Prediction of Music Pairwise Preferences from Facial Expressions

Marko Tkalčič
Faculty of Computer Science
University of Bozen-Bolzano
Bolzano, Italy
marko.tkalcic@unibz.it

Nima Maleki
Vodafone Research, Italy
Milano, Italy
nima.maleki@vodafone.com

Matevž Pesek
Faculty of Computer Science
University of Ljubljana
Ljubljana, Slovenia
matevz.pesek@fri.un-lj.si

Mehdi Elahi
Faculty of Computer Science
University of Bozen-Bolzano
Bolzano, Italy
meelahi@unibz.it

Francesco Ricci
Faculty of Computer Science
University of Bozen-Bolzano
Bolzano, Italy
fricci@unibz.it

Matija Marolt
Faculty of Computer Science
University of Ljubljana
Ljubljana, Slovenia
matija.marolt@fri.uni-lj.si

ABSTRACT

Users of a recommender system may be requested to express their preferences about items either with evaluations of items (e.g. a rating) or with comparisons of item pairs. In this work we focus on the acquisition of pairwise preferences in the music domain. Asking the user to explicitly compare music, i.e., which, among two listened tracks, is preferred, requires some user effort. We have therefore developed a novel approach for automatically extracting these preferences from the analysis of the facial expressions of the users while listening to the compared tracks. We have trained a predictor that infers user's pairwise preferences by using features extracted from these data. We show that the predictor performs better than a commonly used baseline, which leverages the user's listening duration of the tracks to infer pairwise preferences. Furthermore, we show that there are differences in the accuracy of the proposed method between users with different personalities and we have therefore adapted the trained model accordingly. Our work shows that by introducing a low user effort preference elicitation approach, which, however, requires to access information that may raise potential privacy issues (face expression), one can obtain good prediction accuracy of pairwise music preferences.

CCS CONCEPTS

• **Human-centered computing** → *Interaction devices*; *Empirical studies in HCI*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

pairwise scores, implicit preference elicitation, facial expressions, emotions

ACM Reference Format:
Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2019. Prediction of Music Pairwise Preferences from Facial Expressions. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI '19).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3301275.3302266

1 INTRODUCTION

User preference elicitation is a critical step any recommender system to perform well. Traditionally, it has been implemented either by (i) explicitly asking users to enter ratings from an ordinal scale (e.g., five stars) or (ii) using implicit signals, such as clicks, previews, purchases, playcount, listening time etc. Ratings are considered a more reliable measure of the user preferences whereas implicit signals are only approximate indicators of the user preferences. For example, if a user skips a song after having listened to 75% of its duration, can one conclude that she likes that song or not? It is hard to say for sure. Another important aspect of the distinction between explicit and implicit preference elicitation is intrusiveness and the required user effort. Asking the user to provide explicit ratings is intrusive because it interferes with the interaction flow, whereas implicit signals are acquired in the background without disrupting the interaction. Moreover, the user must spend time to evaluate an item and to enter her evaluation into the system; this required effort may have a negative effect on the correctness of the entered evaluation.

Independently from whether the feedback is acquired implicitly or explicitly, it can be acquired as a single judgment against an absolute benchmark or in the form of a pairwise comparison, hence a relative evaluation of the compared items [20]. When pairwise comparisons are acquired the user is shown pairs of items and she has to indicate which one she prefers in each pair. Each comparison can be coded in a numerical scale and is called a *pairwise score*. For instance, if we consider the pair of items (A, B) and the user prefers A to B, this can be coded as 1, while if B is preferred to A it can be coded as -1. One can also consider more nuanced comparisons, such as, A is much more preferred to B, it could be coded as +2. There are cases, where pairwise preferences and their elicitation may be more appropriate than single item judgments. In the aforementioned example related to the inference of user preferences from listening

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302266

time, if we have two songs, where one has been listened to only 15% and the other 75% of their respective durations, we can safely guess that the user prefers the second song to the first one. As a matter of fact, relative comparisons bring less information than absolute, single item, evaluations. But, deriving pairwise comparisons from implicit signals may be easier and more reliable.

Moreover, pairwise comparison helps the users to reflect more on their preferences and require less cognitive effort especially when the items that are compared are actually *comparable*. Two items are comparable when they belong to the same category, i.e., they are rather similar, and we can justify the comparison on the base of their features. For instance, hotel A is cheaper than hotel B, or track A is mellower than track B, are direct justifications for preferring one item to the other. Obviously, reflecting on the item features and determining which one is important for the user is easier and appropriate when the user task is clear. For example, two songs can be compared based on many attributes, such as their mood, tempo, danceability, energy, loudness etc [4]. However, when choosing between two songs for a party, energy and danceability might be more important than other attributes. Hence, the context of the music consumption is also important when determining the comparability of two songs. Since our aim in future work is to build a context-aware recommender system for music we have assumed that pairwise comparisons could be a more appropriate preference elicitation mechanism than ratings.

However, asking the user to provide pairwise preferences explicitly is cumbersome. We were therefore interested in finding out how one can possibly use implicit feedback, that is, data collected without requesting an explicit action to the user, to signal her preference. We conjectured that these signals could be effectively used for predicting "explicit" pairwise preferences, i.e., pairwise preferences that would be given by the user if explicitly requested. Traditionally, implicit feedback-based approaches are leveraging signs of the user-system interaction that are easy to acquire and are collected also for other reasons, e.g., system monitoring. Examples are play-counts and listening time in the music domain [26, 31].

In this scenario, we also observe that music is a particular application where emotions are important. Research has shown that music listening evokes emotional responses [39]. These emotional responses are coupled with bodily responses, among which facial expressions are ones of the strongest predictors of the current emotion [35]. Two pieces of music that evoke different emotions should cause the user to produce different facial expressions, too. Furthermore, it has been found that the strength of an emotion (i.e. the arousal) is correlated with the music preferences of the listener [23]. Hence, we here conjecture that the difference in facial expressions of a user while listening to two songs is a predictor of the pairwise preference of that user for the two songs.

The research question we address is hence: **Can we infer (implicitly) pairwise music preferences of a user from her facial expressions during the listening of songs?** Figure 1 summarizes the approach we take to address the research question.

In this paper we present a novel approach for implicit pairwise music preference elicitation. The novelty lies in (i) the usage of new implicit signals and (ii) the fact that these features provide a better prediction of pairwise preferences than a baseline method that uses listening time. The user facial expressions when listening to songs are leveraged as implicit signals of her preferences. These are more difficult to obtain than other implicit signals, such as listening time, but are conjectured to yield a higher prediction accuracy of the user's pairwise preferences.

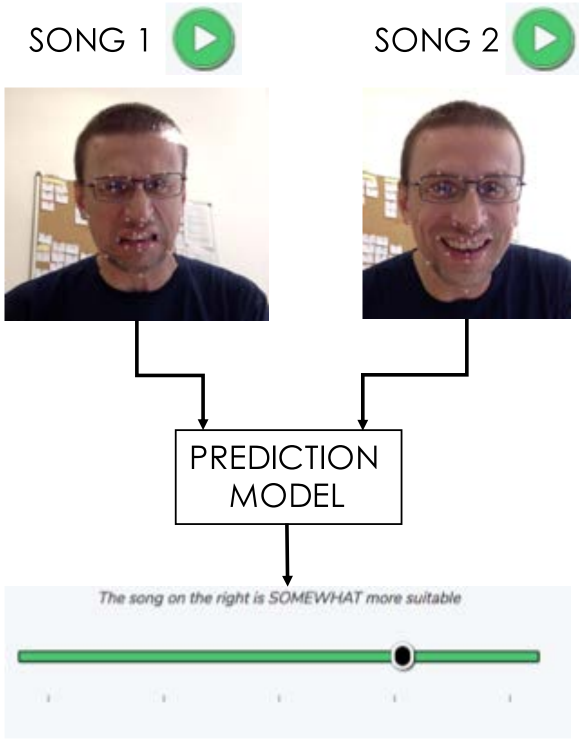


Figure 1: We hypothesize that the difference between the facial expressions while listening to song 1 and the facial expressions while listening to song 2 are related to the user preference, hence are informative features for predicting the pairwise score of songs 1 and 2.

The experimental results show that our facial expressions-based approach does predict the pairwise preference better than a listening time-based approach. The prediction error is decreased by 17%.

With the proposed approach we are introducing a more complex acquisition mechanism compared to listening time but we improve the accuracy of the prediction of the users' pairwise preferences. However, the proposed method raises privacy issues. Although a thorough evaluation of the privacy perspective is out of the scope of this paper, we discuss some approaches that can be used to tackle this issue.

Finally, previous research has also shown that there are differences between users in the way they perceive music [30]. In our work we have also explored these individual differences and found that the prediction accuracy differs in groups of people with common personality traits (e.g. people who are extroverted).

2 RELATED WORK

The conducted work stands on a diverse set of related work. In this section we cover topics that support the rationale of our research hypothesis and the chosen experimental approach. Our general contribution is about techniques for predicting *pairwise preferences* in music. The prediction technique we present is based on *implicit feedback* from the user. Since related work shows that *music is related to emotions* and *emotions are related to user preferences* then we used emotions as implicit signals of music preferences. Emotions are manifested through a variety of bodily signals, among which are *facial expressions.* We use facial expressions as a proxy for *acquiring the user's emotions* felt during the listening of a song. As emotions and preferences depend on *personal user characteristics* we provide an overview of that related work as well. Finally, we summarize the work done on *privacy in recommender systems* and position our work in this context.

2.1 Pairwise Preferences

Preferences can be acquired as single judgments about items (e.g. a rating on an ordinal scale from 1 to 5) or as pairwise judgments comparing two (or more) alternatives (e.g. I prefer item A to item B). Bockenholt [6] showed that pairwise judgments are useful when users have difficulty assessing the utility of a single item. This is especially true when the utility of an item changes with context or when the preference elicitation method itself influences the judgment of a user [6]. This holds in the music domain. We usually listen to music in a specific context, e.g., when working or driving. One may find it difficult to give a precise (scaled) judgment about a particular song. But, when asked to simply compare two songs for a given usage scenario, the user is often able to provide a more reliable preference.

Once the pairwise preferences are acquired, recommender system algorithms can take advantage of such type of preferences. For instance, Bayesian Personalized Ranking (BPR) [27] takes pairwise scores and generates a ranked list of recommended items. The algorithm proposed by Kallori et al. [20] takes pairwise scores to predict missing pairwise scores. Then it uses the predicted pairwise scores to generate a ranked list of recommended items.

2.2 Implicit Preference

As asking users to explicitly express their music preferences is intrusive, researchers have started to focus on the implicit acquisition thereof. Jawaheer et al. state that there is a substantial difference in the cognitive effort requested for providing explicit vs. implicit feedback [18]. They claim that acquiring implicit user feedback is seamless, whereas explicit user feedback requires some cognitive effort. Parra and Amaitrian [26] have shown that existing music listening traces can be used for generating good predictions of actual music ratings. One of the most popular implicit signals is playcount, i.e., the number of times a user has played a song [18, 26]. Listening time is a similar measure and has been shown to correlate positively with the user preferences expressed in the form of ratings. The study carried out by Dunn et al. [9] showed that the more the users like a song the longer they listen to it. Moling et al. [25] used the listening time as an evaluation metric for their music recommender system. In the work presented here we have used the listening time as the main predictive feature of a baseline predictor of pairwise scores. To the best of our knowledge, there are no other implicit techniques for pairwise score acquisition. The only approach that we are aware of comes from Information Retrieval, where a clicked result in a ranked list is treated as the preferred one in comparison to the other items on the list of results [27]. However, this approach is not applicable in our usage scenario.

2.3 Music and Emotions

There is a vast body of research showing that music and emotions are related. Emotions are short-lasting bodily responses to stimuli, as opposed to mood, which is a long-lasting experience without an identifiable stimulus [11]. When an emotion is experienced, a set of bodily responses are triggered. These bodily responses include physiological changes (e.g., change in heart-beat or sweating) and expressions (e.g., facial and vocal expressions) [33]. Emotions have been also linked to decision making. Research has shown that the physiological responses that occur with emotions influence the decisions taken [1].

There are two main approaches to characterize emotions: the *categorical* and the *dimensional.* In the categorical approach one looks at emotions as distinct categories. A widely known categorization of emotions has been proposed by Ekman[12] with the six *Ekman basic emotions.* These are *anger, disgust, fear, surprise, happiness,* and *sadness.* The categorization was done in such a way that these six emotions have distinctive facial expressions and physiological responses. In a separate line of research, Zentner et al. [42] conducted a series of studies on music listeners and asked them to describe the emotion they feel on a wide range of musical stimuli. The outcome of these studies was the identification of music-oriented emotional categories, concretely *wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension,* and *sadness*, and the Geneva Emotional Music Scale (GEMS), an instrument for measuring these emotions. The dimensional approach assumes that a single emotion is not a category but a point in a space. The most common dimensions used are *valence, arousal* and *dominance*, as proposed by Russel and Mehrabian [29]. Valence describes whether an emotion is pleasant or unpleasant, arousal describes the strength of the emotion and dominance describes how much we are in control of the emotion.

For each musical piece, Juslin et al. distinguish between *expressed, perceived* and *induced* emotions [19]. The expressed emotions are those that the performer or composer wishes to express. The perceived emotion are the ones perceived, but not necessarily felt, by the listener. For example, a listener might recognize a song as sad while being happy. The induced emotion is the one that is felt by the listener, i.e., the user experiences physiological and expressional reactions. In our work we focus on the induced emotions. We measure the facial expressions of the user, which are the result of the felt emotions, induced by the songs. There are a lot of studies that demonstrate how music influences the emotional responses of listeners. Schedl et al. [32] show that certain audio descriptors related to loudness, timbre, harmony, and rhythm correlate with the perceived emotions. They also show that there are additional variables, such as demographics, expertise, and personality, that influence the type of emotional response to a musical stimulus.

Music and emotions are related also in terms of motivation for the consumption of music. Lonsdale and North have shown that one of the main reasons why people listen to music is emotion regulation, i.e., as a tool to change or keep an emotional state at will. [22, 39].

2.4 Emotions and Preferences

There is evidence that supports the thesis that emotional responses are related to music preferences. Luck et al. [23] conducted an experiment where they let users to listen to music excerpts, give ratings to the excerpts, and dance. The analysis of the data showed a relationship between music preferences and the amount of movement, which was deemed as a proxy for emotional arousal. Furthermore, they observed that personality accounts for the variance in the emotional response as measured through movement.

2.5 Acquisition of Emotions

Emotions can be measured in an explicit or implicit way. Explicit emotion acquisition is done through questionnaires, such as the Self-Assessment Manikin (SAM), which are intrusive and time-consuming [5]. We can take advantage of the fact that emotions are manifested through physiological and bodily responses and expressions and measure these. Research has shown that the manifestation of emotion in humans, especially facial expressions, is quite universal across individuals and cultures [10]. Charles Darwin speculated that the expression of emotions in humans and lower species share a common single origin due to the similarity of emotion expressions across species [8]. Ekman, for example, has devised the Facial Actions Coding System (FACS), a method for assessing the observable facial expressions and mapping them to emotions. For example, happiness is expressed with Action Unit (AU) 6 (contraction of zygomatic major) and AU12 (contraction of the inferior part of orbicularis oculi) [13].

The field of affective computing is developing methods for inferring the emotional state of the user from various bodily signals. This is done using various modalities (i.e. various sensors, such as cameras, heart-rate sensors, skin conductance sensors etc.) [34]. Substantial progress has been made on detecting emotions from facial expressions [35–37, 41] and these technologies have become mature enough to yield robust off-the shelf solutions. In our experiments, we used such a library, the Affectiva Software Development Kit[1]. That library takes as input the video stream of the user's face from a web camera located on a computer and detects a total of 21 facial expressions, six emotions and some additional user characteristics[2], which are summarized in Tab. 1.

2.6 Personality and Music

A user characteristic that is often considered in the analysis of music preferences is personality. Personality accounts for the individual differences in our long-term emotional, interpersonal, experiential, attitudinal and motivational styles [24]. A frequently used model of personality is the five-factor model (FFM), which is composed of the factors: *Openness to new experience, Conscientiousness, Extraversion, Agreeableness* and *Neuroticism.* FFM of users is usually measured using a validated questionnaire, such as the Ten Item Personality Measure (TIPI) [16]. Several studies have shown that personality and music preferences are correlated. Rentfrow et al.[28] found that all five personality factors correlate with music preferences across a wide range of genres. For example, people who score high on Openness tend to prefer more reflective and complex genres such as blues, jazz and classical music. These findings have been confirmed and extended by other similar studies, such as [7, 9]. Another study showed that there is correlation between user characteristics, such as personality and music education, and the emotional perception of music [32]. In a study conducted by Ferwerda et al. [14] the authors induced different emotional states in users and asked them which emotionally-laden music they would prefer to listen to in that situation. The analysis showed that users with different personalities had different preferences. For example, users who scored high on neuroticism and felt disgusted tend to listen more to sad music whereas low-neuroticism users tend to prefer happy music. Similar findings that show that people with different personalities have different emotional responses to the same stimuli have been found by Sachs et al. [30].

2.7 Privacy in User Modeling

In general, users are willing to give up some personal data for getting a personalized service [21]. However, the acquisition of user data poses a privacy threat. The acquired data, i.e. the user profile, can be used to identify the user and thus to reveal preferences for items that may be embarrassing or compromising for the user. In classical recommender systems the user profiles usually consist of ratings for items. Various privacy-preserving techniques have been devised. Berkovsky et al. [2] showed that decentralizing the user profiles can mitigate privacy issues while retaining the accuracy of a collaborative filtering (CF) recommender system. Heitman et al. [17] proposed a portable architecture, where each user carries her own user profile on her own device. This approach limits the access and exchange of user data. Another technique is called differential privacy, which revolves around introducing randomness in the ratings. In their work, Berlioz et al. [3], apply three differential privacy techniques in CF and matrix factorization (MF) recommender systems, and discuss the trade-offs between privacy and accuracy. They found that CF algorithms are more resilient to the introduced randomness, compared to MF, and are more suitable when there are high privacy requirements. In low privacy requirements, however, MF has a better accuracy than CF. The preference prediction method that we present in this paper does not acquire ratings but facial expressions. Although we do not address the privacy aspects experimentally, the same techniques of decentralization and differential privacy could be used in the method we present. Furthermore, if we compare an explicit rating about an item with facial expressions during the consumption of an item, it is clear that the explicit rating reveals more information about the user's attitude towards the item than the facial expressions. We would also like to stress that the proposed approach does not store any images or video of the user, but only system-generated predictions of facial expressions in the form of textual information (see Tab. 1 for an example). These predictions are computed by the used library.

3 EXPERIMENT

In order to build the proposed pairwise preference predictor we first collected observational data. We set up a pre-study surveying

[1] http://developer.affectiva.com/

[2] http://blog.affectiva.com/the-emotion-behind-facial-expressions

Feature	Value
number of faces found	1
gender	Male
glasses	Yes
age	35 - 44
ethnicity	Caucasian
joy*	0.001814206363633275
sadness*	0.02542627975344658
disgust*	0.42628759145736694
contempt*	0.19689859449863434
anger*	0.004459045361727476
fear*	0.0046778046526014805
surprise*	0.19569388031959534
valence*	-0.2734917402267456
engagement	0.07985058426856995
smile	4.354387073135513e-9
innerBrowRaise	0.7271478176116943
browRaise	0.012029489502310753
browFurrow	0.1216883435845375
noseWrinkle	0.006316573824733496
upperLipRaise	0.000019096871255896986
lipCornerDepressor	0.00048577284906059504
chinRaise	0.02014032006263733
lipPucker	0.017056530341506004
lipPress	0.03073771297931671
lipSuck	0.0005146527546457946
mouthOpen	0.01045703049749136
smirk	0.18308372795581818
eyeClosure	5.5701749879233375e-9
attention	98.3320541381836
lidTighten	0.003822572296485305
jawDrop	0.03755816072225571
dimpler	0.0046720667742192745
eyeWiden	0.00891738198697567
cheekRaise	0.00021206788369454443
lipStretch	0.003288324223831296
emoji	:-|

Table 1: List of facial expression raw features $f \in F$ acquired from a webcam video and processed by the Affectiva SDK with example values. The features included in the reduced set SF are marked with an asterisk.

alternative usage scenarios and a main user study where users compared pairs of songs. While they were listening to songs in the main study we collected their facial expressions through a web camera. After listening to each pair of songs, the users provided explicit pairwise scores, which were used as ground truth for training the preference prediction model. We also measured the listening time for each song and used it in a baseline preference predictor. We compared the accuracy of the preference predictions of the proposed and baseline models using Root Mean Squared Error (RMSE), precision, recall, F-measure and accuracy.

3.1 Data Acquisition

3.1.1 Pre-study. We ran a pre-study in order to identify a suitable music listening scenario for the main study. The scenario should be experienced quite frequently and the users should really listen to music in the scenario. We ran the pre-study on a sample of 145 subjects (85 males, mostly under- and post-graduate students). The subjects were asked to provide free-text answers to the following questions: (1) which periods of day do you normally listen to music and in those periods (2) which activities do you perform while listening to music. After coding the answers, we found that the most popular activities accompanied by music listening are *daily cognitive activities*, such as *studying, working, programming, reading* (76% of participants) and *commuting, e.g. car driving, bus riding, cycling, walking* (50 % of participants). We finally chose *working (cognitively demanding tasks)* as the target usage scenario for our main study.

3.1.2 Music selection. For the main study, we have chosen to use the Moodo music dataset [38]. The dataset contains a total of 200 song snippets lasting 15s each. The songs are unknown to the large public, which eliminates the familiarity bias in expressing preferences. They span across a variety of music genres. Moreover, for preference elicitation, the songs were assigned to users randomly, however, some songs were presented more often than others. Namely, we artificially generated a short head/long tail distribution of the probabilities for each song to be presented to a user. The number of occurrences of individual songs are shown in Fig. 2.

There are two reasons for this choice: (i) in real scenarios some songs are actually played more often than others and (ii) ratings in real data sets are not uniformly distributed. In fact, in our future work we will implement a recommender system that will exploit a collaborative filtering algorithm. These algorithms are commonly trained on real usage data sets where songs from a smaller part of the full catalog are more frequently rated than the others. This has the beneficial consequence that when users' profiles are compared, e.g., in user-based collaborative filtering approach, a target user have a larger number of neighbors with whom a similarity can be computed, compared to the case when ratings are uniformly distributed on items.

3.1.3 Main Study. The main study was carried out in a controlled environment. We first explained to the users the goal of the experiment and trained them. We had a total of 75 users (they were different from those involved in the pre-study). The average age was 29.8 years ($SD = 9.5$) and there were 49 males. We asked the users to imagine having to choose music for the usage scenario selected in the pre-study, i.e., *working (cognitively demanding tasks).* By using this approach (to imagine a listening context) we put constraints on the ecological validity of the experiment; users may have overestimated the importance of the scenario. However, we believe that this has no impact on the comparison of alternative pairwise score prediction algorithms, as we have done. Using the web interface presented in [40] and depicted in Fig. 3, each user u was shown two play buttons, one for each song $s = \{l, k\}$, being l the song on the left and k the song on the right. Each user was forced to listen to at least 10 seconds of each song and to adjust the slider into the proper position, depending on their preference. The

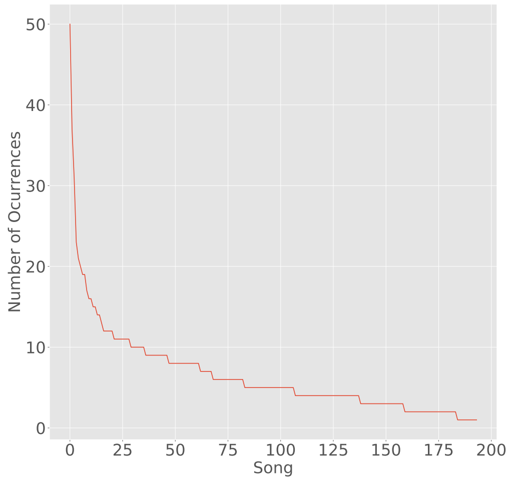
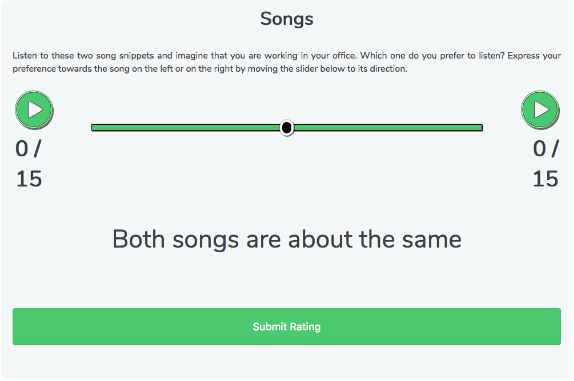


Figure 2: The number of times individual songs have been presented to the users in the experiment.

selected position was converted into a pairwise preference score $p(u, k, l) \in \{-2, -1, 0, 1, 2\}$ where -2 meant *I prefer strongly the left song* and +2 meant *I prefer strongly the right song*. The user could listen to each song multiple times, so the time of user u listening to the song l or k was in the range $t(u, s) = [10, \infty]$. The observed distribution of the listening times is shown in Fig. 4.


Songs

Listen to these two song snippets and imagine that you are working in your office. Which one do you prefer to listen? Express your preference towards the song on the left or on the right by moving the slider below to its direction.

Figure 3: The user interface of the pairwise preference acquisition.

While listening to the songs, the user’s facial expressions were captured by the web camera installed on the laptop. The flow of the acquisition is shown in Fig. 5. The video stream of the facial expressions was sent to the Affectiva SDK for the extraction of facial expression features. For each video frame, the Affectiva SDK returned a set of 47 features related to the recorded facial expression in JSON format. We denote these raw features as $f \in F$. The complete list of raw features F with example values is listed in Tab. 1. The Affectiva SDK managed to process on average 11 frames per second. We stored the raw facial features F, the listening times, the song IDs and the user IDs in a database.

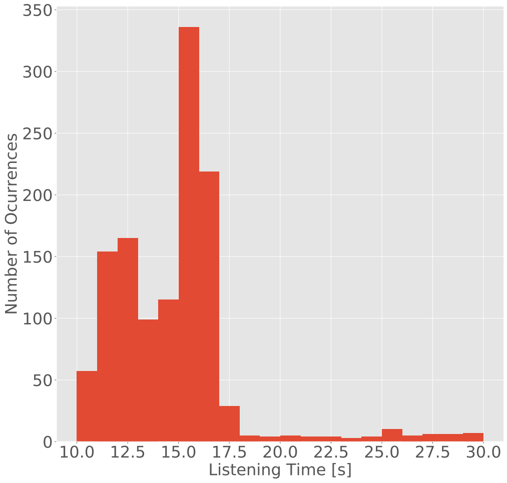


Figure 4: Distribution of listening times. In 3% of the cases the listening times were longer than 30s.

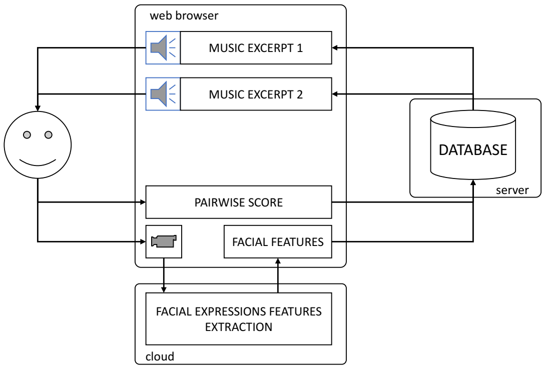


Figure 5: Flow of the data acquisition interaction: the user listens to two music snippets and then provides a pairwise score. During the listening to the music, a web camera streams the video to an API that extracts facial features.

After having listened to the target songs, each user was required to give pairwise scores to at least 10 pairs of songs. Each song could occur only once per user. The users also filled-in two questionnaires: (i) a demographics questionnaire, and (ii) the TIPI questionnaire [16] with ten questions measuring the five personality dimensions of the FFM.

4 PAIRWISE SCORES PREDICTION

We tackled the prediction of pairwise scores both as (i) a regression problem, where we predict numeric pairwise scores $\hat{p}(u, k, l) \in$

$[-2, 2]$ and as (ii) a classification problem, where we predict pairwise scores as alternative discrete class (left is preferred, right is preferred, equally preferred).

4.1 Regression

We designed a model that predicts the pairwise scores $\hat{p}(u, k, l) \in [-2, 2]$ the user u would give to a pair of songs k and l based on the features engineered from the facial expressions. Since this is a regression problem we used RMSE as metric to assess the quality of the prediction.

For each of the 638 pairwise scores provided by the users we had at least 20 seconds of facial features returned by the Affectiva SDK, i.e., at least 10 seconds per song in the pair. Figure 6 shows an example of the time series of the feature *surprise* for a pair of songs. We needed to map the variable-length time series of facial raw features F during listening to a song to a fixed number of features per song-user in order to be able to use them in a regression algorithm. We denote these as engineered features E. It is important to note that the selected features should be invariant with respect to some modification of the raw observations. For instance, small time delays of a raw feature behavior should not affect the engineered features and the regression algorithm.

We experimented with several approaches for aggregating variable-length time series F into fixed length feature vectors, such as average values, standard deviations, peak values, position of the peak values, monotonicity and polynomial fitting. The best performance in terms of prediction accuracy was yielded when we used polynomial fitting and monotonicity features. We performed a second degree polynomial fitting for modeling the raw features' changes during the listening to a song, i.e., we approximate a raw feature with the following polynomial:

$$y^f(u, s, t) = \alpha_s^f + \beta_s^f t + \gamma_s^f t^2 \tag{1}$$

where f is the facial feature $f \in SF = \{joy, sadness, disgust, contempt, anger, fear, surprise, valence\}$, s is either the left or the right song $s \in \{l, k\}$, and α_s^f, β_s^f and γ_s^f are the parameters to fit (which become the new set of engineered features E). We note that in the equation above, t is the time and y is the fitted value of an f feature.

When determining the α, β and γ engineered features, we tested several subset of the raw features from Tab. 1 and found that the α, β and γ engineered features extracted from the subset of features $f \in SF \subset F$ yield results that are as good as the ones from the full set of features F (i.e. including other features from Tab 1, such as smile, browRaise etc.). The values α, β and γ (for each raw feature) were used as features for the prediction of the pairwise score. We also calculated the differences between the α, β and γ values of the left song l and the right song k (for the same raw feature) thus obtaining additional engineered features for training the predictor of the pairwise score. We denoted these features as $\Delta\alpha^f$, $\Delta\beta^f$ and $\Delta\gamma^f$. Besides the engineered features α^f, β^f, γ^f, and their respective differences, we calculated also features that describe the monotonic relationship between time and the facial features SF. To do that we used the Spearman's Rank-Order Correlation. We denote it with ρ_s^f. We also calculated the difference of the ρ_s^f of the left and the right song and we denoted it with $\Delta\rho^f$. So, we state again that in the predictor we did not use the raw features $f \in SF \subset F$ but the engineered features $E = \{\alpha_s^f, \beta_s^f, \gamma_s^f, \Delta\alpha^f, \Delta\beta^f, \Delta\gamma^f, \rho_s^f, \Delta\rho^f\}$ for each $f \in SF$ calculated from the raw features SF.

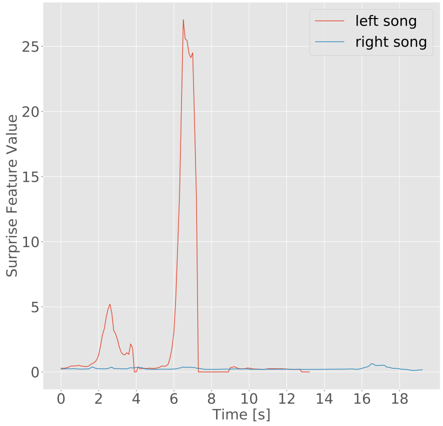


Figure 6: Example of the time series of the *surprise* feature for a pair of songs.

Feature	Absolute Correlation
$\Delta\beta^{contempt}$	0.277868
$\Delta\beta^{valence}$	0.262055
$\beta_l^{contempt}$	0.242743
$\Delta\beta^{joy}$	0.229939
$\Delta\rho^{sadness}$	0.215670
$\rho_l^{sadness}$	0.202872
$\Delta\rho^{valence}$	0.202689
β_l^{joy}	0.196904
$\beta_l^{valence}$	0.189589
$\Delta\alpha^{joy}$	0.175126
$\Delta\beta^{sadness}$	0.173816
$\rho_l^{valence}$	0.167465
$\Delta\rho^{joy}$	0.165396
$\Delta\alpha^{contempt}$	0.164943

Table 2: Absolute values of correlations between features and pairwise scores.

We have calculated the correlations between each new engineered feature and the pairwise score and found several significant correlations. The features with the strongest correlation were those related to contempt, sadness, joy, and valence. They are reported in Tab. 2. As expected, the engineered features with higher correlation are those that model differences of SF features. There are, however, some features related to the left song, i.e. $s = l$. We speculate that this is related to the fact that in most cases (79%) the users started to listen first to the left song.

After having performed the above described feature engineering step we performed the prediction task by using the features describing the facial expression of the user while comparing two songs to predict the pairwise score of the two songs. We ended up, for each pair of songs, with 96 engineered features: 12 engineered features for each of the eight $f \in SF$ features, i.e. α, β, and γ features for the left song, α, β, and γ features for the right song, three Δ features, a ρ for the left song, a ρ for the right song and a $\Delta\rho$. We experimented with several prediction algorithms and ended up with considering Random Forest and the Gradient Boosting that yielded the best results. We used stratified splitting in order to preserve the distribution of the users' data in the training and test sets. This was done by keeping, for each user, approximately 60% of her comparisons in the training set and the remaining 40% in the test set. Then we used random sampling from each stratum with 60% in the training set and 40% in the test set. We repeated this procedure 5 times and averaged the results. Each time we optimized the hyper-parameters using five-fold cross validation on the training set.

4.1.1 Baseline Predictor. Our method for predicting pairwise scores from implicit signals is novel and there is little prior work to compare with. Popularity features, such as playcount, do not fit in our experiment design because each user has listened to each song at most one time. The global playcounts are also not suitable because they were enforced artificially in order to ease our future work on recommendations. We chose Random Forest and Gradient Boosting baseline predictors that used the duration of the listening to each song and the difference in the duration between the two songs in the pair. The Spearman's Rank Correlation Coefficient between the duration differences in listening times (non-normal distribution) in a pair of songs is correlated with the pairwise scores (discrete and non-normal distribution). The correlation is $r = 0.156$ with the p-value $p = 0.000025$. This means that the bigger the difference in listening times (i.e. the user listens to one song from the pair longer than to the other song) the more the user is likely to prefer the song she listened longer to.

4.1.2 Results. The results of the predictions are reported in Tab. 3. The RMSE of the predictors that use the proposed facial features are lower than the RMSE of the baseline methods which use listening time features. Hence, although the proposed method requires to collect the facial features it pays off with a substantially better prediction accuracy.

Reggressor	Features	RMSE
Random Forest	Facial Features	1.06
Gradient Boosting	Facial Features	**1.04**
Random Forest	Listening Time	1.25
Gradient Boosting	Listening Time	1.27

Table 3: RMSE of the tested predictors using facial features (the proposed method) and listening time features (baseline).

4.1.3 Exploratory Analysis. Besides the global prediction model, we explored whether the music preferences of some user groups are easier to predict than those of others. To achieve this we split the users in two groups, trained two separate models (one for each group) and compared the RMSE using the t-test. We perform the splitting in two groups several times, each time along one of the five personality factors. We used median splitting on each factor to get the two groups of users that we compared.

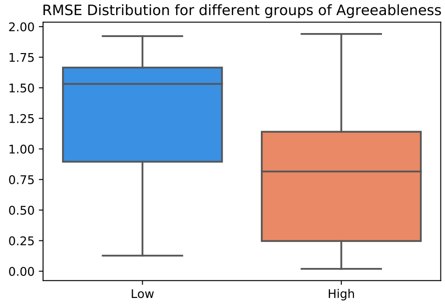


Figure 7: RMSE distribution for users with high and low agreeableness.

We found that there were significant differences in the mean RMSE of the split groups when the splitting was done on agreeableness, conscientiousness, and openness. For each group we used the predictor that yielded the best result. The results are summarized in Tab. 4 and shown in Figs. 7 through 9. We speculate that users, who score low on agreeableness, high on conscientiousness and/or low on openness, tend to either show less emotions through their facial expressions or have generally lower variance in their preferences.

	FaceFeatures GradBoost	FaceFeatures RandForset	Baseline
High Openness	**0.91**	0.99	0.99
Low Openness	**1.33**	1.39	1.58
High Conscientiousness	**0.97**	0.99	1.05
Low Conscientiousness	1.43	**1.37**	1.51
High Agreeableness	0.96	**0.94**	0.98
Low Agreeableness	**1.39**	1.40	1.68
Global Model	**1.05**	1.07	1.26

Table 4: RMSE of the prediction of the pairwise score on the scale from -2 to 2 for different groups of users. Bold values represent the best (i.e. lowest) RMSE score among the three considered prediction methods.

4.2 Classification

In a second set of experiments we designed a classification model that predicts pairwise scores $\hat{p}(u, k, l) \in \{-1, 0, 1\}$ as a three-class classifier. The users in the main study assigned pairwise scores as discrete values $p(u, k, l) \in \{-2, -1, 0, 1, 2\}$. We mapped the scores $\{-2, -1\}$ to -1 and $\{1, 2\}$ to 1 for two reasons: (i) we wanted to transform the regression problem to a classification one where the classes are -1: *k is preferred to l*, $+1$: *l is preferred to k*, and 0: *no one*

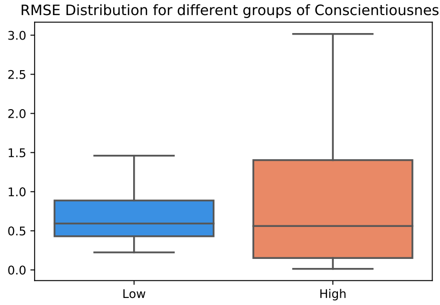


Figure 8: RMSE distribution for users with high and low conscientiousness.

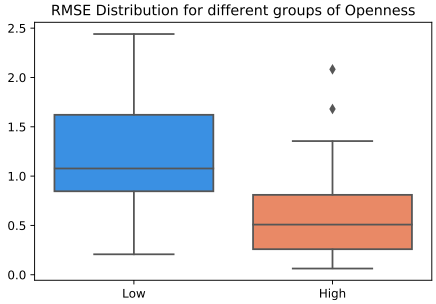


Figure 9: RMSE distribution for users with high and low openness.

is preferred and (ii) the distribution among the classes was skewed by having much more elements in the class 0 than in the other classes.

We considered the same features used in the regression problem. We also used the same data splitting technique. The results are reported in Tab.5. Since the classification was in three classes, we calculated the aggregated precision, recall and F-measures using a weighting scheme. The average precision, recall and F-measure were calculated by weighting these scores of each class by the number of true instances for each class to account for class imbalance. We note that in this setting F-measure may not take a value between precision and recall [15].

We note that Gradient Boosting with facial features performs the best in terms of accuracy, recall and F-measure while the Random Forest classifier is slightly better in terms of precision.

5 CONCLUSION AND FUTURE WORK

In this paper we have proposed a new approach for using implicit preference signals to infer precise pairwise preferences in the form of pairwise scores of the user. Our approach exploits features extracted from the analysis of raw features describing the time evolution of the facial expressions captured while the user is listening to the compared songs. Compared to a baseline method, which uses listening time to predict pairwise preferences, our method has a lower RMSE, when the problem is modeled as a regression task, and higher precision, recall, F-measure and accuracy when it is modeled as a classification task.

Classifier	Features	Accuracy	Precision	Recall	F-Measure
Majority Classifier	None	0.476	0.227	0.476	0.307
Random Forest	Facial Features	0.642	**0.617**	0.642	0.610
Gradient Boosting	Facial Features	**0.646**	0.616	**0.646**	**0.617**
Random Forest	Listening Time	0.593	0.545	0.593	0.557
Gradient Boosting	Listening Time	0.593	0.545	0.593	0.557

Table 5: Accuracy, precision, recall, and f-measure of classifiers using facial features (the proposed method) and listening time features (baseline).

These results support the working hypothesis of considering facial expression and pairwise scoring prediction as a viable solution for preference elicitation tasks, especially in forthcoming scenarios of affective interactions with computers and robots, which are already a reality in many emerging domestic applications.

We are aware that privacy is an important issue of the proposed approach. Even if we do not store any image of the user, images are taken in order to extract facial expression features. Although we did not address the potential privacy issues raised by our approach, we surveyed existing work on privacy and suggested solutions that could be adopted to tame them.

Interestingly, we have shown that there are several facial expressions-derived features that correlate well with the user’s pairwise preferences. Furthermore, we observed that personality factors account for differences in the accuracy of prediction, and certain type of users may be better served with the proposed solution for preference elicitation.

We finally note that we plan to use the proposed pairwise scores prediction method in a pairwise recommender system for music. The goal is to reduce the number of explicit pairwise scores that the system must elicit in order to obtain a given accuracy. We are already working on an algorithm that combines explicitly acquired pairwise scores with implicitly acquired ones. We will evaluate the impact of the here presented pairwise score predictor on the performance of the recommender system. We will use ranking-based metrics as we did in our prior work [20].

REFERENCES

[1] Antoine Bechara, H Damasio, and Antonio R Damasio. 2000. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex (New York, N.Y. : 1991)* 10, 3 (mar 2000), 295–307. http://www.ncbi.nlm.nih.gov/pubmed/10731224

[2] Shlomo Berkovsky, Yaniv Eytani, Tsvi Kuflik, and Francesco Ricci. 2007. Enhancing privacy and preserving accuracy of a distributed collaborative filtering. *Proceedings of the 2007 ACM conference on Recommender systems - RecSys ’07* (2007), 9. https://doi.org/10.1145/1297231.1297234

[3] Arnaud Berlioz, Arik Friedman, Mohamed Ali Kaafar, Roksana Boreli, and Shlomo Berkovsky. 2015. Applying Differential Privacy to Matrix Factorization. *Proceedings of the 9th ACM Conference on Recommender Systems - RecSys ’15* (2015), 107–114. https://doi.org/10.1145/2792838.2800173

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.*

[5] M.M. Bradley and P.J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59. http://www.sciencedirect.com/science/article/pii/0005791694900639

[6] Ulf BÃűckenholt. 2001. Thresholds and Intransitivities in Pairwise Judgments: A Multilevel Analysis. *Journal of Educational and Behavioral Statistics* 26, 3 (2001), 269–282. https://doi.org/10.3102/10769986026003269

[7] Tomas Chamorro-Premuzic and Adrian Furnham. 2007. Personality and music: can traits explain how people use music in everyday life? *British journal of psychology (London, England : 1953)* 98 (may 2007), 175–85. https://doi.org/10.1348/000712606X111177

[8] Charles Darwin. 1872. *The expression of the emotions in man and animals*. D. Appleton and company. 5–196 pages.

[9] P. G. Dunn, B. de Ruyter, and D. G. Bouwhuis. 2012. Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music* 40, 4 (2012), 411–428. https://doi.org/10.1177/0305735610388897

[10] Paul Ekman. 1993. Facial expression and emotion. *American Psychologist* 48, 4 (1993), 384. http://psycnet.apa.org/journals/amp/48/4/384/

[11] Paul Ekman. 1994. Moods, Emotions, and Traits. , 56–58 pages.

[12] Paul Ekman. 2005. Basic Emotions. In *Handbook of Cognition and Emotion*, Tim Dalglesish and Mick J. Power (Eds.). Number 1992. John Wiley & Sons, Ltd, Chichester, UK, 45–60. https://doi.org/10.1002/0470013494.ch3

[13] Paul Ekman, Wallace V. Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of Personality and Social Psychology* 39, 6 (1980), 1125–1134. https://doi.org/10.1037/h0077722

[14] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Personality & Emotional States : Understanding Users ' Music Listening Needs. In *UMAP 2015 Extended Proceedings*, Alexandra Cristea, Judith Masthoff, Alan Said, and Nava Tintarev (Eds.). http://ceur-ws.org/Vol-1388/

[15] Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative Methods for Multi-labeled Classification. 22–30. https://doi.org/10.1007/978-3-540-24775-3_5

[16] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (dec 2003), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

[17] Benjamin Heitmann, James G. Kim, Alexandre Passant, Conor Hayes, and Hong-Gee Kim. 2010. An architecture for privacy-enabled user profile portability on the web of data. *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '10* (2010), 16–23. https://doi.org/10.1145/1869446.1869449

[18] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling User Preferences in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (jun 2014), 1–26. https://doi.org/10.1145/2512208

[19] Patrik N. Juslin and Petri Laukka. 2004. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research* 33, 3 (2004), 217–238. https://doi.org/10.1080/0929821042000317813

[20] Saikishore Kalloori, Francesco Ricci, and Marko Tkalcic. 2016. Pairwise Preferences Based Matrix Factorization and Nearest Neighbor Recommendation Techniques. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16* (2016), 143–146. https://doi.org/10.1145/2959100.2959142

[21] Bart P. Knijnenburg and Shlomo Berkovsky. 2017. Privacy for Recommender Systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17* (2017), 394–395. https://doi.org/10.1145/3109859.3109935

[22] Adam J Lonsdale and Adrian C North. 2011. Why do we listen to music? A uses and gratifications analysis. *British journal of psychology (London, England : 1953)* 102, 1 (feb 2011), 108–34. https://doi.org/10.1348/000712610X506831

[23] Geoff Luck, Suvi Saarikallio, Birgitta Burger, Marc Thompson, and Petri Toiviainen. 2014. Emotion-driven encoding of music preference and personality in dance. *Musicae Scientiae* 18, 3 (2014), 307–323. https://doi.org/10.1177/1029864914537290

[24] Robert R McCrae and Oliver P John. 1992. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality* 60, 2 (1992), p175 – 215.

[25] Omar Moling, Linas Baltrunas, and Francesco Ricci. 2012. Optimal radio channel recommendations with explicit and implicit feedback. In *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*. ACM Press, New York, New York, USA, 75. https://doi.org/10.1145/2365952.2365971

[26] Denis Parra and Xavier Amatriain. 2011. Walk the Talk. In *UMAP 2011*, Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver (Eds.). Lecture Notes in Computer Science, Vol. 6787. Springer Berlin Heidelberg, Berlin, Heidelberg, 255–268. https://doi.org/10.1007/978-3-642-22362-4_22

[27] Steffen Rendle and Christoph Freudenthaler. 2009. BPR: Bayesian personalized ranking from implicit feedback. *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009* (2009), 452–461. http://dl.acm.org/citation.cfm?id=1795167

[28] Peter J. Rentfrow and Samuel D. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

[29] JA Russell and A Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality* 294 (1977), 273–294. http://www.sciencedirect.com/science/article/pii/009265667790037X/

[30] Matthew E Sachs, Antonio Damasio, and Assal Habibi. 2015. The pleasures of sad music: a systematic review. *Frontiers in Human Neuroscience* 9, July (2015), 404. https://doi.org/10.3389/fnhum.2015.00404

[31] Markus Schedl, Arthur Flexer, and Julián Urbano. 2013. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems* 41, 3 (jul 2013), 523–539. https://doi.org/10.1007/s10844-013-0247-6

[32] Markus Schedl, Emilia Gomez, Erika Trent, Marko Tkalcic, Hamid Eghbal-Zadeh, and Agustin Martorell. 2017. On the Interrelation between Listener Characteristics and the Perception of Emotions in Classical Orchestra Music. *IEEE Transactions on Affective Computing* (2017), 1–1. https://doi.org/10.1109/TAFFC.2017.2663421

[33] K. R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (dec 2005), 695–729. https://doi.org/10.1177/0539018405058216

[34] Björn W Schuller. 2016. Acquisition of Affect. In *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir (Eds.). Springer International Publishing, Cham, 57–80. https://doi.org/10.1007/978-3-319-31413-6_4

[35] Mohammad Soleymani, Sadjad Asghari Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 3045, c (2015), 1–1. https://doi.org/10.1109/TAFFC.2015.2436926

[36] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14. https://doi.org/10.1016/j.imavis.2017.08.003

[37] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing* 3, 2 (apr 2012), 211–223. https://doi.org/10.1109/T-AFFC.2011.37

[38] Gregor Strle, Matevž Pesek, and Matija Marolt. 2016. Towards User-Aware Music Information Retrieval: Emotional and Color Perception of Music. In *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir (Eds.). 327–353. https://doi.org/10.1007/978-3-319-31413-6_16

[39] Myriam V Thoma, Stefan Ryf, Changiz Mohiyeddini, Ulrike Ehlert, and Urs M Nater. 2012. Emotion regulation through listening to music in everyday situations. *Cognition & emotion* 26, 3 (jan 2012), 550–60. https://doi.org/10.1080/02699931.2011.595390

[40] Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2017. A Research Tool for User Preferences Elicitation with Facial Expressions. In *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*. ACM Press, New York, New York, USA, 353–354. https://doi.org/10.1145/3109859.3109978

[41] Marko Tkalčič, Ante Odić, and Andrej Košir. 2013. The impact of weak ground truth and facial expressiveness on affect detection accuracy from time-continuous videos of facial expressions. *Information Sciences* 249 (nov 2013), 13–23. https://doi.org/10.1016/j.ins.2013.06.006

[42] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion (Washington, D.C.)* 8, 4 (2008), 494–521. https://doi.org/10.1037/1528-3542.8.4.494