(3) (2007) 499–515
17. Pereira, C.S., et al.: Music and emotions in the brain: Familiarity matters. PloS one **6** (2011)
18. Grimm, M., et al.: Primitives-based evaluation and estimation of emotions in speech. Speech Communication **49**(10-11) (2007) 787–800
19. McKay, C., et al.: jSymbolic 2.2: Extracting features from symbolic music for use in musicological and MIR research. In: Proc. of ISMIR, Paris, France (2018) 348–354
20. Eyben, F., et al.: Opensmile: The Munich versatile and fast open-source audio feature extractor. In: Proc. of ACM Multimedia, Florence, Italy (2010) 1459–1462
21. Shen, T., et al.: Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. In: Proc. of the AAAI Conf. on AI, New York, NY, USA (2020) 206–213
22. Dormann, C., et al.: Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography **36** (2013) 27–46
23. Wasserstein, R.L., Lazar, N.A.: The ASA's statement on p-values: Context, process, and purpose. The American Statistician **70** (2016) 129–133
24. Baayen, R.H., et al.: Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language **59**(4) (2008) 390–412
25. Kruskal, J., Wish, M.: Multidimensional Scaling. Sage University, London, U.K. (1978)
26. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12** (2011) 2825–2830
27. Gabrielsson, A., Lindström, E.: The role of structure in the musical expression of emotions. In: Handbook of Music and Emotion. Oxford Uni. Press, Boston, MA, USA (2010) 187–221
28. Atmaja, B.: Predicting valence and arousal by aggregating acoustic features for acoustic-linguistic information fusion. In: Proc. of TENCON, Osaka, Japan (2020) 1081–1085
29. Eerola, T., Vuoskoski, J.K.: A comparison of the discrete and dimensional models of emotion in music. Psychology of Music **39**(1) (2011) 18–49