Technische Universität Berlin

Institut für Sprache und Kommunikation im Fachgebiet Audiokommunikation

Fakultät I - Geistes- und Bildungswissenschaften
Marchstraße 23, 10587 Berlin

Master Thesis

# Prediction of Audio Features of Self-Selected Music by Situational and Person-Related Factors

Towards a Context-Aware Music Recommendation System

Christoph Karnop

[redacted]

01.März 2019

Erstgutachter: Prof. Dr. Stefan Weinzierl

Zweitgutachter: Dr. Jochen Steffens

# Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und verstanden habe. Die im betroffenen Fachgebiet üblichen Zitiervorschriften sind eingehalten worden. Eine Überprüfung der Arbeit auf Plagiate mithilfe elektronischer Hilfsmittel darf vorgenommen werden.

________________________________

Datum/ Unterschrift

# Abstract

Music listening behavior has changed by means of the latest technological advances like widespread use of smartphones, streaming-services, and headphones. This allows for the active selection of music in any situation to accomplish personal goals. Still, it is not well understood what implications that might have on the selection of music with certain objective characteristics (audio features). This work aimed at investigating relationships between music-selection behavior and its corresponding functional and situational context of music listening as well as person-related factors. Therefore, a dataset was evaluated that included 1021 real-life listening situations of 101 participants assessed in an *Experience Sampling Method*. For each listening situation the audio features *valence, energy, danceability* and *loudness* of the corresponding song were retrieved from the *Spotify Web API*. Various model selection procedures (*protocols)* all based on the *percentile-Lasso* were assessed by *Nested Cross-Validations.* Finally, the best generalizing *protocols* were used to select the *$L_1$-penalized Linear Mixed Effects Models* to predict the audio features. Replicating several related works, the results suggested that the music-selection behavior is highly situational and music listening functions play an important role in predicting musical properties of self-selected music. Implementing these findings in context-aware music recommendations could benefit users' satisfaction and also enable the assessment of the real-life significance of personal and situational factors and the functional use of music listening.

*Keywords: Music-Selection Behavior, Experience Sampling Method, Linear Mixed Effects Models, Percentile-Lasso, Nested Cross-Validation, Spotify Audio Features, Context-Aware Music Recommendation*

# Zusammenfassung

Das Musikhörverhalten hat sich durch den technologischen Wandel von Smartphones, Musik-Streamingdiensten und Kopfhörern stark verändert. Es ist nun möglich, Musik in jeder erdenklichen Situation aktiv auszuwählen, um persönlichen Bedürfnissen gerecht zu werden. Bis jetzt ist jedoch noch nicht klar, welche Auswirkungen dies auf die Auswahl von Musik mit bestimmten objektiven Charakteristika (Audio Features) hat. Diese Arbeit zielte darauf ab, Beziehungen zwischen dem individuellen Musikauswahlverhalten und dem dazugehörigen funktionalen und situativen Kontext und persönlichen Eigenschaften des Hörers zu untersuchen. Dafür wurde ein Datensatz mit 1021 Alltagssituationen des Musikhörens von 101 Teilnehmern analysiert, der mittels einer *Experience Sampling Methode* erfasst wurde. Für jede Situation wurden die Audio Features *Valence, Energy, Danceability* und *Loudness* des dazugehörigen Liedes von der *Spotify Web API* abgerufen. Verschiedene Modellauswahlverfahren (*Protokolle*), die alle auf dem *Perzentil-Lasso* basierten, wurden mit Hilfe von *verschachtelten Kreuzvalidierungen* auf ihre Generalisierbarkeit hin untersucht. Das jeweils beste *Protokoll* wurde für die Erstellung der *$L_1$-penalisierten Linearen Gemischten Modelle* ausgewählt, um die Audio Features vorherzusagen. Die Ergebnisse replizierten vorherige Erkenntnisse aus anderen Untersuchungen und zeigten, dass das Musikauswahlverhalten stark von situativen Variablen abhängt und Musikhörfunktionen in der Lage sind, objektive Musikeigenschaften vorherzusagen. Diese Ergebnisse können in kontext-basierten Musikempfehlungen eingesetzt werden, um die Nutzerzufriedenheit zu erhöhen, sowie die hier gefundenen Zusammenhänge auf ihre Alltagsrelevanz hin zu überprüfen.

---

*Schlüsselwörter: Musikauswahlverhalten, Experience Sampling Methode, Lineare Gemischte Modelle, Perzentil-Lasso, Verschachtelte Kreuzvalidierung, Spotify Audio Features, Kontextbewusste Musikempfehlung*

# Content

# Figures

# Tables

# 1 Introduction

The existence of music is a very fascinating reality and its implications go far beyond current human understanding. Theories about the origin of music are highly speculative but evolutionists generally agree on the importance of music in the development process of humanity. To date, there is no doubt about the immense value that music has for all people's lives across the world and in all different cultures. The omnipresence of music in all aspects of human history and present therefore is the ultimate manifestation that musicality is a fundamental human property. Shedding light into all those implicit questions regarding the nature of music involves the profound understanding of the nature of mankind and will always continue being a thriving research area in various disciplines. Music consumption has considerably changed throughout the time and nowadays it is increasingly controlled by modern technological developments. In contrast to the past, people are now able to listen and select music in any imaginable situation. Thus, questions regarding aspects of music-selection behavior have never been more relevant than they are now. Considering the trends of our modern digital way of living and the consequential constant information flow we are exposed too, intelligent applications for supporting our personal needs are about to become more and more valuable. Investigating the factors of modern music-selection behavior and deriving models to accurately determine and predict our music listening goals are the next steps to an adaptive, intelligent music recommendation that supports modern technological challenges and fulfills our ever changing existence.

This work considers research findings regarding contextual and psychological aspects of everyday music listening to build prediction models of objective musical characteristics (audio features) based on person-related, situational and functional factors of music-selection behavior. The following sections present the theoretical background regarding the motives and goals of music listening as well as predictive factors that describe our everyday music-selection behavior. Afterwards, facets of musical content are discussed by explaining the discrepancy of modern music information approaches and psychologically relevant effects in the course of music listening. This discrepancy is generally referred to as the *semantic gap*. In the end, state-of-the-art statistical modeling procedures as well as best practices of music recommendation systems are presented. All those aspects build the foundation of the pursued modelling process of this work and hint to future directions.

## 1.1 Predictors of Music-Selection Behavior in Everyday Life

As Mithen (2005) argues, musicality has probably already appeared in very early human evolution stages as part of a pre-linguistic multi-modal communication style, the so-called *Hmmmmm* (holistic, manipulative, multi-modal, musical and mimetic). In accordance with his theories, *Neanderthals* and other hominin ancestors might have used their musicality for emotion expression and induction as well as for social communication and especially for mother-infant communication (Mithen, 2005). This theory is in line with a lot of other (theoretical) evolutionary and anthropologic explanation approaches about the functionality of music listening as being a parallel way of communication besides proper language (Huron, 2001; Bicknell, 2009; Falk, 2004; Longhi, 2008) and its emotional and cognitive use (Schubert, 2009; Hargreaves & North, 1993). The referenced works only constitute a fraction on the amount of theories about the origin and development of musical functions, which are partly based on vague speculations. Notably though, they are not very far off from some empirical findings about modern music listening functions. Probably the most comprehensive overview of possible musical functions was laid out by Schäfer, Sedlmeier, Städtler, & Huron (2013) - ranging from theoretical explanation approaches to empirical listening behavior studies. Since their work integrated the most important empirical findings, specific research is not named here individually, but special focus is put on the outcome conveyed by their study. After aggregating all possible functions of music listening that were found in the existing literature, Schäfer and colleagues asked 834 people to rate on the functionality of music by 129 items. The *principal component analysis* (PCA) that was conducted afterwards suggested that three dimensions mainly accounted for the variance (40%) in the level of agreement with those items that were rated by the participants. Those are namely the *regulation of arousal and mood*, *achievement of self-awareness* and *expression of social relatedness*. Potentially, those functionalities of music listening are either person-related or situation dependent (Greb, Schlotz, & Steffens, 2017). Based on technological trends like widespread use of smartphones, headphones, portable speaker devices and music streaming services, it seems plausible to say that people are more than ever before able to select and listen to music in just any kind of situation. While the situational influences have considerably changed following the modern technological trends (North, Hargreaves, & Hargreaves, 2004), evolutionary explanation approaches still account to a certain extend for the person-related functional use of music listening, considering the results of Schäfer et al. (2013). Despite the fact that modern everyday music listening takes place in nearly any situation of the modern life, so far most research has focused on personal factors. Less research has investigated situational impact and almost no studies have combined both dimensions at the same time, as stated by Greb and colleagues (2017). Summing up, so far the list of investigated personal factors on the functional use of music includes basic variables like age, gender and nationality, more complex concepts like personality traits, intelligence, physical and mental health as well as more music specific factors like musical expertise, musical preferences and strength of preferences for style and genre (Schäfer, 2016). Situational variables of interest have been the location of music listening, the core activity besides music listening, the presence of others, the level of choice, the mode of presentation, the momentary mood, and time of day (North, Hargreaves, & Hargreaves, 2004; Greasley & Lamont, 2011; Krause, North, & Hewitt, 2016; North & Hargreaves, 2000; North & Hargreaves, 1996).

From investigating person-related and situational factors separately it does not become clear how much those facets actually contribute to the functional use of music. Thus, to quantify the relative impact of individual differences of listeners and situational variables on the functionality of music listening, Greb and colleagues (2017) asked participants about three typical listening situations they could remember. Person-related questions (gender, *Big Five*, musical taste and others) and items describing the situation and functionality of music listening were answered. Explained more in detail, the situation was meant to be described by each individual first in an open response format in retrospective. After the initial open description of three typical listening situations, participants answered situational questions regarding the presence of others, the level of choice of selecting the music, the location where this situation would usually occur, the momentary mood and the importance of this mood for the decision to listen to music, the time of day, the attention paid to the music in this situation, and finally the frequency of occurrence of this listening situation in everyday life. Musical functions were mainly derived from the work of Schäfer et al. (2013) and were rated by twenty-two items on a 7-point rating scale for each situation ("I listen to music because… " 1 = Not at all, up to 7 = Completely). In contrast to most other researches, this work resulted in an integrated survey in which situational, person-related and functional variables were investigated at the same time. The open responses were evaluated in the following manner: at first, listening situations were classified by 11 activity categories (e.g. being on the move, housework, working & studying…). Listening locations were assigned to seven categories (e.g. at home, workplace, transportation vehicle…). Activities and locations were then found to be highly correlating, therefore listening locations were excluded from further analyses. A factor analysis on the 22 items of functions of music listening finally yielded a five-factor solution. Those factors were labeled respectively as: *intellectual stimulation*, *mind wandering & emotion involvement*, *motor synchronization & enhanced well-being*, *updating one's musical knowledge*, and *killing time & overcoming loneliness*. The hierarchical data structure containing three distinct measurements of listening situations nested in each participants also allowed the analyses of *within-subject* (W-S) predictors and *between-subject* (B-S) predictors for different functions of music listening. W-S differences can be referred to as situational influences, whereas B-S differences arise from personal differences between the participants. *Intra-class correlation coefficients* (ICC) based on *random intercept-only* models delivered insights on the proportion of variance of music listening functions explained by the grouping factor (participants) with the result that in average, *between-subject* differences accounted for 36% of the variance of the functions of music listening. Accordingly, in average 64% of variance was explained by *within-subject* differences, meaning that the functions of music listening are mainly controlled by situational factors. The explained variance by W-S predictors and B-S predictors varied across the different factors, therefore different functions might be influenced more or less by the situation or personal factors. These findings strongly indicate that functions of music listening are highly situational and need to be considered in future experiments. Finally, all variables were fitted in a *Linear Mixed Effects Regression Model* and the most important variables were selected by a variable selection procedure based on a backward elimination procedure to predict the five functionalities. The contribution of the different predictors is explained thoroughly for each factor in Greb et al. (2017). The models explained on average 34% of the variance components of the functionalities, which differed between 29% and 42% for the different functions. It is important to note that the most important *within-subject* predictors were the core activity in the course of music listening, the possibility to choose the music and the corresponding degree of attention paid in this situation to the music. For each factor, at least one activity was shown to have an individual effect. The results of Greb and colleagues supported different earlier findings, e.g. as conveyed by Greasley and Lamont (2011). They found out that music is very differently used by individuals in the context of different activities. Some people, such as engaged music listeners, would chose music to enhance certain activities, while others would never listen to music to accompany another activity. This certainly indicates personal differences in the functional use of music, but as Greb and colleagues as well as earlier findings of Krause and colleagues (2016) and North et al. (2004) suggest, the functionality of music listening is also significantly controlled by the situation, leading to the conclusion that more work is needed to investigate the situational influences of music listening behavior and the functionality of music listening.

Knowing about the functions of music listening behavior does not directly incorporate the knowledge about the implications different listening functions have on the selection of music. And so far, little research has shed light in the complex interactions of functional use of music listening considering situational and person-related variables and the proper music-selection behavior. Most prominently, the works of Krause and North (2017), Randall and Rickard (2017) and Greb and colleagues (2018; 2019) may be named to have investigated both aspects. Krause and North (2017) tried to predict music listening situations and different judgments of the music, such as level of choice, liking, engagement and experienced arousal from the music, by personal and situational variables. Those aspects were related to the PAD dimensions (*pleasure, arousal* and *dominance*) of Mehrabians and Russels (1974) *interaction and interpretation theory* of contextual parameters by people and were now applied in the music listening context. Krause and North (2017) revealed connections of personal and situational variables with the presence of music in contrast to be not present (e.g. music importance ratings, average listening hours, music education level, time of day, location, and activities). Also judgements of the current listening situation were found to be dependent on personal and situational variables. Different locations, activities and also listening devices were associated with different properties of the music listening situation and influenced the liking, level of choice, engagement and *arousal* in the situation. The results gave evidence to the notion that contextual and personal features are playing an important role in everyday listening situations and its perception, and therefore also might contribute to the music-selection procedure. Most prominently mentioned by Krause and North (2017) is the importance of the *dominance* dimension that was tested according to Mehrabian and Russel (1974) and showed relations in peoples experiences with music listening situations. This can be related to modern technological use of playback devices that support the active choice of music at any imaginable place and time and therefore supports the perception of *dominance* in the situation. Randall and Rickard (2017) have investigated the personal music listening behavior rather with regard to its emotional functionality by listener-related and contextual predictors. As many researches have shown before, people use music for emotion regulation purposes and mood enhancement (Schäfer, Sedlmeier, Städtler, & Huron, 2013; Greb, Schlotz, & Steffens, 2017), but the factual outcome of the intended use might have diverse implications. Investigating how situational and person-related variables manipulate the emotional response to music was one important goal of Randall's and Rickard's work. They found out that the initial affective state was the strongest predictor for a corresponding emotional state regulation towards a neutral position, both for *valence* as well as *arousal* of the mood[1]. While in the case of *arousal* those results supported the theories of *arousal regulation* as a listening function, it was rather surprising that people also seek towards neutral *valence* although being initially more positive. Randall and Rickard therefore argued that personal music listening behavior generally is rather being used to return to an emotional state of equilibrium. They could also show that selected music had mostly congruent properties to the initial mood state and those properties were able to reduce the effect of neutralization while music listening. This mood-congruency principle has already found other theoretical support (Skånland, 2013; Thoma, Ryf, Mohiyeddini, Ehlert, & Nater, 2012). In some critical cases this might also sustain a negative initial mood in the short term but potentially provide mood enhancement in longer terms, which would require more time-dependent research to clarify this theory (Randall & Rickard, 2017). Furthermore they found that a higher cognitive functionality of music listening was negatively related to the perceived musical *valence* (negative valence), which indicates that listening behavior follows certain music listening functions. As a critical conclusion of the research, Randall and Rickard showed that the emotional outcome of music listening is mostly determined by contextual factors, which is in line with Greb et al. (2017). Again, this supports the need for more ecologically validated research methods capturing contextual and person-related variables at the same time. Probably the most comprehensive approach of predicting music listening and selection behavior was recently conducted by Greb, Steffens and Schlotz (2018). They asked participants to report three typical music listening situations and answer questions regarding the situation (e.g. presence of others, level of choice, time of day, initial mood), functions of music listening (as defined by Greb and colleagues

[1] The terms *valence* and *arousal* refer to the dimensional emotion representation by Russel (1980). *Valence* indicates the pleasantness (positive versus negative) and *arousal* describes the intensity of an emotion (low versus high).

(2017)), musical characteristics and person-related items (e.g. gender, age, *Big Five*, musical taste etc.). Seven musical characteristics were rated on a 7 point bipolar rating scale by the following qualities: *calming vs. exciting, less melodic vs. melodic, less rhythmic vs. rhythmic, slow vs. fast, sad vs. happy, simple vs. complex,* and *peaceful vs. aggressive*. In the course of predicting music-selection behavior, they only investigated situations that indicated an active music-selection. A variable selection method, the *percentile-Lasso* (Roberts & Nowak, 2014), was used to select the most important contextual and person-related predictors for the subjectively rated music properties. Again, Greb and colleagues were able to show a bigger impact of *within-subject* differences on music-selection behavior than for *between-subject* differences (in average 77% versus 23%), underlining the crucial importance of situational characteristics on the music listening behavior. Similar to the previous research that Greb et al. presented, they found varying relative impact of situational and person-related predictors on musical properties as well as different numbers of predictors, meaning some perceived characteristics were mainly affected by situational variables, others by personal variables. Because functions of music listening were part of all prediction models, the results underlined their importance as predictors for music-selection behavior. With varying impact, the most important situational predictors were the degree of attention to the music, current activity and presence of others, similar to Randall & Rickard (2017). A relevant person-related predictor was the musical taste as an important one, while personality traits (*Big Five*) were rather underrepresented in the prediction models. Greb et al. (2019) then went on investigating music-selection behavior by firstly, developing an *Experience Sampling Method* (ESM) to generate real-life data as generally largely unbiased in ESM and which delivers a nested data structure to evaluate situational and person-related factors by using *multilevel models*. Secondly, they investigated the mediating role of functions of music listening on situational and person-related predictors of music-selection behavior. In this course, replication of earlier research was intended. The questionnaire therefore remained very similar with the inclusion of specific questions to the current listening situation (e.g. artist and name of the song, playback volume, liking of the song, familiarity, vocal vs. instrumental). Also the statistical learning procedure, the *percentile-Lasso* was used for variable selection. Intra-class correlation measurements again showed that the music-selection behavior was mainly influenced by situational factors (accounted in average for 84 % of the variance for all musical characteristics), while functions of music listening were almost equally explained by person-related and situational predictors (44 % versus 56 % explained variance). In accordance with earlier research, music-selection was mostly explained by the function of music listening, the initial *arousal* state and degree of attention paid to the music in the current situation. Only a few person-related variables remained in some prediction models for the different music characteristics, which was again not the case for personality traits or musical sophistication scores. In addition, the mediation analysis revealed aspects that had not been discovered before. Certain functionalities worked as mediators for situational and person-related predictors. This also revealed indirect associations of variables through functions of listening that did not have any direct effect on the music-selection. Considering the different measurement techniques (ESM vs. retrospective reporting), the results Greb et al. (2018) and (2019) were still mostly consistent. Most differences were explained by Greb and colleagues due to response biases in retrospective reports. For example in the case of simple and complex music people would generally tend to report listening to more complex music to appear to be more intellectual. Also musical taste was found to be less important in the latest research. Furthermore, it stands out that mediation analysis should be put emphasize on in future studies. As Greb et al. interpreted the results for *happy* versus *sad* music, the direct positive effect of an initial positive mood (*valence*) would be interpreted as a mood-congruent selection, similar to Randall and Rickard (2017). The indirect negative mediation of *valence* through the function of *intellectual stimulation* that was found indicates a mood-incongruent selection behavior. Those aspects clearly underline the complex interactions involved in music-selection processes.

As it was laid out in this section, the music-selection behavior in everyday life is highly dependent on the situational circumstances and additionally on person-related differences. It is argued that music listening fulfills certain functions, such as emotion and arousal regulation, and it was shown that music-selection is prominently driven by the pursued function of listening. What is already clear to this point is that mostly situational variables are able to predict music-selection behavior, which is contrary to the effort that was put into research of those aspects. Furthermore, music-selection behavior to this date has only been assessed by subjective measurement procedures. However, now it is adequate to move prediction tasks from subjective music ratings towards objective characteristic predictions. Those also would be more handy to implement in commonly used applications (e.g. music recommendation) and also would reduce rating biases of participants and therefore produce more reliable results.

## 1.2 Musical Content – Subjective and Objective Characteristics

The research area concerned with retrieving the aforementioned objective musical characteristics as data units is the *Music Information Retrieval* (MIR) domain. It can generally be assumed that responses to music partially mirror perceived musical content (Husain, Thompson, & Schellenberg, 2002). Hence assuming the functional use of music, those characteristics take part in the daily music-selection behavior. Music comprises a wide range of properties which can primarily be divided into subjective and objective characteristics. They can be derived from the audio signal itself by signal processing techniques or exist in the form of *metadata. Metadata* refers to external information given by intentional tags that cannot directly be concluded from the signal, such as artist name or the year of publishing. Audio signal content is described by various taxonomies covering the multiple facets that are integrated in MIR research (Downie, 2003; Yang & Chen, 2011; Peeters, 2004; Manjunath, Salembier, & Sikora, 2002).

Those are partly congruent but differ in some of their categories and perspectives. So far, there is not a unique taxonomy available for MIR which is most likely due to the fact that it is a huge and relatively young research area (Schedl, Gómez, & Urbano, 2014). But in all aspects is seemingly implied, that properties of music can have different levels of complexity and/or abstractness, either computational or interpretational. Thus, it makes sense to organize those properties by means of their complexity levels. On one hand there are the very objective low-level *audio features* (e.g. zero-crossings of the audio signal), in Gouyon et al. (2008) referred to as "signal-centered descriptors" (p.88). On the other hand there is a very subjective high-level human perception (e.g. perceived positive emotions of a song), "user-centered descriptors" (Gouyon, et al., 2008, p.88), and in between yawns the so-called *semantic gap* (Yang & Chen, 2011). Closing this gap remains one of the most pursued endeavors in the MIR research society. Therefore, mid-level *descriptors* are increasingly developed to combine objective signal processing techniques of low-level *features* with semantic information by implementing musical and perceptual knowledge in the feature construction process (Dittmar, Bastuck, & Gruhne, 2007). Considering the different levels of complexity, this work proposes to distinguish musical properties semantically by means of their measurement procedure. Measures of the signal properties are called *features*. Those can be computed by clear mathematical operations or transformations. *Audio features* contribute to a corresponding subjective higher order perception, determined by a *descriptor*. Different *features* might contribute to the same *descriptor*. *Features* and *descriptors* can vary in their semantic complexity. It is assumed that complexity rises to a certain extent with higher subjectivity.

A second dimension for characterizing musical characteristics refers to the temporal scope that is depicted by a *descriptor* or *feature* (Schedl, Gómez, & Urbano, 2014). At each complexity level, characteristics can be measured by overall/global qualities, such as the overall mood of a song or the overall averaged loudness. In contrast, properties can be calculated periodically based on a frame rate or sample rate which results in a discrete stream of values and allows for other statistical and interpretational models, such as analyzing the change of arousal from the beginning of a song to its end, or the perceived loudness of different parts of a song. Instantaneous properties can add additional complexity, computational wise as well as from its level of abstraction. For instances, the globally measured mood of a song is interpretational wise very different from a time dependent emotion measure throughout the song. The same applies for subjective descriptions of music properties. Having participants annotate emotional values on *valence* and *arousal* scales results in very different interpretations for global and continuous ratings.

Additionally, low-level features are often categorized by their perceptual dimensions into energy, rhythm, temporal, spectrum and melody (Yang & Chen, 2011). It is out of the scope of this work to provide a complete overview of low- to high-level characteristics of musical content. In general, low-level features become very useful in big sets of features contributing to the same prediction/classification task of subjective high-level or mid-level *descriptors*. It is noted that music perception is mostly determined by a combination of several low-level features. Therefore, statistical learning, classification or regression approaches are usually used on big sets of processed features to find a perceptual correlate. Significant importance in MIR research was achieved by features like the root-mean-squared energy (RMS level) of the signal, onset-rate of events, zero-crossings, spectral features like spectral centroid, roll-off or flux that are derived from the frequency transformed signal by STFT (Short Term Fourier Transformation) or FFT (Fast Fourier Transformation) and MFCC (Mel-Frequency Cepstral Coefficients). Perceptual correlates were found for perceived *valence* and *arousal* by energy and tempo measures of songs (Gabrielsson & Lindström, Gabrielsson, Lindström 2001 – The Influence of Musical Structure, 2001). MFCCs are often used in timbre related tasks, such as speech recognition or instrument identification and also genre recognition (Humphrey, Bello, & LeCun, 2012) just to mention a small selection of features and their application in MIR to fill the *semantic gap* by predicting high-level music perception.

## 1.3 Prediction of Music Perception

Music and audio perception imply many different individual tasks that can also differ substantially in their complexity levels. In principle, music perception is subjective to the individual but most tasks underlie certain physiological and psychological factors (e.g. perception of loudness) and are additionally influenced by acquired knowledge of the listener (e.g. in the case of musicians) and situational cues. The research area of music perception includes facets as pitch perception, instrument classification, genre recognition but also musical emotion recognition, music quality perception as well as music similarity perception. As described earlier, basic low-level audio features are often used in MIR to model and predict high-level music perception but conventional methods have often reached their maximal performance already years ago. As pointed out by Humphrey and colleagues (2012), content-based MIR has mainly been making use of hand-crafted audio features built on specialists domain-knowledge and applied relatively simple processing procedures. They argue for advanced feature design and deep learning architectures to further improve music perception tasks. But methods for content-based predictions of music perception should remain less important in the context of this work for now. Particular focus should be raised on the fact that music perception is not exclusively determined by objective musical characteristics. According to Schedl, Flexer and Urbano (2013), understanding music perception rather requires the inclusion of a second approaches in MIR research, besides a *systems-based* also a *user-centric* approach. This implies the merging of the mostly separated research areas MIR and music psychology by means of current statistical learning procedures. As it is depicted in Figure 1, music perception and recognition are not only functions of musical content descriptors but also musical context in form of metadata impacts the individual perception process, for example as it was shown in the case of music similarity perception (Schedl & Knees, 2009). What has seemingly been ignored in most past MIR research is the role of the listener in the process of perception. Although it is known already that the individual music-selection behavior is determined by contextual and person-related factors (Greb, Steffens, & Schlotz, 2018), little is known about the impact of context on the individual perception. *User context* refers to all situational or dynamic aspects that are active in the course of music listening (e.g. mood, activities, time) and *user properties* are static properties of the listener such as music preferences or personality traits. Figure 1 implies that the user has mostly been object to music psychological research, whereas MIR concentrated on the system only. What is needed is a merged modeling approach of music recognition and perception tasks.

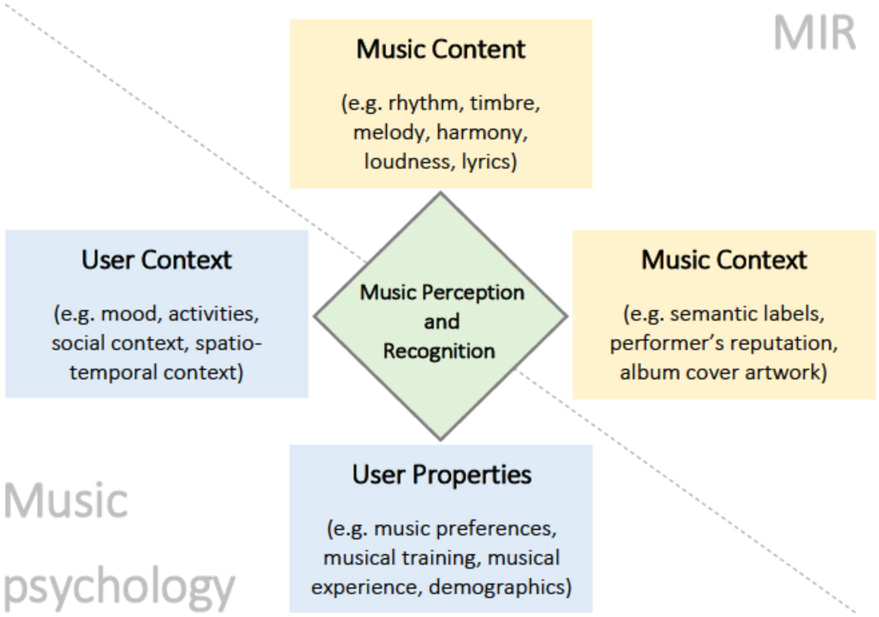
Music Content
(e.g. rhythm, timbre, melody, harmony, loudness, lyrics)
MIR
User Context
(e.g. mood, activities, social context, spatio-temporal context)
Music Perception and Recognition
Music Context
(e.g. semantic labels, performer's reputation, album cover artwork)
User Properties
(e.g. music preferences, musical training, musical experience, demographics)
Music psychology

*Figure 1 - Emerging factors contributing to the music perception. Musical as well as user characteristics impact on music perception and recognition. (Figure was taken from Schedl et al. (2013) and slightly modified.)*

The annual evaluation campaign for MIR algorithms *MIREX* (Music Information Retrieval Evaluation eXchange) has established benchmarks for several prediction tasks of musical perception (Downie, 2006). As reviewed by Weigl and Guastavino (2011), research strategies in the context of one of the most important MIR conferences *ISMIR* (International Symposium on Music Information Retrieval) have mostly been focusing on *system-based* evaluation frameworks. Besides all music recognition processes, this seems to be coming especially very short in the context of emotion recognition tasks because it has been widely acknowledged that musical emotions are additionally recognized very differently by means of the listener context and properties. Subsequently, several aspects of emotions in regards to music are considered.

## 1.4 Music and Emotion

Because this work aims at predicting several objective musical characteristics and especially includes emotional representations, the complex interactions between music and emotions shall be expounded in the following. Despite many high-level recognition tasks of musical properties, the understanding and prediction of musical emotions, *perceived* as well as *felt* emotions, seem to be very appealing, considering the amount of research that was done during the last decades. All those very promising results are supporting the researchers with a lot of good reasons to proceed. At this point, this work aims on giving an introduction in general emotion theory. After that, special phenomena of musical emotions are depicted before state-of-the-art emotion recognition techniques are presented.

Many theories try to explain the foundation and development of emotions in mankind. Nothing can be stated with complete confidence, but certain theories seem to be convincing and have been the bases for many ways of argumentations and investigations in recent research. In the context of music listening, often stated are evolutionary approaches considering emotions to be an integral component of the survival instinct of hominin ancestors that provided the individual with fast responding underlying mechanism to produce a fast and reliable reaction on certain events (Juslin, 2013). This can be closely related to evolutionary theories about the origin of music, in the sense of music being an emotion expression tool in very early prelinguistic development stages, as it was mentioned before (Mithen, 2005). The complexity of emotions has certainly overwhelmed researchers ever since but supported the creation of various emotion models, interested in dimension reduction and finding the principal components of emotions. The two most common approaches are the *categorical* and the *dimensional* emotion representations. Affective terms are used to denote the corresponding emotion in *categorical* approaches. Despite the fact, that people use those affective terms in everyday situations and therefore the concept is easily understood, difficulties in research scenarios lie in selecting the right terms and the necessary number of distinct terms to describe the emotional space, which is referred by Yang and Chen (2011) as the *ambiguity* and *granularity* issue of *categorical* emotion taxonomies. Paul Ekman has arguably been one of the most influential emotion researchers by proposing the six *basic emotions* (anger, disgust, fear, happiness, sadness and surprise). Many researchers, also in music sciences, have used the basic emotions to avoid *ambiguity* but never overcame the *granularity* issue (Yang & Chen, 2011). However, dimensional approaches do not have the dilemma with finite emotion classes. Therefore, *ambiguity* and *granularity* issues are less prominent. Most commonly used is the two-dimensional space of emotions in the *valence-arousal*-plane (positive/negative - exciting/calming) (Russell, 1980). Notably, Ekman discarded his prior dimensional approach that described emotions on a *pleasant-unpleasant* scale after finding universal cues of emotion expressions in the 1970s (Ekman, 1992), which was not very far off Russel's dimensions and other dimensional approaches. It is possible to approximate positions of different affective terms on a circumplex in the *valence-arousal* (VA) plane as shown in Figure 2. This makes the VA-approach quite comprehensible for users.

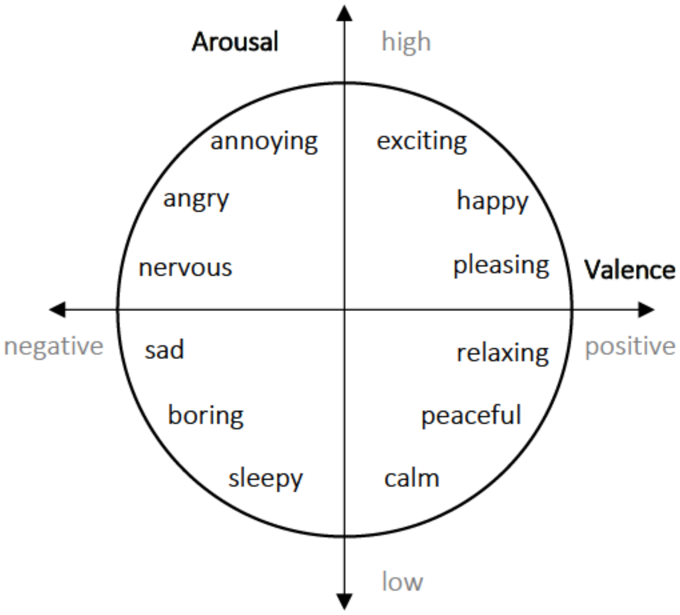

*Figure 2 - The circumplex model with approximated positions of affective terms in the 2D valence-arousal plane c.f. Russel (1980).*

Nevertheless, there are considerable discussions about the nature of *dimensional* approaches. In studies of other researchers, not only the terms of the dimensions but also the number of only two dimensions are vexed. For example Fontaine et al. (2007) identified four dimensions on the basis of 144 emotion descriptions from different contexts (1. Evaluation-pleasantness, 2. Potency-control, 3. Activation-arousal, 4. Unpredictability). Schimmack and Reisenzein (2002) tested the two-dimensional activation approach from Thayer (1990). This theory claims that *energetic arousal* (awake vs. tired) and *tense arousal* (tense vs. calm) are two independent types of activation and both include *valence* components pointing to opposite directions. One could interpret *awake* as a strong positive *arousal* activation and *tense* as a strong negative *arousal* activation. The so called "valence activation hypothesis" (Schimmack & Reisenzein, 2002, p. 413) was tested by measuring *energetic arousal*, *tense arousal* and *valence* independently to be able to remove the shared variance components of *valence* in both activation dimensions. It was concluded that *energetic arousal* and *tense arousal* are independent dimensions, sharing mostly variance due to the *valence* component. Eerola et al. (2009) carried out a mapping of basic emotions in soundtracks (happiness, sadness, tenderness, anger and fear) onto a three dimensional space containing *valence*, *activity* and *tension*. This was based on the findings of Schimmack and Reisenzein (2002) and considered the notion of Bigand and colleagues (2005) that especially *musical emotions* are not mirrored completely in just two-dimensional models. Eerola and colleagues (2009) had experts rating 360 soundtrack excerpts using the *basic emotions* categories and the dimensional representation. By mapping the dimensional ratings onto the corresponding *basic emotions*, Eerola and colleagues realized that the three-dimensional model could not significantly explain more variance of most of the affect terms, except of anger, than the two-dimensional VA-model. To sum up, *valence* and *arousal* are in certain ways always included in dimensional approaches and seem to be two principal dimensions to describe emotions, which should not generally exclude other possible dimensions.

The distinction of *everyday emotions* and *musical emotions* has already been part of many considerations. Often, music listening is assumed to create an *aesthetical context*. Since emotions are highly situational and subjective, music listening also owns its proper emotional response pattern, so-called *aesthetic emotions* (Juslin, 2013). It is argued, that the *basic emotions*, by having an evolutionary origin, do not explain emotional responses to music to their full extend (Zentner, Grandjean, & Scherer, 2008). Therefore, a theory of Juslin (2013) combines evolutionary emotions and *aesthetic emotions* to a unified concept of *musical emotions*. This theory also attempts to explain the existence of *mixed emotions* as a result of different interacting mechanisms reacting on the same event. A good example for *mixed emotions* is the existence of a sort of positive sadness like *sweet sorrow*. While sadness generally is considered to be a rather negative emotion, *sweet sorrow* also carries positive aspects within. In 2008, Zentner, Grandjean and Scherer presented a 9-factor model (Geneva Emotional Music Scale, GEMS) for *induced musical emotions*. Part of their studies was investigating the often discussed divergence between *perceived* and *felt* emotions in the course of music listening (Gabrielsson, 2001) and drawing the line between *everyday emotions* and *musical emotions*. Results showed that the domain-specific model was able to measure emotions induced by music better than the *basic emotion* model and dimensional emotion models. The outstanding finding in their study however, is the discrepancy of negative emotions being *perceived* as opposed to those being *felt* by music listening. The scale also shows proportionally few negative affect terms compared to positive ones. This might indicate that music usually induces more positive associated emotions, which could explain the emotional value of music in terms of mood enhancement and regulation. This view also is in line with a study of Kawakami et al. (2013). They presented results that indicate that *perceived* sad music makes people feel more romantic, more blithe and less tragic than the music would actually imply. However, other studies also supported the contrary point of view that negative *musical emotions* (grief, melancholia and sweet sorrow) indeed are often induced by sad music and contextual aspects contribute to their specific form (Peltola & Eerola, 2016). In this context, personality traits are often considered to have an individual impact on the appreciation of sad music. For example *trait empathy* and *openness to experience* were encountered to support preferences for sad music (Vuoskoski & Thompson, 2012; Garrido & Schubert, 2011; 2013). All these aspects that were now examined on negative musical emotions but of course exist in a certain way for all *musical emotions*, like complex *mixed emotions*, discrepancy of *perceived* and *felt* emotions and paradox interpretations of affect terms, lead to the conclusion that the emotional perception of music is highly situational and person dependent and emphasizes the need for more ecologically validated research about affective music use and consequently selection behavior in everyday life situations.

Using one or another emotion representation has implications on the actual emotion recognition prediction task. In categorical approaches this is referred to as a *classification* task (e.g. positive versus negative emotion). Dimensional approaches imply a *regression* prediction of the outcome variables on each dimension (e.g. value of valence on a continuous scale). There are several comprehensive state-of-the-art reviews that summarize the efforts of the last decades (Kim, et al., 2010; Yang & Chen, 2012). It is out of the scope of this work to give a complete overview of the research area but some main findings shall be shared now. Generally researchers would collect *ground truth* data, which means collecting subjective ratings of emotional qualities of music (categorical or parametrical). Collecting this kind of data remains a big challenge but a variety of solutions have been found to overcome this issue, such as listener surveys, using social tags (e.g. from Last.fm) or data collection games (Kim, et al., 2010). Despite the fact that the distinction between *perceived* and *felt* emotions is well noted by MIR researchers, emotion recognition tasks have mainly focused on *perceived* emotions, which are argued to be less prone to contextual factors (Yang & Chen, 2012). This obviously is a critical point with regard to investigating the functional use of music as an emotion regulator, which would benefit from information about the actually *induced* emotions, too. Reading through several MIR related works, it quickly becomes obvious that *mood* and *emotion* are mostly treated synonymously (Yang & Chen, 2012), in contrast to common psychological taxonomies that distinguish the different affect terms precisely (Juslin, 2013). As it was already pointed out, different MIR approaches to music emotion recognition might also have very different interpretations. This leads to the fundamental distinction of music emotion recognition (MER, c.f. Prediction of Music Perception) systems in static ones and dynamic ones. Static approaches would apply a single-label (affect term, or parameter value) accounting for the emotional content of the whole song, whereas in dynamic MER systems continuous emotion values are set on a frame rate or time rate. The latter approach also is considered as *music emotion variation detection* (Aljanaki, Yang, & Soleymani, 2017). Aljanaki and colleagues started to create a benchmark framework including a dataset with dynamic emotion annotations. It is understood that pieces of music vary in their emotional content changing with time which gives reason to support the dynamic approaches. But it remains particularly difficult to obtain a large and meaningful dataset of continuous emotion annotations. Categorical approaches are even more difficult to assess dynamically. The most obvious approach would be to create continuous annotations for example on a *valence-arousal* scale. Additionally to ratings or annotations of emotional properties via surveys, emotion tagging etc., contextual information sometimes is used to support analyses. These information include artist biography, album reviews, but also social tags (Kim, et al., 2010). Strategies collecting those music-context information are referred to the area of *web mining*. Especially social tags are gaining increased attention due to their high ecological validity and accessibility on platforms like *Last.fm*. Still mostly used for music recognition tasks is the processing of musical content properties, so-called audio features. In accordance with earlier sections (c.f. Musical Content – Subjective and Objective Characteristics, Prediction of Music Perception), MER seeks to predict musical emotions by low-level audio features. Being able to assess musical moods computationally is from great economic importance for many current services since the world wide music content is increasing drastically and content-based approaches are quite scalable. MIREX, already introduced earlier, is a benchmark platform for MIR researchers and is also involved in content-based MER classification tasks. Classifications are made on five different clusters of mood adjectives (Hu, Downie, Laurier, Bay, & Ehmann, 2008) as displayed in Table 1.

*Table 1 - Mood clusters of MIREX classification task and corresponding mood adjectives (Hu, Downie, Laurier, Bay, & Ehmann, 2008).*

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Rowdy<br>Rousing Confident<br>Boisterous<br>Passionate | Amiable/Good natured<br>Sweet<br>Fun<br>Rollicking<br>Cheerful | Literate<br>Wistful<br>Bittersweet<br>Autumnal<br>Brooding<br>Poignant | Witty<br>Humorous<br>Whimsical<br>Wry<br>Campy<br>Quirky<br>Silly | Volatile<br>Fiery<br>Visceral<br>Aggressive<br>Tense/anxious<br>Intense |

Typical classification algorithms are *Support Vector Machines* (SVM), *Gaussian Mixture Models* (GMM), *Neural Networks* (NN) and *k-Nearest Neighbors* (*k*-NN), just to mention some (Yang & Chen, 2012). Generally, performances of content-based classification tasks remained relatively low in the past. As Humphrey (2015) mentioned it is a general trend in content-based music recognition tasks, such as chord estimation or genre recognition, that performances are reaching their limits, which has never been very high for music emotion recognition with best classification rates of around 70% in the *MIREX Music Mood Classification Task*. However, it has been argued recently that classification tasks are usually outperformed by *regression* tasks, which is not due to different feature selection. Also methods like *Support Vector Regression* (SVR) or *Gaussian Mixture Model Regressions* are very similar approaches to the classification approaches. The *granularity* and *ambiguity* issue seems to hold up for this tendency. However, typical audio features are summarized in Table 2. Those features belong to different complexity levels, very basic low-level features (e.g. RMS energy, spectral shape) but also higher level features incorporating musical and psychoacoustical knowledge (e.g. rhythm, MFCC's). Again this collection uses a slightly different taxonomy for the different categories of audio features in contrast to earlier descriptions. It should be considered to be a small fraction of audio features that are used.

*Table 2 - Collection of typical audio features for content-based music emotion recognition (Kim et al., 2010).*

| Dynamic | Timbre | Harmony | Register | Rhythm | Articulation |
|---|---|---|---|---|---|
| RMS energy | MFCC's, Spectral shape, Spectral contrast | Roughness, Harmonic change, Key clarity, Majorness | Chromagram, Chroma centroid, Chroma deviation | Rhythm strength, Regularity, Tempo, Beat histograms | Event density, Attack slope, Attack time |

With high advances in the signal processing and machine learning community another musical property has become very valuable for MER systems, namely song lyrics. Lyrics exist somewhat in between the fields of content and context and are still generally assumed to be highly complex, in both steps generating *ground truth* and *feature design*. As many comparisons showed, using multi-modal approaches with audio content and lyrics has improved most emotion recognition tasks (Kim, et al., 2010). A very recent research has shown that in the case of *valence*, a combined approach outperformed single content or lyric prediction tasks with a middle fusion model of a *Convolutional Neural Network* (content input) and a SVM (lyrical input) (Delbouys, Hennequin, Piccoli, Royo-Letelier, & Moussallam, 2018). That was not the case for *arousal*, which was not improved by combined modal approaches compared to a content-only *Convolutional Neural Network*.

In summary, trends indicate that multi-modal prediction techniques, combining lyrics, content, metadata, tags etc. can still improve system-based recognition tasks, also machine learning advances in deep learning and increased efforts for dynamic MER systems support this direction. As this overview showed, it is still poorly recognized to implement the user in such systems. Thus, MER systems continue to lack individuality, interpretability and real-life situation accuracy. The *user properties* and *user context* need to be considered in the future MIR approaches because it is widely acknowledged that emotion perception and induction by music is highly situational (North, Hargreaves, & Hargreaves, 2004; Juslin, Liljeström, Västfjäll, Barradas, & Silva, 2008; Sloboda, O'Neill, & Ivaldi, 2001), and person-related (Vuoskoski & Thompson, 2012; Garrido & Schubert, 2011). *User-centric* MER approaches would definitely support contextual music recommendation approaches. Since the frequent emotional use of music is highly evident, such approaches would be very valuable to model music listening and selection behavior. As the next section shows, music recommendation systems incorporating contextual information have also not reached wide attention so far, and are almost not existent in commercial solutions, despite the contextual importance for the music listening behavior.

## 1.5 Music Recommendation

Since the everyday listening behavior is changing by means of technological trends (North, Hargreaves, & Hargreaves, 2004) and people are confronted with a tremendous amount of music, it makes sense to provide listeners with music recommendations on their playback devices for easy access and practical use. The relevance of recommendation systems (RS) is increasing dramatically. Streaming services for music and audio content became global players in the last decades. Most prominently named by *Spotify*[2], *Pandora*[3] or *Last.fm*[4] just to mention three out of a wide range of music streaming applications that would also recommend music from their large music data bases. Music listening is not determined by the choice of the radio station, concerts and discos or the limited amount of physical media one holds at hand like CD's or vinyl anymore. On the other hand it has almost become impossible to handle the amount of possibilities a user gets. People suffer from constant information overload. A system that provides an accurate music recommendation to fulfill the listeners needs, however could overcome some aspects of modern technological stress factors.

Music recommendation systems (MRS) are approached in very different ways nowadays but *collaborative filtering* (CF) and *content-based filtering* (CBF) are still the most commercially common techniques. Those traditional methods would make use of *user ratings* towards specific *items* and recommend other, unseen items according to the users' preferences. Thus, the recommendation results from previous behavior and is based on a rating function that describes a two-dimensional matrix: *Users x Items - - > Ratings* (Ricci, Rokach, Shapira, & Kantor, 2011). CF refers to the user in comparison to its social environment. Ratings or preferences are compared and they provide a prediction or recommendation of items that are unknown to the user but known in the "neighborhood" by similar users and possibly turn out to be favorites (Kaur & Kumari, 2017), whereas CBF considers specific properties of items that were rated positively or were actively chosen by the user. Therefore, a content-based recommender provides items with similar characteristics (Aucouturier & Pachet, 2002). In music recommendation, MIR would provide this musical content, namely audio features like tempo, mode, timbre, rhythm etc. (Li, Kim, Guan, & Oh, 2004). Both approaches, CF and CBF, have complementary properties with obvious drawbacks, which makes it highly recommended to use *hybrid* versions

[2] www.spotify.com
[3] www.pandora.com
[4] www.lastfm.com

combining different recommendation approaches (Yoshii, Goto, Komatani, Ogata, & Okuno, 2008). CF recommender suffer from *cold-start* problems, which means, new items that no one ever rated before have little chances to be encountered. Another problem is the *popularity bias*, meaning popular items constantly remain being likely to be recommended. Both effects result in a *sparsity* problem, meaning only a fraction of possible items do have a rating and therefore can possibly be recommended by *collaborative filtering* approaches. Many researches have tried to overcome those problems, e.g. (Chen, Wu, Xie, & Guo, 2011). However, CF turned out to be quite effective (Yoshii, Goto, Komatani, Ogata, & Okuno, 2008) and CBF can overcome *cold-start* problems and *popularity bias* but instead suffers from accuracy problems since *similarity* of song properties has not been proven to correspond with musical preferences and is contrary to other quality metrics of recommendation systems, such as *novelty* and *diversity* (Song, Dixon, & Pearce, 2012; Ricci, Rokach, Shapira, & Kantor, 2011). Thus, *hybrid* approaches that combine different methods usually outperform the single conventional approaches (Song, Dixon, & Pearce, 2012; Kaur & Kumari, 2017). More recently, other techniques were also considered in different recommendation tasks. Those seem to make sense especially in the context of music recommendation. *Emotion-based recommendation* (EBR) is similar to CBF by using calculations of emotional content of music for recommendation (e.g. *valence*/arousal values) which is applied by the online recommendation services *Musicovery*[5] as well as *Sourcetone*[6]. This refers to the widely investigated emotional value and functionality of music and is based on the progress in computational emotion recognition (Schäfer, Sedlmeier, Städtler, & Huron, 2013; Greb, Schlotz, & Steffens, 2017; Yang & Chen, 2011). Another, very promising approach is the use of *context-aware recommendation systems* (CARS). Context can be distinguished into *user-context* and *music-context*. *Music-context* refers to all information relevant to the music that is not directly available in the data, such as political background of the artist, specific meanings of songs, etc. (Schedl, Flexer, & Urbano, 2013). The *user-context* was categorized by Göker and Myrhaug (2002) in five categories: *environment context, personal context, task context, social context* and *spatio-temporal context*. In music listening, the *task context* refers to the main activity during and even functionality of music listening, which are considered to be integral components in the process of music-selection (Greb, Steffens, & Schlotz, 2018) and consequently must be contemplated in a MRS. The *personal context* is also highly related to the functionality of music listening and music-selection and considers the physiological and mental context of the individual. Knowing about the physiological and psychological state of the user can for example be useful for mood and *arousal* regulation purposes, since research found associations supporting mood-congruent music-selection and *arousal* regulation functions of music listening (Randall & Rickard, 2017). CARS scan result in a *hierarchical* recommendation data structure, where recommendations are based on groups of

[5] http://b2b.musicovery.com/

[6] http://www.sourcetone.com/

contexts and can be represented by trees (Ricci, Rokach, Shapira, & Kantor, 2011). Another, often parallel implemented approach is *demographic based recommendation* helping by initial groupings in *hybrid* approaches by age, gender, nationality etc. (Liu, Lai, Chen, & Hsieh, 2009).

Contextual recommendation systems for music have by far not been as popular as conventional approaches. But considering the results of Greb and colleagues (2017; 2018; 2019), CARS for music would improve several key quality factors of recommendation systems, such as *accuracy*, *coverage*, *confidence*, *serendipity*, *diversity*, *utility* and *adaptivity*, as defined in Ricci et al. (2011). While the commercial successful recommendation systems use *collaborative filtering* (*Last.fm*) and *content-based* models (*Pandora*), there are already some academic context-based recommenders. It can be said that music recommendation research on context-effects had provided many useful insights on the contextual use of music, but certainly from a very practical point of thinking. Su and colleagues (2010) integrated different physiological dimensions (heartrate, body temperature), weather information (air temperature, humidity) and additional contextual information (time, location, motion etc.) into one context-aware music recommender, called *uMender.* By conducting an experiment using semi-real data and a prototypical real-life application, they achieved results indicating that the inclusion of contextual parameters outperformed *user-based* and *item-based* recommendation approaches. This is underlining the real-life significance of contextual factors in the music recommendation processes, thus also supporting the relevance in music-selection behavior. Wang et al. (2012) collected contextual data about daily activities (walking, relaxing, running, sleeping, shopping) with smart-phones and implemented a probabilistic model to integrate contextual-data with musical content for recommendation. Dias et al. (2014) presented the *hybrid* recommender *Improvise* that associates musical content with the same daily activities from Wang and colleagues using an initial generic system to overcome the *cold-start* problem and improving accuracy by user feedback. Other contextual recommenders incorporated parameters like time of day, seasons, locations, urban environment information (traffic), weather, temperature and motions (Reddy & Mascia, 2006; Lee & Lee, 2007; Liu, Lai, Chen, & Hsieh, 2009). Baltrunas and Amatriain (2009) introduced the *micro-profiling* technique to implement contextual parameters. Each user receives several sub-profiles corresponding to different contextual parameters. This is an extension to CF or *user-based* approaches and acts as a context filter. Baltrunas et al. (2011) implemented a context-aware music recommender for driving situations. Different context variables (e.g. landscape, sleepiness, mood, traffic conditions) were shown to have an impact on the propensity for different music styles, implying the relevance of those factors.

Similar to music recommendation is playlist generation. It is usually a rather static approach, since playlists are sets of pre-selected music. Also in this community, aspects of context dependencies have been considered. For example Oliver and Kreger-Stickles (2006) introduced an automatic playlist sorting principle considering physiology and purpose of music listening while working out.

This short introduction into music recommendation systems shows, that there are academic research approaches existing also integrating contextual information of the listener. But as Zheng (2017) argues, that despite being quite effective, most contextual systems were created in the spirit of machine learning strategies, and therefore lack of interpretability and relevance assessment of the different contextual parameters. Making more use of current behavioral research that clearly indicates relevant contextual predictors of music-selection behavior would therefore beneficially support CARS for music. Thus, this work argues to support the creation of contextual music recommendation systems by identifying the statistically most relevant functional, situational and person-related predictors of audio features of self-selected music supported by state-of-the-art statistical learning procedures as introduced further in the next section. However, implementing a context-aware music recommendation system can reversely be object to a real-life relevance and significance assessment of theoretical and statistical models about music listening and selection behavior.

## 1.6 Statistical Learning of Selected Musical Properties

Combining the fields of psychology and machine learning inevitably evokes the controversial discussion about choosing *explanation* or *prediction* as main objective in the statistical analysis. As earlier chapters showed, music psychology has rather intuitively chosen explanatory approaches. Complex theories were supported by findings of statistical relationships giving insights into the present data but usually lack to explain unseen data or predict potential future behavior. This is stated in the well-known *replication crisis* of psychological research (Open Science Collaboration, 2015). Despite the fact that results of explanatory research often were not able to be held up in similar assessments, they were generalized to explain human behavior. Machine learning contrary to that, usually focuses on *prediction* tasks. The goal is predicting the outcome of unseen input variables. Potentially, every measure of input data underlies certain measurement errors introducing *noise*. This notion is generally acknowledged, but models that accurately fit sample-specific noise components ignore this fact in the spirit of evaluating the model's performance by *the goodness-of-fit*. This problem is known as *overfitting*. But also potentially noise-free data can be *overfitted* by applying fits of very high order. Those models would not generalize well on unseen data. By reducing the overfit, the prediction accuracy of present data decreases but increases drastically for *out-of-sample* data. The outcome allows a more reliable generalization, which has actually always been the main objective of psychological research. In fact, prediction tasks would contribute to the understanding and explanation of psychological phenomena on the long term and therefore complement explanation approaches (Yarkoni & Westfall, 2017).

To prevent models from *overfitting* and enable reliable generalized predictions of unseen data, various parameters that can affect the amount of *overfitting* need to be considered: one, a weak effect size of predictors, and two, a low ratio of sample size to predictors support *overfitting* (Yarkoni & Westfall, 2017). Also statistical procedures can produce *overfitting* by intervening in the model building process to reach a better fit on the present data, known as *p-hacking* (Simmons, Nelson, & Simonsohn, 2011). In general, errors that contribute to the overall predictive error can be decomposed into the components of *bias* and *variance*. *Bias* refers to systematic tendencies of over- or underestimation and *variance* explains the fluctuation of a model's parameters (Harlow & Oswald, 2016). The attempt to minimize the overall error requires the adjustment of both types. Unfortunately, *bias* and *variance* depend on each other. Thus, a certain compromise, so-called the *bias-variance-tradeoff*, is pursued in the model building process to reduce the overall error. It can be generally assumed that too simple models are strongly *biased* but have little *variance*, therefore those models underfit the data, whereas models showing high *variance* and low *bias* tend to be too complex, which leads to *overfitting* (Chapman, Weiss, & Duberstein, 2016). Figure 3 shows that *bias* and *variance* errors could be balanced by setting up an ideal model complexity to minimize the total prediction error on unseen data. The optimal test sample prediction error is obviously not achieved by optimizing the training sample fit error.

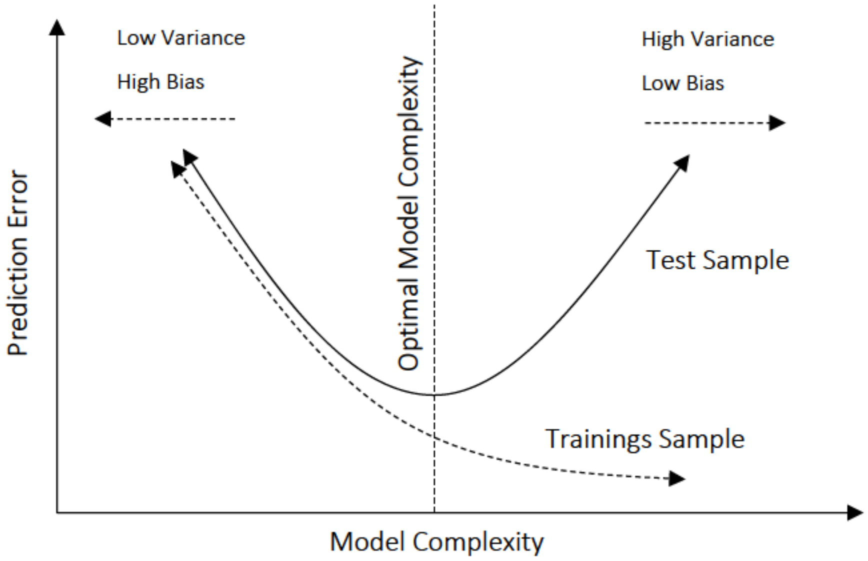

*Figure 3 - Bias-Variance Tradeoff. Finding the optimal model complexity leads to an overall minimized predictive error for unseen data. Variance and Bias components contribute both to the total error. By minimizing the trainings sample error the test sample error would suffer. (Figure was taken from (Hastie, Tibshirani, & Friedman, 2017) and slightly modified.)*

The machine learning community has presented some very effective methods to overcome *under- and overfitting* problems. Three main aspects can help to control the total error in statistical learning procedures, a) using an adequate sample size (*n*) (aims on increasing the *n* to *p* ratio by increasing *n*), b) constraining or regularizing the complexity of models (aims on increasing the *n* to *p* ratio by decreasing *p* or decreasing the effect of *p*), and c) assessing trained models by their predictive performance on unseen data, for example by *cross-validation* the most common technique (supports the best model selection to prevent overfitting) (Yarkoni & Westfall, 2017).

*Big data*, as a buzzword, refers to point a) and has been one of the hot topics in recent years for data scientists but has surprisingly gotten little attention in psychological research. By giving an “Introduction to the Special Issue” (Harlow & Oswald, 2016, p. 1) only in 2016, it is obviously implied that *big data* is still stuck in its infancy regarding psychological applications, despite the fact that a reasonable amount of data would substantially improve the model building process and validation of theories. Of course, acquiring data is highly cost intensive, which is a major drawback, but it is required to produce models that are generalizable and do not suffer from *overfitting*. Also, modern technologies allow for the acquisition of data better than ever by using smart-phone apps, web applications or tools like *Amazon Mechanical Turk*[7] to hire *human intelligence* and measure human behavior. The assessment of sufficient data is a necessity in modern psychological research.

> *Big data science can be instrumental in collaboratively working to uncover and illuminate cogent and robust patterns in psychological data that directly or indirectly involve human behavior, cognition, and affect over time and within sociocultural systems.*
>
> Harlow & Oswald (2016, p. 448)

Reducing the complexity of models is still a more cost efficient way to improve the prediction accuracy for unseen data and should be applied besides collecting much data. Complexity can be handled by means of the number of contributing variables, referred to as *regularization.* The *regularization* of model parameters that have very limited predictive value reduces the *variance* component of the total error but introduces a more *biased* prediction. By finding the best compromise, it results in a better fit for *out-of-sample* data, the data that has not been part of the modeling process. A relatively recent developed method became a more and more popular approach to intentionally *bias* the model fit, the so-called *Lasso* (least absolute shrinkage and selection operator). Introduced by Tibshirani (1996), the

[7] https://www.mturk.com/

*Lasso* regression seems to be very suitable for the needs of psychological research and concurrently makes use of a modern machine learning technique. By controlling *dense* structures (low *n* to *p* ratio) (Yarkoni & Westfall, 2017), the linear *Lasso* regression still provides the, by psychologists so much pursued, interpretability of the model. In general, linear regression predicts the outcome variable $Y$ by $p$ input variables given as a transposed vector $X^T = (X_1, X_2, \dots, X_p)$ by solving the equation (1):

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \, \hat{\beta}_j \tag{1}$$

$\hat{\beta}_0$ is called the *intercept* and is the estimated value where the regression line crosses the ordinate. $\hat{\beta}_j$ is the estimated coefficient of the independent variable $X_j$. The equation can simultaneously be written in vector form by including a constant variable $1$ in $X$ and $\hat{\beta}_0$ in the coefficient vector $\hat{\beta}$:

$$\hat{Y} = X^T \hat{\beta} \tag{2}$$

To fit the linear model, each coefficient $\beta$ is selected by minimizing the *residual sum of square* between the estimated value and the real value of the output variable. This approach is called *least squares* and minimizes the following equation:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{3}$$

where $N$ denotes the number of observations. The *least squares* method can also be regulated by a penalty term $P$.

$$P_1 = \sum_{i=1}^{p} |\beta_j| \tag{4}$$

The *Lasso* would make use of a penalty term considering the absolute value of the coefficient $\beta$. This property supports the selection of *sparse* models by the L$_1$-norm (used in *Lasso)* and is therefore able to perform a variable selection by shrinking coefficients of meaningless variables to zero. Another *regularization* technique is the *ridge regression*. It uses another penalty term (L$_2$-norm) that shrinks coefficients towards zero but never exactly to zero.

$$P_2 = \sum_{i=1}^{p} {\beta_j}^2 \tag{ 5}$$

Both terms can also be mixed in a more generalized approach called *elastic net*.

$$P_{\alpha} = \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right] \qquad (6)$$

In the case of *Lasso*, $\alpha$ is equal to one. *Ridge regression* would mean to have an $\alpha = 0$. Finally, the penalty terms can be controlled by a tuning parameter $\lambda$. The *Lasso* coefficients $\hat{\beta}^{lasso}$ can then be estimated by minimizing $\lambda P_1(\beta)$ and $RSS(\beta)$.

$$\boldsymbol{\hat{\beta}^{lasso}} = \underset{\boldsymbol{\beta}}{\mathbf{argmin}} \ (\boldsymbol{RSS(\beta)} + \boldsymbol{\lambda P_1(\beta)}) \qquad (7)$$

Besides *regularization*, a second technique contributes crucially to the model’s success in predicting unseen data. *Cross-validation* (CV) is the attempt to estimate the generalized performance of a model and can be used to select a model that generalizes the best. Applying CV means simply assigning different observations of the dataset into a *training-set* and *validation-set* and *test-set*. Thus, the model is trained and tested repeatedly on different samples. The model’s performance can be assessed by measuring the *test error* (e.g. average mean-squared-error over all folds) and not by the *training error* itself which can be manipulated in the spirit of the *goodness-of-fit*. One of the most common CV techniques is the *k-fold cross-validation* that applies a pseudo-random assignment of the data into *k* almost equally sized folds. Each fold is iteratively being used to test the prediction error of the model that has been trained in the remaining folds. A generalized performance can be estimated by averaging over all prediction errors (e.g. mean-square-error). Effects that are by chance existing only inside the test or training set do not survive in the model selection process. CV is often used as model selection technique and for *hyperparameter* tuning applying a *grid-search*, which applies a range of values in each *cross-validation* loop to find the best parameter. For example in the case of the *Lasso*, different values for the shrinkage parameter λ could be cross-validated and in the course of model selection the best tuning parameter is chosen by the smallest prediction error over all *cross-validation* folds.

*Cross-validation* can substantially improve the modeling process in regards of making better predictions on unseen data by selecting the optimal model but does not allow to assess the performance of the model selection procedure on large sample at the same time. According to Stone (1974), CV techniques need to be separated into two different processes, one for model selection and one for model selection procedure assessment. The model selection procedure usually refers to the classifier or regression approach that was used. A modelling procedure assessment can potentially be done on a completely different dataset that had no impact on the modelling process itself. This again requires an adequate/big sample size or a second dataset. Another possibility is the implementation of a *nested cross-validation* (NCV). The NCV performs two *cross-validation* loops whereas the inner CV loop is nested in the outer CV loop. The optimal model for the corresponding prediction task is selected in the inner loop, the assessment of the selection procedure takes place in the outer loop. In the case of *Lasso*, it was pointed out in different papers (Roberts & Nowak, 2014; Krstajic, Buturovic, Leahy, & Thomas, 2014) that the selection of the optimal tuning parameter is sensitive to the fold assignment made by *cross-validation*, which requires a repeated model selection procedure. Roberts and Nowak (2014) approached the *percentile-Lasso*, which suggests the selection of a certain percentile value (e.g. 95th-percentile) out of all optimal tuning values generated in a repeated model selection process. As they showed, this procedure is more consistent in selecting optimal *hyperparameters* ($\lambda_{\text{Opt}}$). Krstajic et al. (2014) suggested a slightly different technique where the *mean loss function* of each tuning parameter is calculated over all repeated *grid-search cross-validations*. Consequently, the optimal cross-validated tuning parameter is selected by the average prediction errors over all repetitions. Either way, the complete procedure in the inner loop for selecting the optimal model should be considered as a *protocol* (Krstajic, Buturovic, Leahy, & Thomas, 2014). The task of the outer loop is to assess the performance of the model selection *protocol* (Krstajic, Buturovic, Leahy, & Thomas, 2014). As in model selection, the *protocol* assessment is sensitive to the fold assignment. Therefore, Krstajic and colleagues also recommend a repeated *protocol* assessment. Finally, the exact modelling *protocol* can be used to make a cross-validated model fit. The model fit gives insights about possible predictors, supported by knowledge about the large-sample performance of the *protocol*.

As this section shows, concepts of machine learning are still very new to the psychological research domain despite the fact that statistical learning procedures can substantially improve the quality of the outcome. Choosing a statistical procedure for the sake of minimizing the prediction error of a model can give very valuable insights about the generalization qualities of the result. Creating models that generalize well is the objective of this work to predict *Spotify* audio features by situational, person-related and functional variables.

# 2 Method

## 2.1 Participants

Greb et al. (2019) collected the data of 119 participants (~55% women; mean = 24.4 years; SD = 4.4) recruited via the participant database of the *Max Planck Institute for Empirical Aesthetics* in Frankfurt am Main (Germany). The dataset only included participants with an indicated daily music consumption of at least two hours a day on five days per week. Participants were rewarded depending on the number of valid answers with up to 25€. In the course of this work, only 101 participants were included in the modeling process due to restrictions explained in the next sections (~53% women; mean = 24.4 years; SD = 4.6).

## 2.2 Measures

First, a pre-screening was set up by Greb and colleagues to evaluate the frequency of music listening of each participant (weekly and daily). Two items were rated on a nine-point scale, referring to the questions: 1) How often do you listen to music during the week?; and 2) How long do you listen to music on average? Additionally, participants shared information about their smartphone systems if they owned one.

### 2.2.1 Predictor Variables – Determinants of Music-Selection Behavior

Typical person-related variables were collected for each participant regarding age, gender and educational level. Additionally, participants answered questions about their *musical sophistication* (German version of the Gold-MSI (Schaal, Bauer, & Müllensiefen, 2014)), the intensity of music preferences (Schäfer & Sedlmeier, 2009), musical taste (Greb, Schlotz, & Steffens, 2017), and *Big Five* personality traits (German version of the IPIP-NEO-120 (Johnson J. A., 2014)).

Various items were measured repeatedly to describe the situation in which the music listening took place, the functionality of music listening and the properties of the currently consumed music. For more details and exact wording, see further Greb et al. (2019) and the Appendix. The complete questionnaire only started if participants positively answered the initial question, whether they were listening to music. The situation was described by the main activity that was done during the course of listening. Those categories were described in Greb et al. (2017) and included e.g. *housework, coping with emotions* and *relaxing*. Participants should also indicate if other people were present and in communication with them, if they actively chose the music, as well as about how much control they had over the music. The personal situation was assessed through ratings of the current mood and excitement at the beginning of the music listening by *valence* and *arousal* ratings (Russell, 1980). Additionally, participants should rate how important they thought their mood might have been in driving the decision to listen to music and if they were paying attention to the music. In the second section of the questionnaire, participants were asked to name the currently playing song, artist and/or musical style and rate it by eight musical characteristics each on a bipolar rating scale with seven scale points (*calming-exciting, slow-fast, sad-happy, less melodic-very melodic, less rhythmic-very rhythmic, simple-complex, peaceful-aggressive, less intense-very intense*). Furthermore, familiarity (*unknown-known*) and type of music (*vocal-instrumental*) were rated and information about the playback volume and liking of the song were asked for. The second section was not evaluated in this work, only the data of the songs' names and artists were processed and playback volume and liking were used. The third section included questions regarding the functionality of music listening. A subset of questions was generated from the functions of music listening in Greb et al. (2017), each of them covered by three questions. The corresponding functions were *intellectual stimulation*, *mind wandering & emotional involvement*, *motor synchronization & enhanced well-being*, *updating one's musical knowledge*, and *killing time & overcome loneliness*. The corresponding items delivered sum scores for each function as described in Greb et al. (2017).

### 2.2.2 Outcome Variables – Spotify Audio Features

Besides the person-related variables and the variables measured repeatedly in the *Experience Sampling Method* by Greb et al. (2019), musical properties for each actively chosen song (N=1021) were retrieved from the *Spotify Web API*[8]*.* The API (application programming interface) endpoints of *Spotify* return information encoded as *JSON* (JavaScript Object Notation) including the following audio features:

*Table 3 – Spotify Audio Features. Descriptions of the Spotify audio features were taken from the Web API website (Spotify, 2019)) (https://developer.spotify.com/documentation/web-api/reference/tracks/get-several-audio-features/)*

| | |
|---|---|
| **Acousticness** | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| **Danceability** | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| **Duration** | The duration of the track in milliseconds. |
| **Energy** | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |

[8] https://developer.spotify.com/

| | |
|---|---|
| Instrumentalness | Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| Key | The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. |
| Liveness | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| Loudness | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| Mode | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| Speechiness | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| Tempo | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| Time Signature | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). |
| Valence | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |

Not all features were used in this work. *Valence, energy, danceability* and *loudness* were selected for the modeling process. *Loudness* was additionally transformed by a simple mathematical operation to have output values normally distributed between 0.0 and 1.0 ($l_{linear} = 10^{\frac{l_{logarithmic}}{20}}$).

## 2.3 Design

As described in Greb et al. (2019) after the pre-screening participants rated on the person-related variables in a laboratory session using a tablet (*Samsung Galaxy Tab A1.7*). The app *movisensXS*[9] *Version 1.0.1* was used to measure the situational, functional and music-related variables in the course of the ESM. Participants with *Android* systems were able to use the app on their own device, others got a loan device (*Motorola Moto G3*). The whole measurement time window was 10 days from Friday to Sunday. Within that time period each participant received 14 random alarms per day (in an individual 14 hour time frame per day). This number was tested before and found to be acceptable. The alarms occurred within a minimum time span of 20 minutes between each other and were able to be postponed by 5, 10 or 15 minutes or just rejected by the participant. Besides the time-based sampling plan also an event-based plan was implemented in the way that participants were able to start the questionnaire proactively by starting the app.

## 2.4 Procedure

Greb et al. (2019) collected people's music-selection behavior in everyday life situations with an *Experience Sampling Method* (ESM). As described in Beal & Weiss (2003), the ESM delivers less biased results compared to common psychological techniques like retrospective reports for measuring dynamic behavior. Besides *between-person* differences it additionally measures the *within-person* differences for different situational settings. The data structure acquired in ESM studies is usually unbalanced and hierarchical (Fisher & To, 2012), which means that many observed datapoints (samples) are nested within the corresponding group (e.g. the individual or participant) but are not necessarily equally distributed to the different classes of output variables. Strong imbalance could possibly affect the outcome of statistical methods. Find the exact steps of the ESM described in Greb et al. (2019). Subsequent to the ESM, audio features were retrieved from the *Spotify Web API* for each actively chosen song. A playlist was created including all songs that were identified uniquely in *Spotify*. Due to not specific naming or ambivalent search results in *Spotify*, some data points were discarded at this stage, which was one reason, why only 101 participants remained in the modelling process. A *Python* script based on the *software development kit* (SDK) *Spotipy*[10] retrieved all *Spotify* audio features. In specific, the code used the *audio_features(tracks=[])* function of the *Spotipy* package which returned a *JSON* file including all *Spotify* audio features that were consequently parsed in the *Python* script.

## 2.5 Data Analysis

The data measured in the ESM was filtered by three rules. First, not complete cases were excluded, second, measurements of songs that could not be identified unambiguously in *Spotify* were excluded

[9] https://xs.movisens.com/

[10] https://spotipy.readthedocs.io/en/latest/#

and lastly, only cases with active music-selection behavior were used for further investigation, leaving 101 participants and 1021 listening situations. Prior to modeling, input variables were scaled and centered. There was one main analysis objective approached by this work:

### Predict Spotify audio features of actively selected music by measured person-related and situational variables and the functional use of music listening.

The data used for the prediction task had a hierarchical data structure accounting for *between-subject* and *within-subject* differences. To make this valuable in the modeling process, average values for each functional and situational variable were calculated to explain *between-subject* differences. Additionally, the deviation of each measurement from the person's mean value was calculated to account for the *within-subject* (situational) differences. Those values were originally assessed on the complete dataset from Greb et al. (2019), to prevent the analyses from outliers due to smaller sample size after removing several data points in consequence of the three data filtering rules as described before.

The implemented learning algorithm is based on the *$L_1$-penalized regression* (*Lasso regression*), proposed by Tibshirani (1996). In specific, the *glmmLasso* R-package from Groll and Tutz (2014) was used to solve the *Lasso regression* task on the outcome variables (audio features) *valence, energy, danceability* and *loudness*. The *glmmLasso()* function conducts a *Generalized Linear Mixed Effects Model* (GLMM) assuming a *normal distribution* and applies the *Lasso* shrinkage *hyperparameter* λ. A repeated 5-fold *nested cross-validation* procedure was implemented. This procedure, firstly, selected optimal tuning parameters for λ by *grid-search cross-validation* corresponding to different *protocols* (model selection procedures). All *protocols* were based on different versions of the *percentile-Lasso*. In the second step, a *protocol* assessment took place to make conclusions about the estimated generalization qualities of each *protocol*. These two steps were repeated several times. All *protocols* were compared to a reference model, the corresponding *random intercept-only* model, which is the model with the λ value that shrinks all variables to zero. The *protocol* with the least estimated generalized prediction error was finally used for a new *cross-validated* model selection and a final model fit on the complete dataset with the *lmer()* function of the *lme4* package of R by REML estimates (*restricted maximum likelihood*). All steps are explained in more detail in the following.

### Step A: Nested Cross-Validation – Protocol Assessment and Estimation of Generalization Qualities

The NCV was repeated 20 times in total which resulted in 100 *cross-validated* prediction error values (20x5-folds). This number was found to be reasonable in means of computational effort and expected test power. Each repetition conducted an individual *nested cross-validation* with five outer folds and five inner folds. The decision for five folds (instead of ten) was driven by the number of observations, respectively the number of participants in the remaining dataset. Furthermore, different researches treat 5- and 10-fold *cross-validation* as equally powerful (Krstajic, Buturovic, Leahy, & Thomas, 2014). To avoid leakage, the fold assignment was done at the level of the participants ($n_{Participant}$=101). Generally, a fold-split at the grouping level of nested data also delivers information about the estimated generalization qualities of models on unseen individuals. 101 participants were assigned to $k_{Out}$=5 folds in the outer loop, which means in average approximately 20 participants were assigned to the test set and 81 participants were assigned to the training set of the outer loop. The outer training set then was used in the inner loop to run several model selection *protocols*.

```
# Pseudo Code: Nested Cross Validation, outer loop
# repeat nested cross-validation by rep times (rep=20)
for r in 1:rep          {
                        outer_folds  = create $k_{out}$-outer folds
                        for k in 1:$k_{out}$        {
                                                outer_training_set = create set from outer_folds – $fold_k$
                                                outer_validation_set = create set with $fold_k$
cross-validate protocols (inner loop) ---> optimal_lambda = model selection (outer_training_set, $k_{in}$)
                                                optimal_model = train(outer_training_set, optimal_lambda)
                                                optimal_model_quality = test(outer_validation_set, optimal_model)
                        }
}
```

The remaining ~81 participants were assigned to $k_{in}$=5 inner folds, resulting in approximately 65 participants for each inner training set. The inner training sets were then trained on a grid of 100 λ-values. The sequence started with the maximal λ value. In general, the $\lambda_{max}$-value of *Lasso regression* models is supposed to shrink all regression coefficients to zero, meaning predictions are only made by the model's *random intercept* value. Consequently, 100 different linear distant λ-values between λ=0 and $\lambda_{max}$ were trained and tested for each $k_{in}$ fold. The resulting 100-by-5 matrix delivered the average mean-squared-error for each λ over each inner fold (also known as *cross-validation error*). The best λ with the smallest CV-error was selected for the trained outer fold. Following the suggestions of Roberts & Nowak (2014), the model selection was stabilized against variances in the fold assignment during *cross-validation* by repeated inner fold assignments (per=100), resulting in 100 optimal tuning parameters for each outer training fold. Afterwards, ten different percentile values were chosen from those optimal tuning values (50th-, 55th-, 60th-, 65th-, 70th-, 75th-, 80th-, 85th-, 90th- and 95th-percentile). Each percentile selection refers to its own *protocol* and was then trained on the outer training set and tested on the outer test set in the course of *protocol* assessment, returning the prediction errors for each $\lambda_{Opt}$(percentile). The same was done with a $\lambda_{max}$-model for each outer training set.

```
# Pseudo Code: Nested Cross Validation, inner loop or protocol
# repeat inner loop many times and select percentile values to stabilize model selection (per=100)
for p in 1:per        {
                inner_folds  = create $k_{in}$-inner folds
                for i in 1:$k_{in}$        {
                                inner_training_set = create set from inner_folds – $fold_i$
                                inner_validation_set = create set with $fold_i$
                                for lambda in sequence(maximal_lambda:0, by= -maximal_lambda/100) {
                                        model_lambda = train(inner_training_set, lambda)
                                        model_lambda_quality = test(inner_validation_set, model_lambda)
                                }
                                optimal_lambda = append lambda with best CV-error
                }
                percentile_lambda = select the 95th or 85th – percentile (and so on) from optimal_lambda
                # each percentile_lambda is used as the optimal_lambda of the corresponding protocol in the
                #outer_loop
}
```

### Step B – Model Selection with Best Protocol

The final *cross-validation* followed the exact *protocol* that was chosen by the *protocol* assessment in the *nested cross-validation* and finally selected the tuning parameter. The choice of the best *protocol* was driven by the smallest estimated generalized prediction error and by inspection of a stable model selection. A model was finally fitted on the complete dataset using the optimal λ. The final re-fit with the *lmer()* function (using REML=TRUE) delivered the estimates and the corresponding *p-values*. The typical significance level of $\alpha=0.05$ was considered. *Conditional* and *marginal* $R^2$ values were computed of the final models with the *r.squaredGLMM()* function of the *MuMin* package of R according to Nakagawa et al. (2017).

$$R^2_{marginal} = \frac{\delta_f^2}{\delta_f^2 + \delta_\alpha^2 + \delta_\varepsilon^2} \tag{8}$$

$$R^2_{conditional} = \frac{\delta_f^2 + \delta_\alpha^2}{\delta_f^2 + \delta_\alpha^2 + \delta_\varepsilon^2} \tag{9}$$

Additionally ICC (Intra-class correlation coefficients) values for each *random intercept-only* model were calculated to observe the relative amount of variance explained by the models.

$$ICC = \frac{\delta_\alpha^2}{\delta_\alpha^2 + \delta_\varepsilon^2} \tag{10}$$

Note:

$\delta_f^2$ – denotes the variance calculated by the fixed effects components of the model

$\delta_\alpha^2$ – denotes the variance calculated by the random effects components of the model

$\delta_\varepsilon^2$ – denotes the residual variance

# 3 Results

The results of the NCV are shown in the Figure 4, Figure 5, Figure 6 and Figure 7. Each figure consists of two boxplots. The upper one shows the distribution of the 100 cross-validated prediction errors of each outer test fold. The averaged prediction error of each *protocol* is indicated by a circle and the averaged error of the reference model ($\lambda_{max}$-model) is indicated by a dashed line. Those values are corresponding to the estimated generalized prediction error of each *protocol*. The lower boxplot shows the distribution of the cross-validated optimal λ values of each outer training fold indicating the stability or instability of each *protocol* in regards of model selection.

As Figure 4 shows, each *percentile-Lasso protocol* in average had a worse generalized prediction error for *valence* than the reference model. Additionally, all *protocols* tended to switch randomly between high and low λ values. The best *protocol* was the reference (*random intercept-only* model). The variance of selected λ values was very high for all *protocols*. The variance of the estimated generalized prediction error got higher the more predictors were included. In general saying, models that tended to not include any predictors had a lower prediction error. Still, there was also a decent chance that the best percentile *protocol*, the *95th-percentile-Lasso*, would select some predictors by means of an unfavorable fold assignment. The high variance in the model selection and worse prediction errors in average of all *protocols* indicate strongly that the investigated predictors do not have an effect on the selection of values on the *valence* dimension measured by *Spotify*. The *random intercept-only* model was selected for the final modelling. Still a complete CV was done for all models to achieve similar representations of the results.

In contrast, the NCV of the variable *energy* showed that all *protocols* were improved by the inclusion of predictor variables. The perfect tradeoff was achieved by the *70th-percentile-Lasso*. Low *percentile-Lasso protocols* (50th to 65th) selected too many variables which did not generalize well, therefore the prediction error on unseen data tended to be worse. From the 80th to the 95th-percentiles, the *protocols* also switched between low and high λ values. High λ-values (low indices) indicated that the exclusion of variables yielded to worse predictive performances. It can be assumed that the inclusion of certain predictors generally improves the prediction. The *protocol* chosen for the final CV of the optimal tuning parameter was the *70th-percentile-Lasso*. This *protocol* had the lowest estimated generalized prediction error, the smallest variance of prediction errors and a well-defined range of optimal tuning parameters chosen during the NCV, thus acted very stable.

*Danceability* was in average predicted the best by the *50th-percentile-Lasso*. However, all *protocols* improved the estimated generalized prediction accuracy. As Figure 6 shows, the plot of indices of the selected tuning parameter indicates that the exclusion of variables yielded to a worse performance. It is possible that a *protocol* with a lower percentile value of optimal tuning parameters (e.g. the 45th percentile) would have yielded a better result. But the *50th-percentile-Lasso* was stable and therefore an adequate choice for the final CV. Compared to all other percentile *protocols*, the *50th-percentile-Lasso* did not choose an outlier λ value by chance due to the fold assignment. Despite the fact that the *55th*- and *60th-percentile-Lassos* had a similar average prediction error, those *protocols* tended to be unstable.

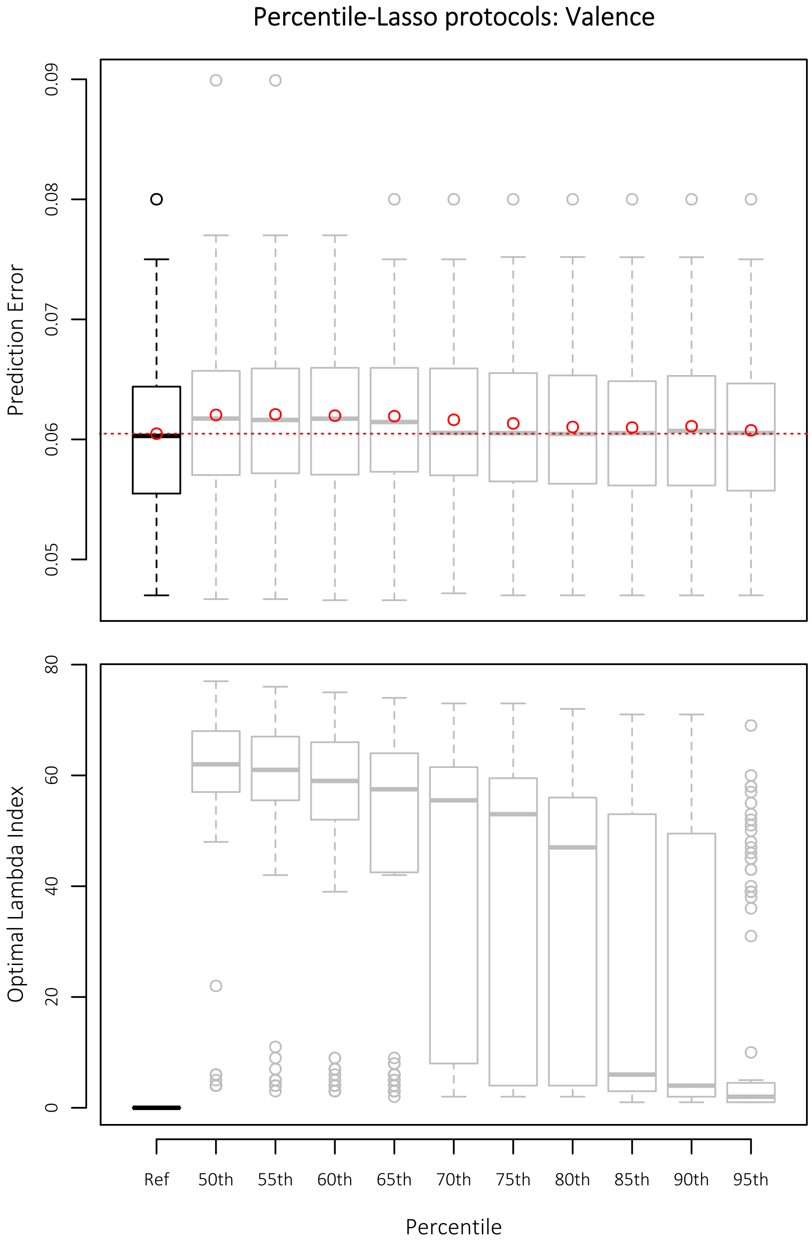

*Figure 4 - Protocol Assessment: Valence. The upper boxplot shows the distribution of cross-validated MSE values for each protocol. The red circles indicate the average MSE for each protocol over all repeated NCVs. The horizontal dashed line indicates the average MSE of the random intercept-only model (or $\lambda_{max}$-model) which is the reference. The nested cross-validation shows that no protocol improves the estimated generalized prediction error in comparison to the reference. This indicates that no variable has any generalized predictive power. The random intercept-only model was selected for the final model selection/fit.*

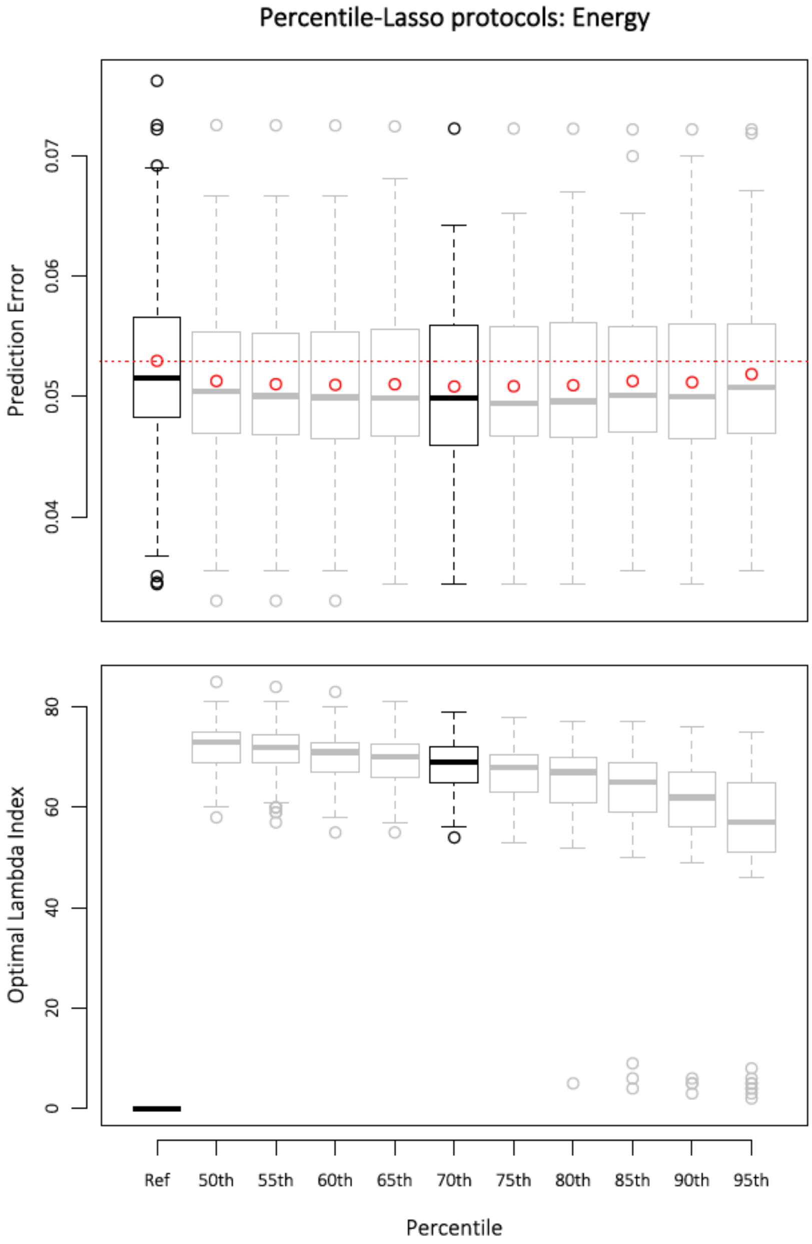

*Figure 5 - Protocol Assessment: Energy. The upper boxplot shows the distribution of cross-validated MSE values for each protocol. The red circles indicate the average MSE for each protocol over all repeated NCVs. The horizontal dashed line indicates the average MSE of the random intercept-only model (or $\lambda_{max}$-model) which is the reference. The nested cross-validation shows that each protocol improves the estimated generalized prediction error compared to the reference model. The 70th-percentile-Lasso was selected as best protocol for the final model selection achieving the best estimated generalized prediction error and a stable range of selected tuning parameters.*

Also the *protocols* for predicting *loudness* had better prediction errors in average than the reference had. Similar to *energy,* a perfect *bias-variance tradeoff* was achieved. Consequently, the *80th-percentile-Lasso* was selected for the final CV with the lowest estimated generalized prediction error and a stable range of optimal tuning parameters chosen in the NCV. Again, higher percentiles tended to be unstable and also jumped between high and low λ values. Lower percentiles seemed to be too passive in shrinking and included too many variables.

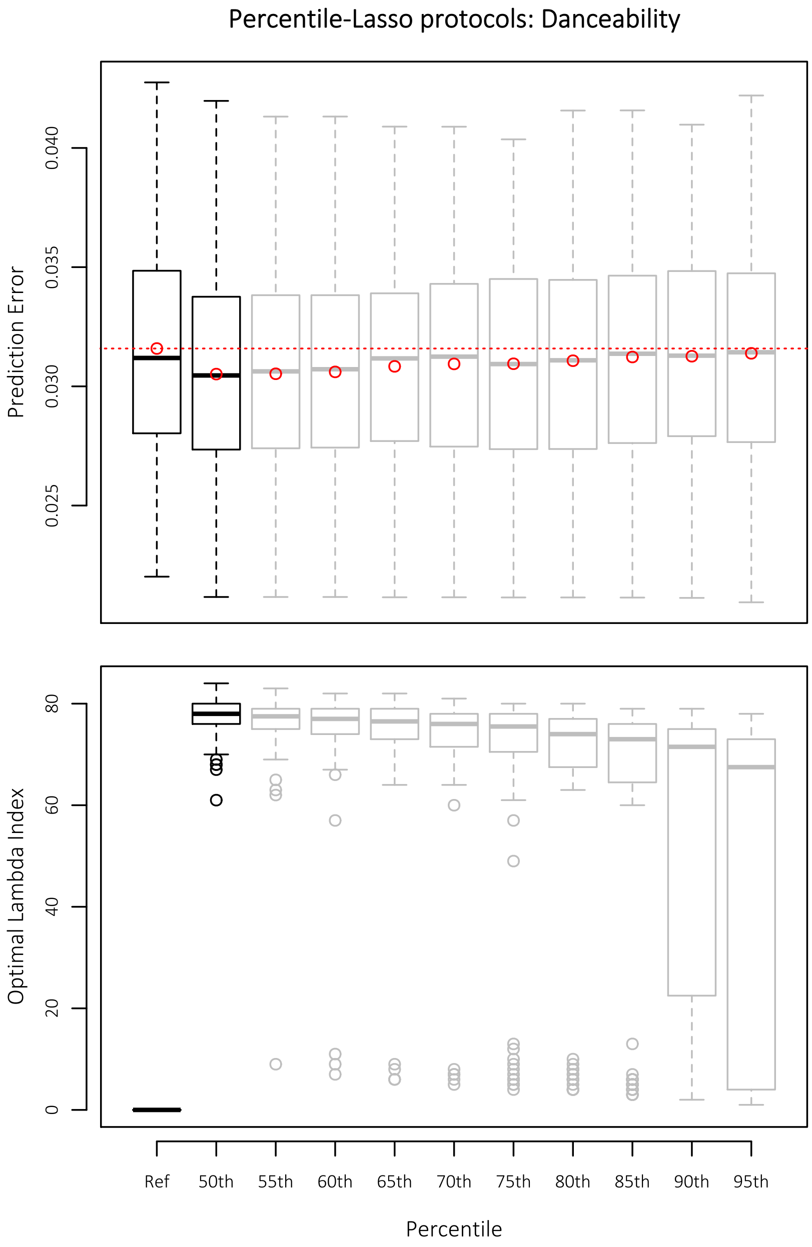

*Figure 6 - Protocol Assessment: Danceability. The upper boxplot shows the distribution of cross-validated MSE values for each protocol. The red circles indicate the average MSE for each protocol over all repeated NCVs. The horizontal dashed line indicates the average MSE of the random intercept-only model (or $\lambda_{max}$-model) which is the reference. The nested cross-validation shows that each protocol improves the estimated generalized prediction error compared to the reference. The 50$^{th}$-percentile-Lasso was selected as best protocol achieving the minimal estimated generalized prediction error in average and a stable range of selected tuning parameters.*

All results show that the choice for the optimal λ value was highly sensitive to the fold assignment but mostly stable for the selected *protocols*. Still it can be assumed that the average prediction error supports a very reasonable choice of the best *protocol*, since this value indicates the estimated generalized prediction error on unseen data. High variance in MSE values can be explained by unfavorable fold assignments. Still this is simultaneously valid for all *protocols* and does not have an impact on the relative comparison. Both attributes, stable selection and lowest estimated generalized prediction error, were considered in the course of *protocol* selection for the next step (B): cross-validated model-selection with the best *protocol*.

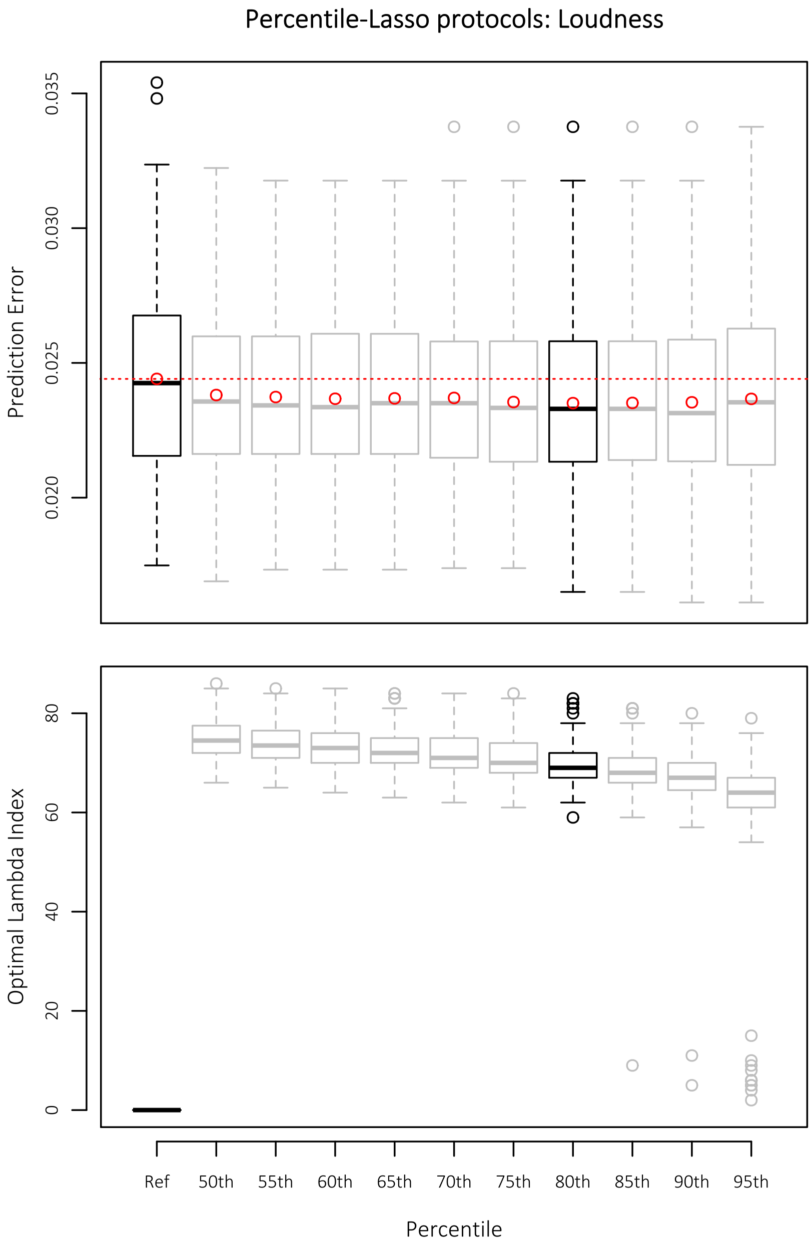

*Figure 7 - Protocol Assessment: Loudness. The upper boxplot shows the distribution of cross-validated MSE values for each protocol. The red circles indicate the average MSE for each protocol over all repeated NCVs. The horizontal dashed line indicates the average MSE of the random intercept-only model (or $\lambda_{max}$-model) which is the reference. The nested cross-validation shows that each protocol improves the estimated generalized prediction error compared to the reference. The 80$^{th}$-percentile-Lasso was selected as best protocol achieving the minimal estimated generalized prediction error in average and a stable range of selected tuning parameters.*

The final cross-validated model selections were done for each outcome variable according to the best *protocols* selected in the NCV. Figure 8 shows the trace plots of the estimated coefficients for each λ of the grid. The vertical red dashed lines indicate the cross-validated optimal λ values corresponding to the selected *percentile-Lasso protocol*. The vertical grey dashed lines indicate the other theoretical λ values of the other *protocols*. In the cases of *valence* and *energy* each *protocol* would have chosen different predictors, whereas for the other models each *protocol* would have yielded a similar result. This underlines the success of controlling the *bias* by selecting different percentile values. The grey trace lines indicate that the corresponding variable did not survive the model selection procedure at the point of the optimal λ. The colored lines indicate the remaining variables at that very point. The resulting models were finally re-fitted with *lmer()* and yielded the models as displayed in Table 4. Additionally, corresponding R² values and ICC were calculated giving insights about the proportions of variance explained by the model itself and the grouping factor (participants).

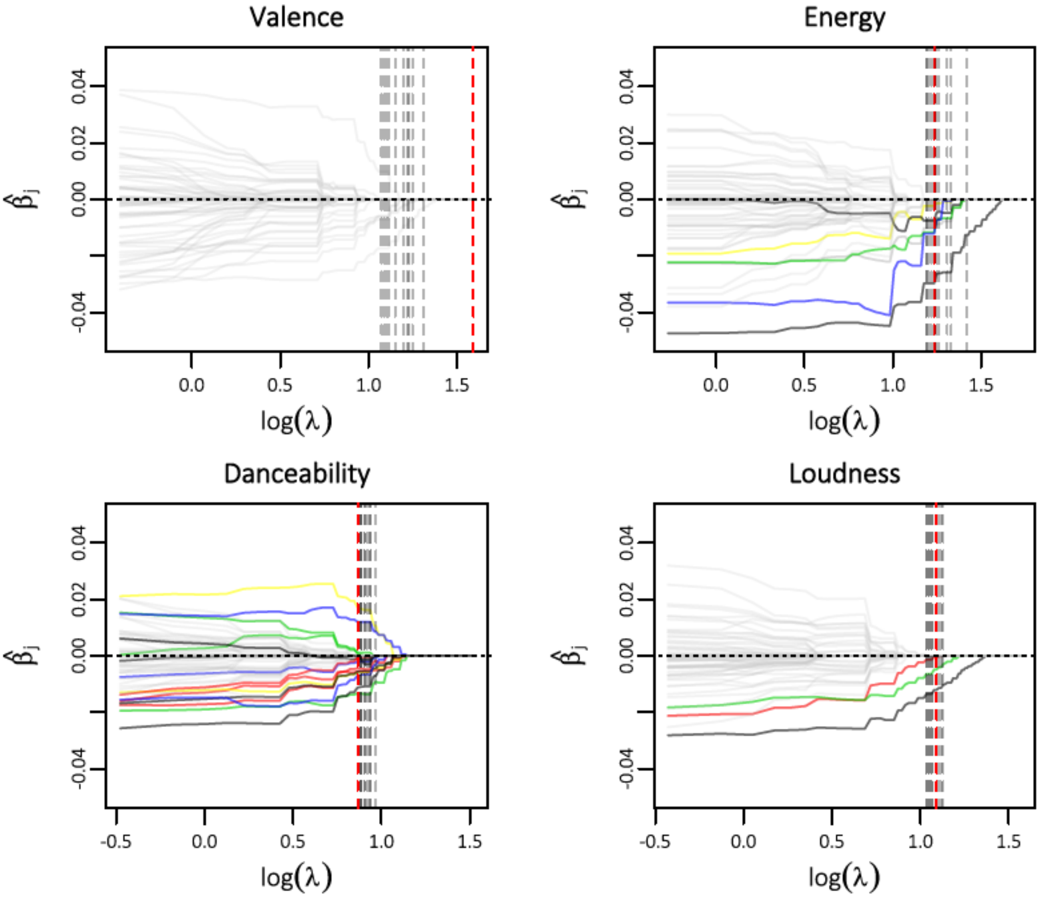

*Figure 8 – Trace plot of coefficients estimates. The final fit for each λ shows how estimated coefficients ($\hat{\beta}_j$) shrink with higher λ values. A colored line indicates that the corresponding variable finally survived the model selection procedure at the point of the red dashed line which corresponds to the optimal λ selected by the best protocol. The vertical grey dashed lines indicate the λ values for the other percentile-Lasso protocols. Comparing the selection procedures for Danceability and Loudness shows that protocols selected fairly similar tuning parameters. In the case of Energy and Valence it shows that different protocols would have yielded to more different models. This underlines the unstable model selection of some percentile-Lasso protocols, which rectifies the control of the percentile value.*

*Table 4 - Summary of Linear Mixed Effects Models after cross-validated model selections.*

| | Valence | Energy | Danceability | Loudness |
|---|---|---|---|---|
| Selected Protocol | Random Intercept-Only Model | Repeated 70th – Percentile – Lasso | Repeated 50th – Percentile – Lasso | Repeated 80th – Percentile – Lasso |
| $\lambda_{max}$ | 39 | 53 | 33 | 37 |
| Index of $\lambda_{opt}$ | 1 | 68 | 78 | 67 |
| $\lambda_{opt}$ | 39 | 17.13 | 7.33 | 12.33 |
| Estimated generalized MSE | 0.061 | 0.051 | 0.031 | 0.024 |
| $R^2_{marginal}$ | 0.000 | 0.104 | 0.151 | 0.087 |
| $R^2_{conditional}$ | 0.186 | 0.292 | 0.196 | 0.305 |
| ICC | 0.186 | 0.251 | 0.159 | 0.296 |

Fixed Effects

| Parameter | Estimate (StdErr) | Estimate (StdErr) | Estimate (StdErr) | Estimate (StdErr) |
|---|---|---|---|---|
| Intercept | **0.4570 (0.0140)***** | **0.6659 (0.0131)***** | **0.6258 (0.0459)***** | **0.5636 (0.0551)***** |
| Age | | | -0.0029 (0.0018) | **-0.0045 (0.0022)*** |
| Gender (Women) | | | 0.0241 (0.0159) | |
| **Personality Traits** | | | | |
| Extraversion (Big5) | | | 0.0082 (0.0079) | |
| **Functions of music listening (Level 1: Situational)** | | | | |
| Intellectual Stimulation | | **-0.0179 (0.0075)*** | **-0.0159 (0.0052)**** | |
| Mind Wandering & Emotional Involvement | | **-0.0252 (0.0076)***** | | |
| Motor Synchronization & Enhanced Well-Being | | **0.0501 (0.0065)***** | **0.0139 (0.0053)**** | **0.0250 (0.0042)***** |
| **Mood (Level 1: Situational)** | | | | |
| Valence | | | 0.0074 (0.0052) | |
| **Functions of music listening (Level 2: Person-related)** | | | | |
| Intellectual Stimulation | | | **-0.0357 (0.0087)***** | |
| Motor Synchronization & Enhanced Well-Being | | 0.0139 (0.0168) | 0.0042 (0.0110) | |
| Updating One's Musical Knowledge | | | 0.0139 (0.0079). | |
| Killing Time & Overcoming Loneliness | | 0.0289 (0.0163). | 0.0020 (0.0101). | **0.0224 (0.0097)*** |
| **Mood (Level 2: Person-related)** | | | | |
| Valence | | **0.0284 (0.0139)*** | **0.0180 (0.008)**** | |
| **Musical Taste** | | | | |
| Techno & EDM | | | **0.0206 (0.007)**** | |
| Volksmusik & Schlager | | | 0.0079 (0.008) | |
| Rock & Metal | | | **-0.0175 (0.007)*** | |
| **Presence of others (Level 1: Situational) in reference to *being alone*** | | | | |
| Others Present & No Communication | | | 0.0031 (0.005) | |

|  |  |  |  |  |
|---|---|---|---|---|
| Others Present & Communication |  |  | **0.0150 (0.005)**** |  |
|  |  | **Random Effects** |  |  |
| **Standard Deviation** |  |  |  |  |
| Participant (Intercept) | 0.1061 | 0.0986 | 0.0378 | 0.0724 |
| Residuals | 0.2217 | 0.1916 | 0.1587 | 0.1291 |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1, Number of observations = 1021, Number of Participants = 101

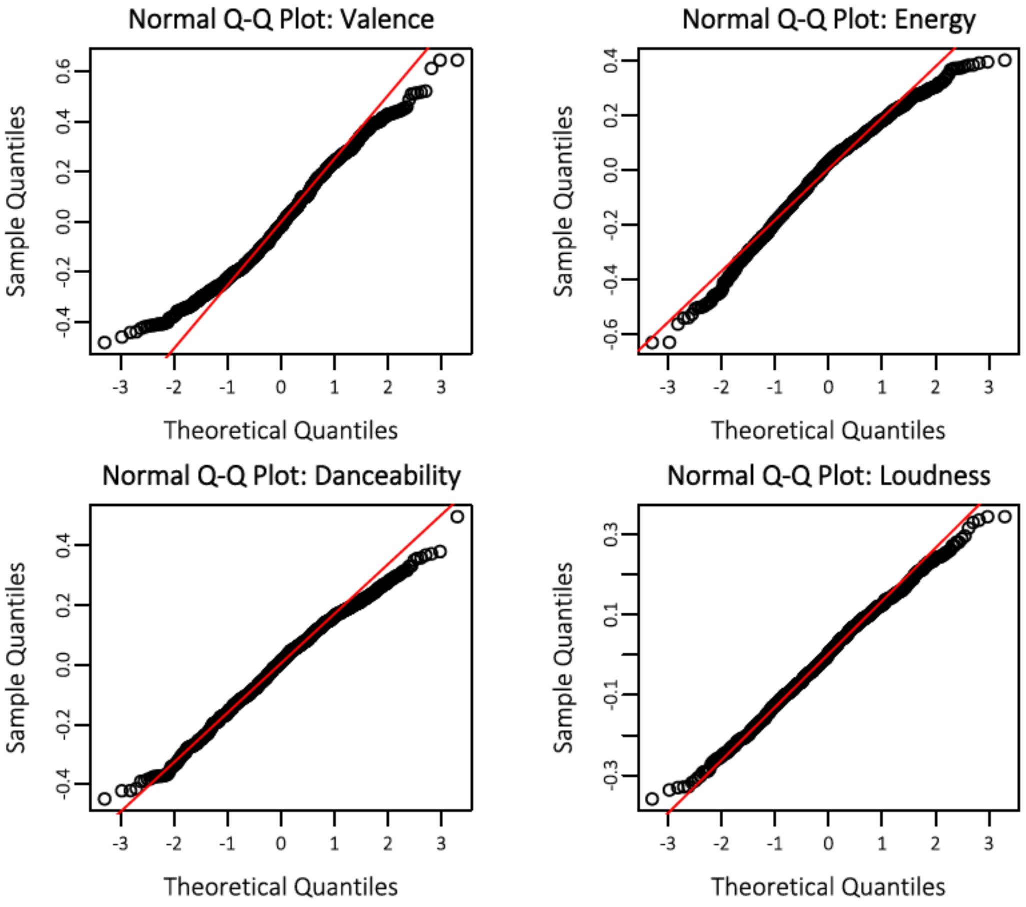

*Figure 9 - Normal Quantile-Quantile residual plots of the final model-fits. The Q-Q plot shows graphically that in the case of Valence the normal distribution of the residuals is not quite given. The same tendency can be seen for Energy. The other models have visually represented normally distributed residuals but as a Shapiro-Wilk normality test showed, no model had normally distributed residuals. With many data points the result can still be trusted and the distributions are not completely off normality.*

At first, all final models were tested in regards of linear regression model assumptions, namely *homoscedasticity* and *normality of residuals*. Models with strong violations should generally be rejected. Although the Q-Q plots (Figure 9) of the residuals of the final models did not display a normal distribution and also Shapiro-Wilk normality tests for each model indicated that the null-hypothesis (saying that the residuals were drawn from a normal distribution) was rejected for all models, it is generally assumed that with a high number of observations, as it is here, a result with only small violation can still be trusted (Kutner, Nachtsheim, Neter, & Li, 2013). *Homoscedasticity* was only judged by inspection of the plots in Figure 10. All models seem to have slight *heteroscedastic* residuals. Especially in the case of *energy* it might indicate *biased standard errors* and hypothesis tests. However, this work reports all coefficients that survived the model selection procedures regardless of their statistical significance. In the cases of *energy, danceability* and *loudness*, the inclusion of coefficients was shown to improve the estimated generalized prediction accuracy. Thus, the *protocol* selected most likely statistically relevant predictors in those models.

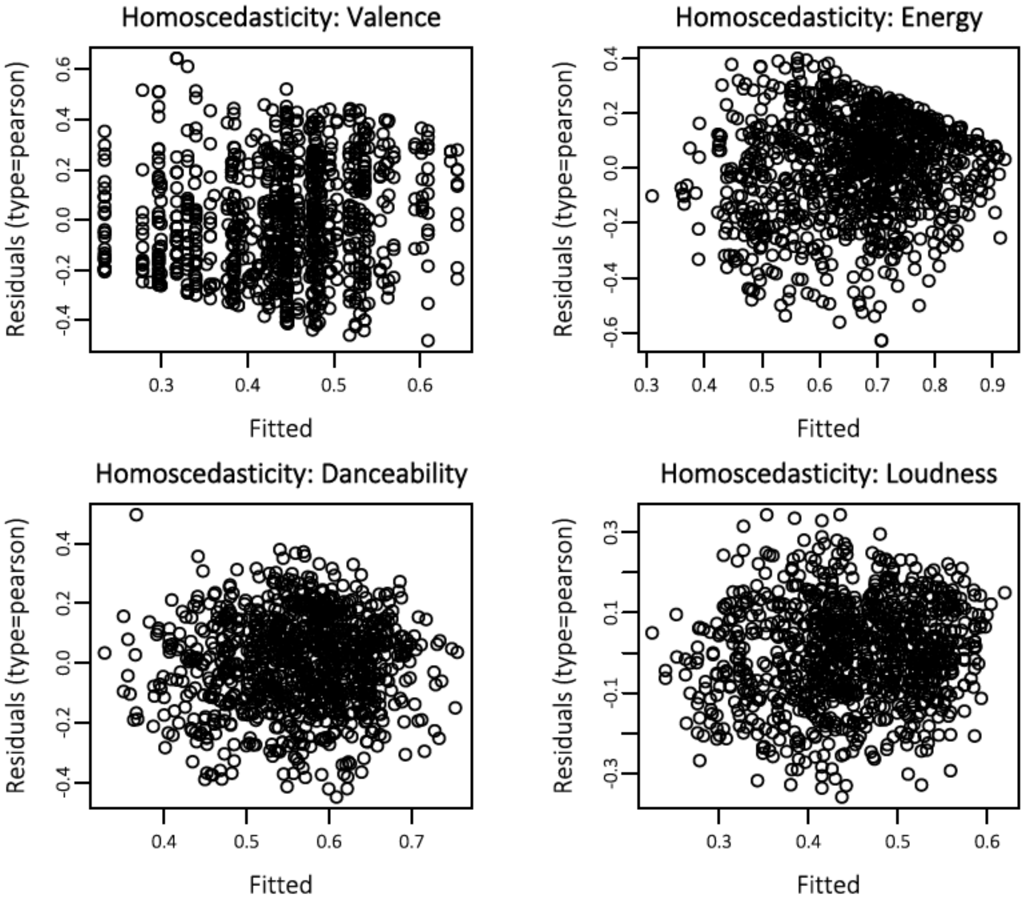

*Figure 10 - Homoscedasticity of residuals versus fitted values. By inspection, all models seem to violate the homoscedasticity rule slightly. Mostly invariant are the residuals of the energy model with very narrow variation towards high energy values and wide variation for smaller values.*

No predictor was selected in the *valence* model. Corresponding to the NCV this result is reliable. The $R^2_{conditional}$ corresponds to the ICC since no fixed effect was included in the final model. Therefore, the proportion of variance explained by the grouping in participants was around 18.6%. Respectively, around 81.4% of the variance was neither explained by the model nor the grouping factor and might be explained by other situational influences.

*Energy* was finally predicted by several functions of music listening with statistically significant effects. On the situational-level, *intellectual stimulation* and *mind wandering & emotional involvement* had negative effects on *energy,* whereas higher values for *motor synchronization & enhanced well-being* also resulted in higher output values for *energy.* The two functions on the personal-level *killing time & overcoming loneliness* as well as the initial *valence state* both had a significant positive effect on the *energy* level of the selected music. Additionally, *motor synchronization & enhanced well-being* as person-related predictor also survived the model selection procedure and showed a positive effect on *energy*, but not significantly. $R^2_{marginal}$ indicates that around 10.4% of the total variance was explained only by the fixed effects of the model, while 29.2% of the total variance was explained by the complete model (random and fixed effects). The ICC value considering the *random intercept-only* model indicated that around 25.1% of the total variance was explained by the grouping factor (participants), supporting the need for a *Mixed Effects Model*. Considering the case that 18.8% of the total variance was explained by the model by means of the random effects, this indicates that a certain amount of explained variance by the grouping factor of the *random intercept-only* model was absorbed by the fixed effects of the model.

The model of *danceability* showed the most interactions between predictors and the outcome variable. While *age* showed a tendency towards a negative effect (p=.11), meaning older people tended to select less danceable music, *gender* showed a tendency towards a positive effect (p=.13). This indicates that women selected more danceable music than men. *Danceability* was the only model that included a personality trait, namely the *Big Five* dimension of *extraversion.* The effect was positive but still not significant. The tendency indicates that more *extraverted* people tend to select more *danceable* music. Situational-wise, two functions of music listening and the initial mood (*valence*) showed effects. Corresponding to intuitive thinking, situations in which music was selected for *intellectual stimulation* had a significant negative effect on the *danceability,* while music selected in situations for *motor synchronization & enhanced well-being* was significantly more danceable. A person being initially more positive in the specific situation than in average also had a positive effect on the selection of more danceable music, although not a significant one. Person-related factors additionally influenced the selection of danceable music: while persons that generally tended to use music for *intellectual stimulation* selected less danceable music, other *between-person* differences of functional music usage also had positive effects on the selection of more danceable music, although those were only tendencies and did not show significant effects (*motor synchronization & enhanced well-being* (p=.70), *updating one's musical knowledge* (p=.08) and *killing time & overcoming loneliness* (p=.05)). A significant person-related effect was found for positive initial mood on the selection of more danceable music. *Danceability* was also the only model that included musical taste factors. High ratings for *Techno & EDM* had a significant positive effect on the selection of more danceable music. The same tendency was existent for high liking ratings of *Volksmusik & Schlager*, although not significantly (p=.32). People rating high on *Metal & Rock* selected significantly less danceable music. Lastly, the *presence of others* also had a positive effect on the selection of more *danceable* music. Those predictors are compared to the reference *being alone*. The *presence of others* accompanied by *communication* had a stronger (significant) effect than *presence of others* without *communication.* Still both effects were positive. Comparing the ICC and $R^2$ values indicates that the fixed effects of the model explained some of the differences incorporated in the initial grouping factor of the *random intercept-only* model. The initially calculated 15.9% of explained variance by the grouping factor only, was reduced in the model to 4.5% explained variance of the random effects, while fixed effects accounted for 15.1% of the variance.

The model of *loudness* included only three predictors, but all had a significant effect on the selection. In situations when individuals selected music rather for the function of *motor synchronization & enhanced well-being* people selected music with higher *loudness* levels. Person-related, people that in average listened more to music in the course of *killing time & overcoming loneliness* also selected louder music. Similar to *danceability*, the *loudness* model included a negative age effect on the selection of less loud music. The ICC value indicates that approximately 29.6% of the total variance was explained by the grouping in participants. While 8.7% of the variance was explained by the fixed effects only, 30.5% were explained by the model including random and fixed effects. Thus, the random effects in the model explained around 21.8% of the total variance. Again, some random effects of the *random intercept-only* model were explained by the included fixed effects in the model.

Differences in estimated generalized prediction errors (Table 4) also need to be considered in regards of the variance of the outcome variable itself. While *valence* had a very spread value range between 0.0 and 1.0 ($Var_{Valence}$ = 0.059), the outcome variable *energy, danceability* and *loudness* had less variance ($Var_{Energy}$ = 0.052; $Var_{Danceability}$ = 0.031; $Var_{Loudness}$ = 0.024). That the selected *protocols* still provide prediction improvements can be seen in relation to the *random intercept-only* models (references) as shown in the boxplots above, of course except for the *valence protocol*.

# 4 Discussion

This work investigated the prediction of *Spotify* audio features of actively selected music in the daily life by situational, person-related and functional variables of music listening. The data was measured by an *Experience Sampling Method* by Greb and colleagues (2019) and the corresponding audio features were retrieved from the *Spotify Web API*. Different statistical learning *protocols* based on the *percentile-Lasso* were evaluated by their estimated generalized performance in a *nested cross-validation*. The best *protocol* for each outcome variable was chosen for the final cross-validated model selection of the most important predictors of *Spotify* audio features. The final model-fits showed several statistically significant effects on the audio features *energy, danceability* and *loudness*. The feature *valence* could not be predicted reliably by any variable. Those results are further discussed in comparison to existing related works. Afterwards, methodological challenges and limitations are considered and future directions are outlined.

## Predictors of Audio Features and Music-Selection Behavior

The prediction of objectively computed audio features by contextual, functional and person-related variables has not been addressed before. Related works have so far considered the prediction of subjectively rated music characteristics by contextual, functional and person-related predictors (Greb, Steffens, & Schlotz, 2018), and possible mediating roles of music listening functions on situational and person-related predictors of music-selection (Greb, Steffens, & Schlotz, 2019). The prediction of music listening situations and corresponding subjective judgements of music and the situation (Krause & North, 2017), as well as the prediction of selected music by their emotional functionality (Randall & Rickard, 2017) were also investigated. Still, all works have assessed subjective music perception characteristics and not objective audio features. In this conglomerate, the works of Greb and colleagues (2018; 2019) are the closest related to this one, methodologically as well as from an analysis point of view. The latest work additionally used the same data assessment technique, the ESM, and collected the behavioral and personal data used in this work. The outcome variables can also be related to each other in several aspects, as described in the following. The *Spotify* audio feature *valence* that was evaluated in this study, corresponds to Greb and colleagues' subjective dimension of *sad-happy* music. *Energy* refers to the *calming-exciting* and in some perspectives also to the *less intense-very intense* dimensions of Greb et al. (2019). Greenberg and colleagues (2016) showed that *intensity* as a psychological attribute was highly correlated with one principal dimension that people use to describe music, namely *arousal*. It was not scoring high on the *valence* and *depth* dimensions they found. Still, *intensity* might be understood differently in different contexts. As Ali and Peynircioğlu (2010) described, either listening to familiar or unfamiliar music changes the intensity of *perceived* emotions. Reversely, this might also result in an emotionally biased rating of the *perceived intensity* of self-selected music (as in Greb et al. (2019)) in contrast to unfamiliar music listening as in Greenberg et al. (2016). Thus, the small amount of *intensity* loading onto the *valence* and *depth* dimensions in Greenberg's and colleagues' study might not hold up in Greb et al. (2019), and might explain differences between *energy* values in this work and *intensity* ratings in Greb et al. By *Spotify's* explanation, high *energetic* values also correspond to the feeling of fast songs, which is congruent with the *slow-fast* dimension in Greb et al. (2018). In line with that, effects on the *perceived arousal* of a song were found by altering the tempo (Husain, Thompson, & Schellenberg, 2002). According to *Spotify*, both variables (*valence & energy*) are perceptual parameters and correspond to the well-established V-A emotion model. *Danceability* might be comparable to the *less rhythmic-very rhythmic* dimension as described by Greb and colleagues. As *Spotify* states, also tempo is considered in this measure. Thus, this variable might also be closely related to the *energy* feature of *Spotify*. Lastly, *loudness* is a subdomain of *energy* and corresponds to the *loudness perception* of humans. This is a lower-level representation of *energy* and can also be related to the *less intense-very intense* dimension according to Greenberg et al. (2016) but might be different for reasons of familiarity bias as explained before.

### Valence

Starting with the most divergent results between this work and the related works, the analyses revealed no predictive power of contextual, functional and person-related variables for the audio feature *valence.* A decent proportion (18.6%) of the total variance was explained by the grouping factor (random differences between participants), which is very similar to the results of Greb et al. (2018: ICC = 18%; 2019: ICC = 17%). In contrast, Greb and colleagues (2018; 2019) found several significant predictors for the subjective ratings on the *sad-happy* dimension. Those included activities during music listening like *relaxing & falling asleep* or *coping with emotions,* both having a negative effect. And functions of music listening like *motor synchronization & enhanced well-being* and *updating one's musical knowledge* positively predicted happier music, while higher values on the function *mind wandering & emotional involvement* predicted more sad music. Greb and colleagues as well as Randall and Rickard (2017) additionally found a mood-congruent selection behavior, which would have implied that the initial mood possibly would have been predicting the feature *valence* in this study, which was also not the case. The divergence of those results might have very different reasons. They are mostly explained by different facets of the *semantic gap* of emotion recognition tasks and also include the inherent problem of confusion between *perceived* and *induced* emotions.

First, *valence* as an audio feature and *sad-happy* as a subjective measure of the *perceived valence* seem to be semantically very closely related, though in fact they are not the same. While subjective measures always depend on personal differences and situational dependencies of perception, a single value of the emotional quality of a song does not cover aspects of inter- and intra-rater differences. Also, people might not select music by its *perceived* emotional value, as measured by *Spotify*, but by its implications on the intentional emotional use, incorporated in the *induced* emotions. Additionally, this is in accordance with the view that the *induction* of emotion by music is not properly described by commonly used emotion representations as the *categorical* one using basic emotions (Zentner, Grandjean, & Scherer, 2008) and immanent critics on the simplifications of a two-dimensional V-A model, as described earlier. As the *sad-happy* ratings would be considered more personal and biased by *induced* emotions, the *Spotify* feature delivers an objective measure of the *perceived valence*. Notably, the notion saying that *perceived* emotions are more consistently rated by subjects in contrast to *induced* emotions (Song Y. , Dixon, Pearce, & Halpern, 2016) cannot necessarily be supported following this argumentation, since both models showed similar proportions explained by the grouping factors.

Second, also the exact computation of the *valence* feature is not clear at all. A tiny hint is given on *The Echonest Blog*[11] about how *valence* is computed in the current *Spotify* API (formerly *The Echonest API)*:

> *One key aspect: We have a music expert classify some sample songs by valence, then use machine-learning to extend those rules to all of the rest of the music in the world, fine tuning as we go.*
>
> The Echonest Blog (2019)

This procedure is certainly not very well defined and leaves a lot of room for speculation. So far there is no comparison of *Spotify* audio features and other approaches of feature computation in terms of accuracy and consistency with subjective ratings. Previous studies also suggested that especially the dimension of *valence* shows less consistency for predictions (Huq, Bello, & Rowe, 2010; Song & Dixon, 2015). Since recent works figured that the performance of *valence* prediction could be substantially improved by the inclusion of multi-modal criteria (Delbouys, Hennequin, Piccoli, Royo-Letelier, & Moussallam, 2018) typical approaches that do not include e.g. lyrics in the prediction process of the *perceived* emotion usually underperform, since aspects of *valence* are not sufficiently mirrored in audio properties. Thus, it also needs to be considered that those *valence* values do not exactly indicate the *perceived* emotional value, not even on a global level. However, it cannot be generalized that people do not select music by *perceived* or *induced* emotional qualities in the dimension of *valence*. It can only be concluded that the feature *valence,* the way *Spotify* calculates it, cannot be predicted. A closely related

[11] http://blog.echonest.com/post/66097438564/plotting-musics-emotional-valence-1950-2013

consideration arises from the results of the context-based recommendation approach of Dias et al. (2014). After retrieving *The Echonest* audio features of songs that were clustered by subjects into corresponding activity categories, they performed a discriminative analysis to find the most important features fitting into those categories. Respectively, *acousticness, energy, loudness* and *tempo* were the most discriminative features for the different activities. It showed that the feature *valence* was not able to sufficiently differentiate the categories of different activities. Again, this does not necessarily imply that people do not choose music by its emotional value, simply the representation could not be suitable or discriminative enough.

Third, that mood-congruent music-selection was not present for the *valence* dimension in this study also puts the assessment of subjectively rated music characteristics up to discussion. Subjective reports are potentially biased and stereotypical. This can also lead to ratings of music characteristics that do actually just mirror the personal situation or socially expected answers. In this regard, the distinction between *perceived* and *induced valence* becomes very fuzzy. Subjects might mix up the distinction unconsciously and project *induced* emotions onto the *perceived* emotions. In Greb et al. (2019), the mood congruent selection behavior was also already partially revised by showing a mediating role of functions of music listening on the initial mood of music listening, which implied a mood-incongruent selection behavior. People would select more negative music although being in a positive mood mediated via *intellectual stimulation.* Finally, it needs to be considered that people do not select music by *perceived valence* characteristics but by expected *induced* emotional value.

### Energy

The *protocol* for the model selection of *energy* showed an improved estimated generalized performance assessed by the NCV. Additionally the final *protocol* was very stable and the cross-validated $\lambda_{\text{optimal}}$ index also turned out to be in the range of the indices that were found in the NCV, therefore the result is statistically reliable. The predictors that survived the cross-validated model selection most likely generally influence the music-selection in regards of the *energy* dimension. The ICC value of the *random intercept-only* model computed in the current work indicated that 25.1% of the total variance were explained by the grouping factor. This is somewhat different from the results of Greb et al. (2019) with an ICC for *calming-exciting* of 10%, but similar to the dimension of *less intense-very intense* with ICC = 22%. Still, all three dimensions are closely related but certainly not the same and can therefore may differ in their individual perception. In regards of Greb et al., the situational differences of functional use of music are very consistent for *energy* and *calming-exciting*. In both studies people selected less *energetic/exciting* music for *intellectual stimulation* and *mind wandering & emotional involvement.* Following the intuition, music with higher *energetic* or *exciting* characteristics was selected for *motor synchronization & enhanced well-being.* In the present study, personal differences between the functional use of *motor synchronization & enhanced well-being* were additionally found. The results showed a positive effect of the function *killing time & overcoming loneliness* with higher *energy* values. This can be related to the investigations of Randall & Rickard (2017). They found out that people would reach a higher *arousal* state by music listening coming from initial low *arousal*. This might be the reason why people select *energetic* music. Those listeners, generally using this function more often might tend to be under-stimulated or bored in general and therefore use *energetic* music to keep busy and pass time. It would seem to be plausible that the initial state of *arousal* would correspond to the same notion. But in fact, Randall & Rickard also showed convergent effects of music listening coming from high and low initial *arousal* towards a neutral *arousal* state. Thus, both effects might level out overall. The function *killing time & overcoming loneliness* might indicate effects of low initial *arousal* states.

Despite many replications of earlier studies, the inverse behavior of the dimension of *intensity* from Greb and colleagues and the feature *energy* are particularly notable. The same tendency compared to *calm-exciting* was already present in their work. Despite the fact that the musical attribute *intense* was found to highly load on the dimension *arousal* (Greenberg, et al., 2016), there still seem to be different understandings of both categories in real-life situations. In Greenberg and colleagues, participants rated mostly unknown music. *Perceived* emotion ratings are potentially less biased by the properly *induced* emotion for unknown songs, since the *intensity* of emotions is less for unfamiliar music (Ali & Peynircioğlu, 2010). This can be connected to the big factor loading of *intense* on the *arousal* dimension and little loading on *valence* and *depth* in the work of Greenberg and colleagues (2016). Therefore, the perception of *intensity* of unfamiliar music would refer to classical *arousal* components, such as higher loudness, faster tempo etc. But in Greb et al. (2019) and this work, people actively selected music to accomplish certain goals. The meaning of *intensity* ratings might differ substantially. Despite the fact, that participants were asked to rate the *perceived intensity* in the questionnaire of Greb and colleagues, those ratings might be biased by *induced* emotions due to the high degree of familiarity and emotional involvement. Thus, *intense* as music characteristic rated on familiar music might be under the influence of the intensity of emotional characteristics, such as *valence*. This part of the theory would support the results of the work of Vuoskoski & Eerola (2011). They investigated music-induced emotions and correlations to intensity ratings of emotions, indicating that *valence* components in music, such as sadness, were also related to intense music emotions. However, *energy* (Spotify) and *calming-exciting* (in Greb et al.) seem to be much more semantically congruent in this case than *energy* (Spotify) and *intensity* (in Greb et al.), since *intensity* is most likely being understood from a more engaged and emotional perspective. The influence of the initial mood (*valence*) predicting more *energetic* music intensifies the difficult relationships presented here. People selected music with higher *energetic* values when they in general tended to have a more positive initial mood than the average of participants. This might indicate connections between emotional and physiological processes that might closely relate to findings of the feature *danceability.*

### Danceability

*Danceability* was certainly less treated than *valence* and *energy/arousal* in this work so far. On first sight, this feature might seem somewhat unrelated to the high emphasize on emotional and mood aspects of music listening, but in fact dancing is strongly related to emotion expression and dance movements of individuals were shown to highly depend on the emotional state (van Dyck, Maes, Hargreaves, Lesaffre, & Leman, 2013). Additionally, dancing and music listening go often hand in hand. As the $R^2$ values show, 15.1% proportion of the total variance of the outcome variable can be explained by the fixed effects of the model and a considerable amount of the initially explained variance by the grouping factor (ICC = 15.9%) was absorbed by the model leaving only 4.5% of how much the random effects of the model explained the total variance. People certainly differed in the selection of *danceable* music and that was explained by several person-related predictors. The initial mood (*valence*) showed a tendency for situational differences and a significant effect on the personal level to be positively influencing the *danceability* of the selected music. This is largely in line with findings about the emotional expression of states of happiness by dancing. Van Dyck and colleagues (2013) investigated the effect of initially induced positive and negative emotions on the dancing of participants. They found out that people would move *faster*, more *accelerated*, more *impulsive* and more *expanded* by initial happiness compared to initial sadness. Those attributes are also considered by *Spotify* in the computation of the *danceability* value. Thus, people use music to express positive *valence* and would select music for that cause. Consequently, this dimension seems to fit the notion of positive emotion expression very well and supports a mood-congruent music-selection behavior in that sense. This is similar to Greb et al. (2019), suggesting a mediating role of the function *motor synchronization & mood enhancement* on a positive initial mood for the selection of *more rhythmic, more exciting* and *faster* music. An effect tendency was also found for gender, resulting in women selecting more *danceable* music. This can be well referred to earlier reports that indicate that women are interacting much more *intense* with emotions in both aspects, experiencing and expressing those (Kring & Gordon, 1998). Thus, *danceable* music might be much more frequently used by women to express positive emotions. Further investigations of a possible mediation of gender and the positive influence of the initial mood on the selection of more *danceable* music might be beneficial to support or discard this theory. Personality traits were also considered in the context of *danceable* music and showed an effect on different dance styles and emotion expression through dancing (Luck, Saarikallio, & Toiviainen, 2009). In this study, *extraversion* as one of the *Big Five* personality traits showed a tendency towards predicting higher values of the *Spotify* feature. This is partly in line with existing reports. Again, the results need to be closely related to the nature of the feature. Corresponding connections to the related literature might explain better what a high value of *danceability* by *Spotify* actually implies. Luck et al. (2009; 2014) observed that *extraversion* was related to people dancing with higher speed of movements. They argued that their result mirrors the notion that *extraverted* people tend to express positive emotions more strongly and are more energetic overall. This certainly underlines the idea that a higher value of the *Spotify danceability* incorporates the capacity of the song to work as a positive emotion expression tool and stimulates the act of dancing. It should not be assumed that people with different traits and taste do not select *danceable* music. Those styles might simply not match the characteristics incorporated in the *Spotify* feature. However, this work is additionally supporting the notion presented by Rentfrow and McDonald (2009). They considered people scoring high on *extraversion* having preferences for pop music and dance genres (besides other genres). In accordance, this study revealed that also the musical preference for *Techno & EDM* was a significant predictor of the *danceability* feature. So did the preference of *Volksmusik & Schlager* (German folk-music). This style is generally acknowledged to convey positive emotions and tends to be party music. That higher ratings of the musical taste for *Rock & Metal* had a negative prediction effect on the value might imply that *Rock & Metal* has less characteristics of *danceability*. Still, people might dance to it also to express emotions. So far, there is no proper reason to believe that people do not select *Rock & Metal* music to dance to, neither that this music is not danceable just because the music does not incorporate the properties reflected in this feature. The assumption finally is that people indicating preferences for *Techno & EDM* also listened to a considerable extend to this style. Characteristics of *Techno & EDM* might be incorporated highly in the *danceability* value. Thus, the selection of this feature is highly reflected by the taste and its corresponding relations to personality, gender and emotion expression goals.

Functions of music listening predicted *danceability* in ways that someone would suggest intuitively. In situations when individuals listened to music for *intellectual stimulation* people selected less *danceable* music. Also, people who tended to listen for *intellectual stimulation* reasons more than others did in average also select less *danceable* music. This is according to the relations between the features *danceability* and *energy*, both representing aspects of fast tempo perception, which seems to be less beneficial for intellectual purposes. Furthermore, it was also shown by Greb and colleagues (2018) that people selected considerable slower music for *intellectual stimulation* purposes. The reports from Greenberg and colleague (2016) pointed in a similar direction but are more controversial compared to this work. Attributes negatively loading onto the description of intellectual *depth* of music were *danceable* attributes and party music. Consequently *danceable* music was not positively related to intellectual cognitive processes. So far, this is in line with the relations towards *intellectual stimulation.* But what they additionally reported is higher preferences for *depth* in music for women than for men. This is implicitly divergent to the present results which indicated that women selected more danceable music, which was argued to convey less *depth*. Again, it could be explained by differences between listening to self-selected (familiar) music and listening to unfamiliar music. While unfamiliar music might be preferred by women if it carries intellectual properties, it might be the emotional value that is stronger preferred in the case of familiar music. Those effects should be further investigated. *Motor synchronization & enhanced well-being* (on a personal and situational level) as a function of music listening is closely related to the explanations before. Clearly, people would select more *danceable* music in the course of dancing. Furthermore, it is the only congruence to the dimension of *rhythmic* music in Greb et al. (2019). Initially it was assumed that both dimensions would be treated very similar. But apparently *rhythmical* properties, as rated subjectively by the participants, are not the main emphasis in the *danceability* feature which gives even more support to the idea that it mostly incorporates characteristics of fast and energetic music, which should be further investigated. On the personal level, two other listening functions showed positive tendencies to interact with the selection of more *danceable* music, namely *updating one's musical knowledge* and *killing time & overcoming loneliness*. Related findings can be found in Greb et al. (2019) for the dimension *intensity*. Again they showed an inverse behavior predicting *less intense* music by higher ratings of people updating their musical knowledge. This could also be explained by previously mentioned differences between the perception and selection of familiar versus unfamiliar music. People seeking for new music more than the average, would tend to use music less for emotional reasons than others, since unfamiliar music that is most likely selected in the course of knowledge updating was shown to have less emotional effects. Activating and arousing characteristics of music might be more rewarding when listening to unfamiliar music, while the emotional value is bigger for familiar music. The age effect in the present work is according to the findings of Greenberg et al. (2016). There was a tendency that older people would select less *danceable* music, although this effect did not reach statistical significance. Greenberg and colleagues found higher preferences for *depth* in music positively associated with age, while the dimension of *depth* was generally negatively associated with danceable attributes. Whether this is actually related to the different musical taste between older and younger people remains unclear. Research investigating musical preferences indeed suggest that younger people would prefer faster and more rhythmic music (Drake, Jones, & Baruch, 2000). Further investigations on the nature of the *Spotify* features are needed to reveal more specific interpretations. As the last variables, the *presence of others with* and *without communication* were positively related to the selection of more *danceable* music on a situational level (in comparison to *being alone)*. That the effect was considerable stronger for *presence of others with communication* could be explained by means of technological trends. People listen to music in any kind of situation nowadays, for example also in public transportation where emotion expression by dancing is generally less common. The results indicate that the selection of *danceable* music also follows social circumstances and the selection of more *danceable* music is also a social related action.

The *Spotify* feature for measuring the *danceability* of songs seems to have several meaningful predictors. However, it might be possible that it mostly covers aspects of arousing and positive dance music according to its relations to *extraversion* and certain musical tastes. Still there may be other types of *danceable* music that serve the expression of other emotions or are favorites to people with another personality or taste. It would be of great interest how the feature *danceability* corresponds to the human perception and how different genres are treated by that feature. It is clear that different music styles like Hip Hop, Techno, or classical dancing music like Tango, Salsa and Waltz have very different musical and acoustical properties. Thus, it would be of interest how this is handled by the *Spotify* audio feature. Comparing the results of the features *valence, energy* and *danceability* leaves many questions unanswered. Although the selection of *danceable* music might be driven by taste and personality, people indeed use this music for physical activation on a personal and situational level. Additionally it seems to involve emotional processes by means of communication and expression. People seem to react on a positive emotional state by physical activation. In this way, aspects of the selection of music with different intrinsic *valence* levels are still incorporated in this work, even if not directly through the *valence* model.

### Loudness

The model of *loudness* only included three predictors but all of them showed a significant statistical effect. The effect directions are in line with findings for *energy* and *danceability,* suggesting the close relationships. The $R^2$ and ICC values indicate high person-related differences in the selection of music on the dimension of *loudness*. 21.8% of the total variance was explained by the random effects of the model and only 8.7% were covered by fixed effects. Still, the model absorbed some of the proportion explained by the grouping factor indicated by the ICC. Similar to the *danceability* model, age also showed a negative effect on the selection of *louder* music. Thus, older people would select less *loud* music than younger people. This is in line with findings concerning musical preferences indicating that older people would listen to less intense/loud music due to hearing losses. Consequently this implies an adjusted perception of music changing with age (Smith, 1989). The age effects for *loudness* and *danceability* support recent findings from Bonneville-Roussy and Eerola (2018) indicating age trends for changing musical preferences moderated by intrinsic music characteristics. Contemporary and intense music, which is comparable to louder and more distorted music, was increasingly disliked by older people. They concluded that musical preferences can be explained by age-related perception differences of dynamics, timbre, and pitch. Thus, this notion is relevant in the context of music-selection, too. On the situational level the function of *motor synchronization & enhanced well-being* showed a positive effect on the selection of *louder* music, on the personal level the function of *killing time & overcoming loneliness* also showed a positive effect. This is in line with the models of *energy* and *danceability* and it might explain partly, that those effects rely on the *loudness perception* of music. Further investigations on features like *tempo* need to be compiled to distinguish the impact of several lower to mid-level representations on the high-level perception of *arousal* and *danceability.*

Several patterns indicated the close relationships of the three dimensions *energy, danceability* and *loudness.* Still, those features act on different complexity levels. The *loudness* feature contributes to the higher level features *energy* and *danceability*. But both also imply associations on other perceptual correlates like tempo. Additionally, *danceability* seems to feature several aspects of *positive, arousing* and *activating* music. That *valence* was not predicted reliably by any variable shows that it is inherently difficult to make generalizations of the emotional perception of unknown participants. Comparing this NCV approach that made use of a fold assignment strategy at the level of the participants with another approach that randomly splits the observations and intentionally creates leakage of participants in all folds, could investigate whether predictions can be improved by knowing about the listener. However, as the NCV indicated the results of the model selection procedures can explain differences in the music-selection behavior, also of unseen participants. The most diverse functionalities were investigated for the selection of *danceable* music related to *physiological activation*, *emotion expression* and *social associations*. Still, listening functions were prominently included in three out of four models, underlining the capability of functional variables to predict music-selection behavior on the personal and situational level.

### Further Considerations on Situational, Functional and Person-Related Variables

Activities were generally not included in any model. This is different to previous works (Greb, Steffens, & Schlotz, 2018; 2019). They found activities predicting the *sad-happy, fast-slow, peaceful-aggressive*, and *less intense-very intense* dimensions. However, activities were treated by the *glmmLasso()* function as grouped variables, with the implication that all or none of the activities would have been included in the model. This selection behavior might be too rigorous. Other investigations should consider to treat those activities individually to gain specific information about individual activities. Wang et al. (2012) and Dias et al. (2014) already investigated the effect of incorporating activities in CARS and the results indicated that they may have real-life relevance. There were no effects regarding time, neither time of day nor weekday/weekend differentiation. This is also largely in line with the findings by Greb et al. (2019). They found an effect for the *sad-happy* dimension (not statistically significant), which was not replicated in this research. It should be mentioned again, that active music listening behavior was investigated. This would not generalize to situations were music is not selected actively. For example someone could imagine that people listen to more *danceable* and *arousing* music on the weekend in the course of going to a party or club. Those situations were certainly not investigated in this work because they were not considered to be active choices of music-selection. In line with Greb et al. (2019), musical sophistication had no predictive value for any model. The initial *arousal* state has not been part of any model. Mediation analysis might explain more about the different interactions between music listening functions and person-related and situational initial *arousal* differences. Another consideration might also explain the not prominent effect of the *arousal* state on the selection of music: according to Thoma et al. (2012), individual emotion regulation styles, such as *hedonistic, distress-augmenting,* and *emotional moderation,* might modulate people's selection behavior in different situations. Therefore, complex person-situation interactions might have compensated the overall effects. Still there might be specific effects for different levels of the initial state. Notably, the *valence* of the initial mood showed overall positive person-related effects on the selection of *energetic* and *danceable* music. That personality traits were not included in most models does not necessarily give reason to believe that music is not selected according to traits. As also different investigation showed, within each factor of personality traits, different facets tend to be divergent (Greenberg, et al., 2016). Thus, the overall effects of trait factors as the *Big Five* could have been leveled out.

This work, in line with the quoted research, showed that the situational impact on the music-selection behavior was much bigger than person-related differences. ICC values indicated that in average only 22.3% of the total variances of the models were explained by the grouping factors (participants). Thus, situational differences might account for most of the rest of the proportion. It should be investigated whether different complexity levels of audio features show different behavior of explained variances. For example, there was the tendency that *loudness,* by having the lowest level of all features, was explained to the highest proportion by the grouping factor (participants). *Valence* and *danceability,* by being much more complex seemed to be much more situational dependent. It supports the need for more situational investigations of objective characteristics of music listening and selection behavior using ecologically validated assessing tools like the ESM.

## Methodological Considerations and Limitations

In the course of this work the fields of music psychology and machine learning were combined to create reliable prediction models of audio features and being able to make generalizable conclusions about situational, person-related and functional predictors of music-selection behavior. Still, all interpretations of the results need to be viewed in the context of the dataset that was used. The sample was a convenience sample mostly consisting of German students. Cultural differences may influence the effects. Also age effects might be more prominent in more general samples. However, this work has especially put methodological emphasize on assessing the quality of the performed model-selection procedures for the sake of generalizability. It resulted in a complex and computationally expensive approach but the outcome was finally very powerful and reliable. The choice of the broad *protocol* assessment procedure by *nested cross-validation* was driven by the complexity of the data structure. To the author's knowledge, there are no considerable studies on *Linear Mixed Effects Models* (LME) showing the exact effects of selection procedures like *Lasso* and different *protocol* assessment procedures but performing the NCV was an intuitive way to gain more information about the *protocols* used in this work. The *percentile-Lasso protocol* introduced by Roberts & Nowak (2014) is certainly able to perform stable selections but as the NCVs showed, the present *Mixed Effects Models* acted differently than the simulation studies that only used *Fixed Effects Models*. Their suggestion using the *$95^{th}$-percentile-Lasso* was never the best choice in terms of estimated prediction errors and model selection stability. This shows the need for further investigations on LMEs and model selection procedures. Also the influences of 5-fold or 10-fold CV on the *percentile-Lasso* is not quite clear for *Mixed Effects Models*. Splitting the data on the participant level of the clustering gave insights about the generalization of the models on unknown participants but in some cases having enough participants to create a NCV might be a critical aspect. This brings up a consideration of Krstajic et al. (2014) who proposed averaging the error rates of the tuning parameters over all *repeated cross-validations* in the inner loop. This would also allow for having more but smaller folds and still less sensitivity to the fold assignments. A corresponding *protocol* could easily be compared by repeating the current NCV technique with a different inner loop. Especially in regards of stable model selection, this *protocol* is expected to be less prone to unfavorable fold assignments and produces less outlier values of tuning parameters. A considerable drawback of the current approach is the high variance of the prediction errors due to the NCV. There are two possible adjustments that could be made: One, applying a version of *stratified cross-validation* in the outer loop, as proposed by Krstajic and colleagues (2014). This might avoid unfavorable distributions of outcome variables and the overall prediction errors might be lower and less variable, making results more comparable. Two, the approach of NCV is not the only technique to estimate the generalized prediction error of classifiers. Although it is argued to assess generalization qualities of model selection *protocols* by NCV (Cawley & Talbot, 2010), this approach is highly computationally expensive. In Ding et al. (2014) different *bias* correction techniques were compared in terms of estimation accuracy and variability of the estimated error. Similar to the results of this work, NCV returned highly variable error estimates and additionally overcorrected the *bias*. The latter was not much of a problem in this work because all *protocols* were similar, thus, the relative comparison was still given. However, further investigations should compare different bias estimation techniques especially in the spirit of LMEs, such as *Tibshirani's procedure* (TT) introduced in Tibshirani & Tibshirani (2009), *weighted mean correction* (WMC/WMCS) as proposed by Bernau, Augustin and Boulesteix (2013), and bias correction by the *inverse power law* (IPL) as presented by Ding and colleagues (2014). This would be more methodological work. More practical implications of the results that were shown here should definitely be accomplished. One way of applying the findings of this work would be the implementation of a context-aware music recommendation system.

### Towards Context-Aware Music Recommendations

This work certainly has found several statistical relationships between audio content characteristics of actively selected music and the intentional use of music depending on person-related variables and the situation. The NCV additionally indicated that the models improve the estimated generalized prediction performance. This corresponds to the notion that people perceive diverse musical properties as beneficial to accomplish certain goals. Still, *content-based music recommendation* has generally been underperforming in agreeableness and satisfaction by users. This problem is not primarily related to the incapacity of signal processing and machine learning techniques to create meaningful audio content descriptors, much more to the little attention paid to contextual and personal differences in music perception and selection so far. The music recommendation community has certainly made efforts to correspond to the obvious effects contextual parameters have on the accuracy of recommenders, but a comprehensive assessment of meaningful predictors has lacked interpretable conclusions in the past. This work argues to present a solution to overcome the "major initial issue" (Baltrunas et al., 2011, p. 90) of not knowing which contextual factors are possibly important in context-aware music recommendation systems. This might still have different implications on different recommendation techniques. It can help in *content-based* recommenders to directly predict musical descriptors by context. But it can also help implementing a *context-based filtering* in addition to *collaborative filtering*. Additionally challenging will be the practical assessment of the factual context factors but technological trends indicate that solutions are about to become more reliable. People use smartphones delivering contextual information about location and activity as provided by the *Google Awareness API*[12]*,* physiological signals measured by smartwatches can be used for increasingly reliable emotional state or physiological state assessment (Jaimovich, Coghlan, & Knapp, 2013) and social information provided by social media platforms can form a comprehensive picture of different facets of taste, activities, and trends of active users. How those factors might interact with music listening, selection, functionality etc. is still open to further investigation. Still, the theoretical models presented in this work show generalizable effects of contextual and personal variables on the music-selection behavior. This should now be tested in terms of real-life practicability and user satisfaction implemented in a proper CARS. Statistical significance may still deceive from lacking real-life significance of certain factors due to small practical effect. In the spirit of Johnson (1999), the author wants to encourage assess the model significance in the real world by real applications based on the results presented in this work.

[12] https://developers.google.com/awareness/

## Acknowledgments

This work was supported by the *Max Planck Institute for Empirical Aesthetics* based in Frankfurt am Main by supplying the dataset used in this work. It was assessed by Fabian Greb, Jochen Steffens and Wolff Schlotz (2019). I want to express especially warm gratitude to Jochen Steffens and Fabian Greb for supporting this work in all aspects from the beginning to the finalization.

# References

Ali, S. O., & Peynircioğlu, Z. F. (2010). Intensity of Emotions Conveyed and Elicited by Familiar and Unfamiliar Music. *Music Perception, 27*(3), pp. 177–182.

Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS one, 12*(3). doi:10.1371/journal.pone.0173392

Aucouturier, J.-J., & Pachet, F. (2002). Music Similarity Measures: What's The Use ? In *Proceedings of the ISMIR* (pp. 157–163).

Baltrunas, L., & Amatriain, X. (2009). Towards Time-Dependant Recommendation Based on Implicit Feedback. *Workshop on context-aware recommender systems 2009*.

Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., . . . Schwaiger, R. (2011). InCarMusic: Context-Aware Music Recommendations in a Car. In C. Huemer, & T. Setzer, *E-Commerce and Web Technologies* (Vol. 85, pp. 89–100). Berlin, Heidelberg: Springer Berlin Heidelberg.

Beal, D. J., & Weiss, H. M. (2003). Methods of Ecological Momentary Assessment in Organizational Research. *Organizational Research Methods, 6*(4), pp. 440–464.

Bernau, C., Augustin, T., & Boulesteix, A.-L. (2013). Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics, 69*(3), pp. 693–702.

Bicknell, J. (2009). Explaining Strong Emotional Responses to Music II. In *Why Music Moves Us* (pp. 89–115). London: Palgrave Macmillan UK.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion, 19*(8), pp. 1113–1139.

Bonneville-Roussy, A., & Eerola, T. (2018). Age trends in musical preferences in adulthood: 3. Perceived musical attributes as intrinsic determinants of preferences. *Musicae Scientiae, 22*(3), pp. 394–414.

Cawley, G. C., & Talbot, N. L. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research, 11*, pp. 2079–2107.

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological methods, 21*(4), pp. 603–620.

Chen, Y., Wu, C., Xie, M., & Guo, X. (2011). Solving the Sparsity Problem in Recommender Systems Using Association Retrieval. *Journal of Computers, 6*(9).

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018). Music Mood Detection Based On Audio And Lyrics With Deep Neural Net. *19th International Society for Music Information Retrieval Conference*.

Dias, R., Fonseca, M. J., & Cunha, R. (2014). A User-centered Music Recommendation Approach for Daily Activities. *CBRecSys 2014*, pp. 26–33.

Ding, Y., Tang, S., Liao, S. G., Jia, J., Oesterreich, S., Lin, Y., & Tseng, G. C. (2014). Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics (Oxford, England), 30*(22), pp. 3152–3158.

Dittmar, C., Bastuck, C., & Gruhne, M. (2007). Novel mid-level audio features for music similarity. In *International Conference on Music Communication Science* (pp. 38-41).

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology, 37*(1), pp. 295–340.

Downie, J. S. (2006). The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine, 12*(12).

Drake, C., Jones, M. R., & Baruch, C. (2000). The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. *Cognition, 77*(3), pp. 251–288.

Echonest. (2019, February 26). *The Echonest Blog*. Retrieved from http://blog.echonest.com/post/66097438564/plotting-musics-emotional-valence-1950-2013

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *ISMIR* (pp. 621–626).

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), pp. 169–200.

Falk, D. (2004). The "putting the baby down" hypothesis: Bipedalism, babbling, and baby slings. *Behavioral and Brain Sciences, 27*(04).

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior, 33*(7), pp. 865–877.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science, 18*(12), pp. 1050–1057.

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae, 5*(1_suppl), pp. 123–147.

Gabrielsson, A., & Lindström. (2001). The Influence of Musical Structure on Emotional Expression. In P. N. Juslin, & J. A. Sloboda, *Music and emotion* (Reprinted. ed., pp. 223–248). Oxford: Oxford Univ. Press.

Garrido, S., & Schubert, E. (2011). Individual Differences in the Enjoyment of Negative Emotion in Music: A Literature Review and Experiment. *Music Perception: An Interdisciplinary Journal, 28*(3), pp. 279–296.

Garrido, S., & Schubert, E. (2013). Adaptive and maladaptive attraction to negative emotions in music. *Musicae Scientiae, 17*(2), pp. 147–166.

Göker, A., & Myrhaug, H. I. (2002). User context and personalisation. In *Workshop proceedings for the 6th European Conference on Case Based Reasoning* (pp. 1-7).

Gouyon, F., Herrera Boyer, P., Gómez Gutiérrez, E., Cano, P., Bonada, J., Loscos, A., . . . Serra, X. (2008). Content processing of music audio signals. In R. D. Polotti P, *Sound to sense, sense to sound: a state of the art in sound and music computing* (pp. 83-160). Berlin: Logos Verlag.

Greasley, A. E., & Lamont, A. (2011). Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae, 15*(1), pp. 45–71.

Greb, F., Schlotz, W., & Steffens, J. (2017). Personal and situational influences on the functions of music listening. *Psychology of Music, 4*(1). doi:10.1177/0305735617724883

Greb, F., Steffens, J., & Schlotz, W. (2018). Understanding music-selection behavior via statistical learning. *Music & Science, 1*(2). doi:10.1177/2059204318755950

Greb, F., Steffens, J., & Schlotz, W. (2019). Modeling Music-Selection Behavior in Everyday Life: A Multilevel Statistical Learning Approach and Mediation Analysis of Experience Sampling Data. *Frontiers in psychology*. doi:10.3389/fpsyg.2019.00390

Greenberg, D. M., Kosinski, M., Stillwell, D. J., Monteiro, B. L., Levitin, D. J., & Rentfrow, P. J. (2016). The Song Is You: Preferences for Musical Attribute Dimensions Reflect Personality. *Social Psychological and Personality Science, 7*(6), pp. 597–605.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L 1-penalized estimation. *Statistics and Computing, 24*(2), pp. 137–154.

Hargreaves, D. J., & North, A. C. (1993). Experimental Aesthetics and Liking for Music. In P. N. Juslin, *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 515–546). Oxford University Press.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological methods, 21*(4), pp. 447–457.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning* (Second edition, corrected at 12th printing 2017 ed.). New York, NY: Springer.

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX Audio Mood Classification Task: Lessons Learned.

Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics.

Huq, A., Bello, J. P., & Rowe, R. (2010). Automated Music Emotion Recognition: A Systematic Evaluation. *Journal of New Music Research, 39*(3), pp. 227–244.

Huron, D. (2001). Is Music an Evolutionary Adaptation? *Annals of the New York Academy of Sciences, 930*(1), pp. 43–61.

Husain, G., Thompson, W. F., & Schellenberg, E. G. (2002). Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. *Music Perception, 20*(2), pp. 151–171.

Jaimovich, J., Coghlan, N., & Knapp, R. B. (2013). Emotion in Motion: A Study of Music and Affective Response. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, . . . S. Ystad, *Computer Music Modeling and Retrieval* (Vol. 7900, pp. 19–43). Berlin: Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The journal of wildlife management*, pp. 763–772.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, pp. 78–89.

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. *Physics of life reviews, 10*(3), pp. 235–266.

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: listener, music, and situation. *Emotion (Washington, D.C.), 8*(5), pp. 668–683.

Kaur, L., & Kumari, N. (2017). A Review on User Recommendation System Based Upon Semantic Analysis. *International Journal of Advanced Research in Computer Science and Software Engineering, 7*(11), pp. 35-43.

Kawakami, A., Furukawa, K., Katahira, K., & Okanoya, K. (2013). Sad music induces pleasant emotion. *Frontiers in psychology, 4*. doi:10.3389/fpsyg.2013.00311

Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., . . . Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the ISMIR* (pp. 255–266).

Krause, A. E., & North, A. C. (2017). Pleasure, arousal, dominance, and judgments about music in everyday life. *Psychology of Music, 45*(3), pp. 355–374.

Krause, A. E., North, A. C., & Hewitt, L. Y. (2016). The role of location in everyday experiences of music. *Psychology of Popular Media Culture, 5*(3), pp. 232–257.

Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: expression, experience, and physiology. *Journal of Personality and Social Psychology, 74*(3), pp. 686–703.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics, 6*(1). doi:10.1186/1758-2946-6-10

Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2013). *Applied linear statistical models* (5 ed.). Boston, Mass.: McGraw-Hill.

Lee, J. S., & Lee, J. C. (2007). Context Awareness by Case-Based Reasoning in a Music Recommendation System. In H. Ichikawa, W.-D. Cho, I. Satoh, & H. Y. Youn, *Ubiquitous Computing Systems* (Vol. 4836, pp. 45–58). Berlin, Heidelberg: Springer Berlin Heidelberg.

Li, Q., Kim, B. M., Guan, D. H., & Oh, D. W. (2004). A music recommender based on audio features. In *Proceedings of the 27th Annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 532-533). ACM Press. doi:10.1145/1008992.1009106

Liu, N.-H., Lai, S.-W., Chen, C.-Y., & Hsieh, S.-J. (2009). Adaptive music recommendation based on user behavior in time slot. *International Journal of Computer Science and Network Security, 9*(2), pp. 219–227.

Longhi, E. (2008). Emotional responses in mother-infant musical interactions: A developmental perspective. *Behavioral and Brain Sciences, 31*(05), pp. 559-575.

Luck, G., Saarikallio, S., & Toiviainen, P. (2009). Personality traits correlate with characteristics of music-induced movement. In *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009).*

Luck, G., Saarikallio, S., Burger, B., Thompson, M., & Toiviainen, P. (2014). Emotion-driven encoding of music preference and personality in dance. *Musicae Scientiae, 18*(3), pp. 307–323.

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG-7.* Chichester; New York: Wiley.

Mehrabian, A., & Russell, J. (1974). *An Approach to Environment Psychology.* Cambridge, MA: MIT press.

Mithen, S. (2005). *The Singing Neanderthals.* Oxford: Oxbow Books.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society, Interface, 14*(134).

North, A. C., & Hargreaves, D. J. (1996). Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition, 15*(1-2), pp. 30–45.

North, A. C., & Hargreaves, D. J. (2000). Musical Preferences during and after Relaxation and Exercise. *The American Journal of Psychology, 113*(1), pp. 43-68.

North, A. C., Hargreaves, D. J., & Hargreaves, J. J. (2004). Uses of Music in Everyday Life. *Music Perception, 22*(1), pp. 41–77.

Oliver, N., & Kreger-Stickles, L. (2006). PAPA: Physiology and Purpose-Aware Automatic Playlist Generation.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251).

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project.

Peltola, H.-R., & Eerola, T. (2016). Fifty shades of blue: Classification of music-evoked sadness. *Musicae Scientiae, 20*(1), pp. 84–102.

Randall, W. M., & Rickard, N. S. (2017). Personal Music Listening. *Music Perception: An Interdisciplinary Journal, 34*(5), pp. 501–514.

Reddy, S., & Mascia, J. (2006). Lifetrak: music in tune with your life. In *Proceedings of the 1st ACM international workshop on Human-centered multimedia* (pp. 25–34). ACM.

Rentfrow, P. J., & McDonald, J. A. (2009). Music preferences and personality. In P. N. Juslin, & J. A. Sloboda, *Handbook of music and emotion* (pp. 669–695). Oxford, England: Oxford University Press.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender Systems Handbook.* Boston, MA: Springer US.

Roberts, S., & Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis, 70*, pp. 198–211.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), pp. 1161–1178.

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe. *Musicae Scientiae, 18*(4), pp. 423–447.

Schäfer, T. (2016). The Goals and Effects of Music Listening and Their Relationship to the Strength of Music Preference. *PloS one, 11*(3). doi:10.1371/journal.pone.0151634

Schäfer, T., & Sedlmeier, P. (2009). From the functions of music to music preference. *Psychology of Music, 37*(3), pp. 279–300.

Schäfer, T., Sedlmeier, P., Städtler, C., & Huron, D. (2013). The psychological functions of music listening. *Frontiers in psychology, 4*. doi:10.3389/fpsyg.2013.0051

Schedl, M., & Knees, P. (2009). Context-based Music Similarity Estimation. In *Welcome to the 3 rd International Workshop on Learning Semantics of Audio Signals.* Graz, Austria.

Schedl, M., Flexer, A., & Urbano, J. (2013). The Neglected User in Music Information Retrieval Research. *Journal of Intelligent Information Systems, 41*(3), pp. 523–539.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval, 8,2/3*, pp. 127-261. doi:10.1561/1500000042

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion, 2*(4), pp. 412–417.

Schubert, E. (2009). The fundamental function of music. *Musicae Scientiae, 13*(2_suppl), pp. 63–81. doi:10.1177/1029864909013002051

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), pp. 1359–1366.

Skånland, M. S. (2013). Everyday music listening and affect regulation: the role of MP3 players. *International journal of qualitative studies on health and well-being, 8*. doi:10.3402/qhw.v8i0.20595

Sloboda, J. A., O'Neill, S. A., & Ivaldi, A. (2001). Functions of Music in Everyday Life: An Exploratory Study Using the Experience Sampling Method. *Musicae Scientiae, 5*(1), pp. 9–32.

Smith, D. S. (1989). Preferences for differentiated frequency loudness levels in older adult music listening. *Journal of Music Therapy, 26*(1), pp. 18–29.

Song, Y., & Dixon, S. (2015). How well can a music emotion recognition system predict the emotional responses of participants? In *Sound and Music Computing Conference (SMC)* (pp. 387-392).

Song, Y., Dixon, S., & Pearce, M. (2012). A Survey of Music Recommendation Systems and Future Perspectives. In *9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012).*

Song, Y., Dixon, S., Pearce, M. T., & Halpern, A. R. (2016). Perceived and Induced Emotion Responses to Popular Music: Categorical and Dimensional Models. *Music Perception, 33*(4), pp. 472–492.

Spotify. (2019, February 26). *Spotify Web API*. Retrieved from https://developer.spotify.com/documentation/web-api/reference/tracks/get-several-audio-features/)

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), pp. 111–147.

Su, J.-H., Yeh, H.-H., Philip, S. Y., & Tseng, V. S. (2010). Music recommendation using content and context information mining. *IEEE Intelligent Systems, 25*(1), pp. 16–26.

Thayer, R. E. (1990). *The biopsychology of mood and arousal.* Oxford: Oxford University Press.

Thoma, M. V., Ryf, S., Mohiyeddini, C., Ehlert, U., & Nater, U. M. (2012). Emotion regulation through listening to music in everyday situations. *Cognition & Emotion, 26*(3), pp. 550–560.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics, 3*(2), pp. 822–829.

van Dyck, E., Maes, P.-J., Hargreaves, J., Lesaffre, M., & Leman, M. (2013). Expressing Induced Emotions Through Free Dance Movement. *Journal of Nonverbal Behavior, 37*(3), pp. 175–190.

Vuoskoski, J. K., & Eerola, T. (2011). Measuring music-induced emotion. *Musicae Scientiae, 15*(2), pp. 159–173.

Vuoskoski, J. K., & Thompson, W. F. (2012). Who Enjoys Listening to Sad Music and Why? *Music Perception: An Interdisciplinary Journal, 29*(3), pp. 311–317.

Wang, X., Rosenblum, D., & Wang, Y. (2012). Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 99-108). New York, New York, USA: ACM Press.

Weigl, D., & Guastavino, C. (2011). User Studies in the Music Information Retrieval Literature. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (pp. 335-340).

Yang, Y.-H., & Chen, H. H. (2011). *Music emotion recognition.* Boca Raton, FL: CRC Press.

Yang, Y.-H., & Chen, H. H. (2012). Machine Recognition of Music Emotion. *ACM Transactions on Intelligent Systems and Technology, 3*(3), pp. 1–30.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on psychological science: A journal of the Association for Psychological Science, 12*(6), pp. 1100–1122.

Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2008). An Efficient Hybrid Music Recommender System Using an Incrementally Trainable Probabilistic Generative Model. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(2), pp. 435–447.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion (Washington, D.C.), 8*(4), pp. 494–521.

Zheng, Y. (2017). Interpreting Contextual Effects By Contextual Modeling In Recommender Systems. *CoRR*, pp. 1-7.

# Appendix

This is the questionnaire used by Greb and colleagues (2019) to assess the music-selection behavior of participants in an *Experience Sampling Method.* The questionnaire was originally presented in German. The questionnaire included items that were not evaluated in this work but maintained below.

1) Do you currently listen to music?

- Yes, I currently listen to music.
- No, I currently do not listen to music.

### Section 1 [Situation]

2) For how long have you been listening to music already? Please indicate the duration in minutes: Free response

3) Please choose your current main activity [Activity].

- Pure music listening
- Housework
- Working/studying
- Coping with emotions
- Exercise
- Social activity (e.g. eating or playing with friends)
- Party
- Making music
- Relaxing/falling asleep
- Being on the move (bus/train/car)
- Personal hygiene
- Other (none of the activity listed is appropriate)

4) Are there currently any other persons present? [Presence of others]

- No, I am alone.
- Yes, I am surrounded by others but do not interact or communicate with them.
- Yes, I interact/communicate with other people.

5) Did you choose the music? [Choice]

- Yes
- No
- Radio
- Club
- Concert
- Playlist

6) How much control do you have in what you hear? [Control]

| | | |
|---|---|---|
| Any control | 1-2-3-4-5-6-7 | Full control |

7) How was your mood at the moment you decided to listen to music? [Valence]

| | | |
|---|---|---|
| Bad | 1-2-3-4-5-6-7 | Good |

8) How awake did you feel at the moment you decided to listen to music? [Arousal]

| | | |
|---|---|---|
| Tired | 1-2-3-4-5-6-7 | Awake |

9) How important was your mood for your decision to listen to music? [Importance of mood]

| | | |
|---|---|---|
| Not at all | 1-2-3-4-5-6-7 | Very important |

10) How much attention are you paying to the music? [Attention]

| | | |
|---|---|---|
| Little | 1-2-3-4-5-6-7 | A lot |

## Section 2 [Music]

11) How loud is the music?

| | | |
|---|---|---|
| Quiet | 1-2-3-4-5-6-7 | Loud |

12) How much do you like the music?

| | | |
|---|---|---|
| I like it less | 1-2-3-4-5-6-7 | I like it a lot |

13) Please name the composer/artist if known: Free response

14) Please name the title of the piece if known: Free response

15) Please name the musical style if known: Free response

16) Which characteristics does the music have? [Musical characteristics]

| | | |
|---|---|---|
| Calming | 1-2-3-4-5-6-7 | Exciting |
| Slow | 1-2-3-4-5-6-7 | Fast |
| Sad | 1-2-3-4-5-6-7 | Happy |
| Unfamiliar | 1-2-3-4-5-6-7 | Familiar |
| Less melodic | 1-2-3-4-5-6-7 | Very melodic |
| Less rhythmic | 1-2-3-4-5-6-7 | Very rhythmic |
| Simple | 1-2-3-4-5-6-7 | Complex |
| Peaceful | 1-2-3-4-5-6-7 | Aggressive |
| Less intense | 1-2-3-4-5-6-7 | Very intense |
| Instrumental | 0-1 | Vocal |

## Section 3 [Functions of music listening]

17) Why do you currently listen to music? [Functions of music listening]

… because it gives me intellectual stimulation. (I)

| | | |
|---|---|---|
| Not at all | 1-2-3-4-5-6-7 | Fully agree |

… because it mirrors my feeling and moods. (II)

| | | |
|---|---|---|
| Not at all | 1-2-3-4-5-6-7 | Fully agree |

… because it makes me feel fitter. (III)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it addresses my sense of aesthetics. (I)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it puts fantastic images or stories in my head. (II)
Not at all 1-2-3-4-5-6-7 Fully agree

… because I can learn about new pieces. (IV)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it enables me to kill time. (V)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it helps me learn about myself. (I)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it reminds me of certain periods of my life or past experiences. (II)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it makes me feel connected to all people who like the same kind of music. (IV)
Not at all 1-2-3-4-5-6-7 Fully agree

… because I can move to the music. (III)
Not at all 1-2-3-4-5-6-7 Fully agree

… because I need it in the background while I do other things. (V)
Not at all 1-2-3-4-5-6-7 Fully agree

… because I want to inform myself about hits and trends. (IV)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it enhances my mood. (III)
Not at all 1-2-3-4-5-6-7 Fully agree

… because it makes me feel less lonely. (V)
Not at all 1-2-3-4-5-6-7 Fully agree

… because I do it out of habit (not included)
Not at all 1-2-3-4-5-6-7 Fully agree

Note: roman letters indicate the corresponding music listening function, they were not shown to participants.

I – Intellectual Stimulation

II – Mind Wandering & Emotional Involvement

III – Motor Synchronization & Enhanced Well-being

IV – Updating One’s Musical Knowledge

V – Killing Time & Overcoming Loneliness