Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features

Elke B. Lange
Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

Klaus Frieler
University of Music Franz Liszt, Weimar, Germany

Music information retrieval (MIR) is a fast-growing research area. One of its aims is to extract musical characteristics from audio. In this study, we assumed the roles of researchers without further technical MIR experience and set out to test in an exploratory way its opportunities and challenges in the specific context of musical emotion perception. Twenty sound engineers rated 60 musical excerpts from a broad range of styles with respect to 22 spectral, musical, and cross-modal features (*perceptual features*) and perceived emotional expression. In addition, we extracted 86 features (*acoustic features*) of the excerpts with the MIRtoolbox (Lartillot & Toiviainen, 2007). First, we evaluated the perceptual and extracted acoustic features. Both perceptual and acoustic features posed statistical challenges (e.g., perceptual features were often bimodally distributed, and acoustic features highly correlated). Second, we tested the suitability of the acoustic features for modeling perceived emotional content. Four nearly disjunctive feature sets provided similar results, implying a certain arbitrariness of feature selection. We compared the predictive power of perceptual and acoustic features using linear mixed effects models, but the results were inconclusive. We discuss critical points and make suggestions to further evaluate MIR tools for modeling music perception and processing.

Received: April 7, 2017, accepted June 12, 2018.

Key words: music information retrieval (MIR), feature extraction, emotion expression, expert ratings, predictive modeling

One of the most impressive and motivating effects of listening to music is the induction and communication of emotion. The question of how music can affect listeners in this way is still under debate (e.g., Cespedes-Guevara & Eerola, 2018; Juslin, 2013; Scherer, 2004). With regard to emotional expression, the basic idea is that emotional meaning is conveyed by means of cues in the auditory information. Composers and musicians construct emotional meaning using such cues and listeners in turn interpret these cues (Balkwill & Thompson, 1999; Eerola, Friberg, & Bresin, 2013; Juslin, 2000; Juslin & Lindström, 2011; Kreutz & Lotze, 2007; Leman, Vermeulen, De Voogdt, Moelants, & Lesaffre, 2005; Peretz, Gagnon, & Bouchard, 1998; Scherer & Oshinsky, 1977).

At least two different approaches have been utilized to investigate whether specific musical features cue emotional expression or affect (see Gabrielsson & Lindström, 2010, for a general overview of methods to investigate emotions in music). In one approach, judgments on musical features are collected and related to judgments of perceived or induced emotions by regression models (Eerola et al., 2013; Friberg, Schoonderwaldt, Hedblad, Fabiani, & Elowsson, 2014; Korhonen, Clausi, & Jernigan, 2006; Leman et al., 2005; Yang, Lin, Su, & Chen, 2008). The other approach is experimentally based on a sophisticated manipulation of the musical features of musical excerpts. Listeners in turn judge the emotional expression and conclusions are drawn based on a factorial design (Balkwill & Thompson, 1999; Bowling, Sundararajan, Han, & Purves, 2012; Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Eerola et al., 2013; Hevner, 1935, 1937; Ilie & Thompson, 2006; Juslin, Friberg & Bresin, 2002; Juslin & Lindström, 2011; Peretz et al., 1998; Schellenberg, Krysciak, & Campbell, 2000; Scherer & Oshinsky, 1977; see also Bresin & Friberg, 2011, on using the production method to select feature characteristics for specific emotion expressions). The two approaches both have advantages and disadvantages. For instance, whereas experimental designs can be controlled for very precisely and the inferences drawn are generally more reliable, the stimuli chosen are often rather simplistic (e.g., a short sequence of synthesized tones). In addition, manipulating one individual feature like mode or tempo out of the musical context is rather artificial and restricts the ecological validity. On the other hand, when real music is used to investigate musical characteristics, the selection of music is subjective and often based on the experimenter's

Music Perception, volume 36, issue 2, pp. 217–242, issn 0730-7829, electronic issn 1533-8312. © 2018 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, http://www.ucpress.edu/journals.php?p=reprints. DOI: https://doi.org/10.1525/mp.2018.36.2.217

evaluation of the music's attributes. To gain more objectivity, evaluations can be collected from a large subject sample in a pilot study (e.g., Eerola & Vuoskoski, 2011). Subjective evaluations and ratings of musical attributes have some methodological disadvantages. One is the precision of measurements (e.g., the retest reliability can sometimes be low, e.g., see Schedl, Eghbal-Zadeh, Gómec, & Tkalčič, 2016; Yang et al., 2008, for retest reliability of valence). Another is that collecting these ratings for a large data set of music is very time-consuming. However, despite the differences in methodologies, there is a fair amount of agreement: studies converge on the finding that—at least in the context of Western tonal music—perceived tempo,[1] mode, dynamics, and mean pitch are the main contributors to predicting musical emotions (e.g., Hevner, 1935, 1937; Scherer & Oshinsky, 1977; see Gabrielsson & Lindström, 2001 or 2010, for an overview).

In this situation, developments in automatized music information retrieval (MIR) seem promising, offering a third approach to relate acoustic cues to emotion perception. Instead of collecting perceptual judgments on musical features, acoustic cues are extracted directly from the audio signal and related to emotion perception (e.g., Gingras, Marin, & Fitch, 2014; Schubert, 2004; Juslin, 2000). For instance, computer algorithms produce musical features as output, with a simple WAV file as input. Musical features can be analyzed at different levels, with features like loudness and pitch at a low level, complexity of harmonic progression at a medium, and intended emotional content at a high level (e.g., Leman et al., 2005; but see Lartillot, 2014, for a slightly different differentiation of levels). Note, that these levels constitute a very rough classification system and—with regard to perception—cannot be uniquely defined by physical or psychological criteria. However, for automatized feature extraction, the number of assumptions implemented in the algorithms typically accumulates with an increasing level. Such assumptions pertain both to the designers of an algorithm and to the developers who are in charge of implementation. The consequence is a larger or smaller factor of subjectivity for selection of assumptions and weighting. For instance, extraction of high-level features is highly dependent on the cultural context (Balkwill & Thompson, 1999). Pushing analysis from acoustics to human perception is the challenge developers of algorithms are facing.

In the last decades, quite some progress has been made by the MIR community in steadily improving the reliability and validity of models, as documented by the MIREX (Music Information Retrieval Evaluation eXchange) competitions, which started in 2005. For instance, accuracy for detection of the rather difficult feature of tempo improved from a range of 71–95% in a test set of audio samples (at least one of two octave-related tempi correct) in 2005 to 99.3% in 2017 (see http://www.music-ir.org/mirex/wiki/MIREX_HOME; February 21, 2018). On the other hand, predicting a complex feature like emotional expression is less reliable. The highest accuracy in the MIREX contest was reached in 2011 with 69.5%. However, due to inevitable variations in human perception it is questionable whether an accuracy of 100% can be attained or whether the upper limit has to be lowered (Friberg et al., 2014; Saari, Eerola, & Lartillot, 2011).

In the MIR community, huge differences are reported with respect to the emotion classification task, e.g., predicting basic emotions or arousal and valence, and the mathematical method used. Whereas modeling arousal by automatically extracted features works quite well (e.g., Gingras et al., 2014; Schubert, 2004), valence can be much less reliably predicted using given features. For instance, Yang et al. (2008) demonstrated that extracted musical features could predict the rated valence of musical excerpts with $R^2 = 28.1\%$ using regression, and Korhonen et al. (2006) achieved $R^2 = 21.9\%$ using a system identification method. The success was higher applying (non-linear) artificial neural network models and highly sophisticated algorithms for feature extraction: with only five features perceived valence (about 43% explained variance in Coutinho & Dibben, 2013, Table 6, test set) or induced valence (about 50% explained variance in Coutinho & Cangelosi, 2011, Table 6, simulation 1) could be satisfyingly predicted. The best results were achieved with classifier methods that assign music to specific emotion categories (e.g., happy, sad, peaceful, and angry). Here, the mean accuracy of classification by automatically extracted features was about 85% (Hwang et al., 2013) and 86.4% (Lu, Liu, & Zhang, 2006). Interestingly, extracted melodic features such as contour and vibrato can considerably increase model accuracy (using support vector machines in Panda, Rocha, & Paiva, 2015), whereas adding psychophysiological measurements of listeners, such as heart rate, etc., only slightly increases model accuracy (artificial neural networks in Coutinho & Cangelosi, 2011).

Our study had two main objectives. First, we carefully evaluated the methods of determining music characteristics by subjective judgments as well as automatized

[1] We will use the term *tempo* when beats per time unit are measured and *speed* for the perceptive evaluation of this tempo throughout the paper to take into account the distinction between the terms (e.g., see Elowsson & Friberg, 2015).

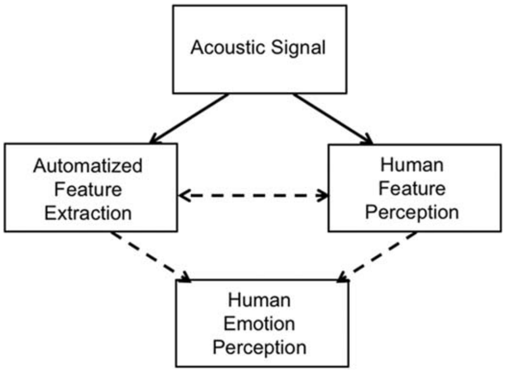

FIGURE 1. Levels of sound description: In our study, the musical excerpts were analyzed as acoustic signal by automatized feature extraction as well as perceptually evaluated by sound engineers. We then asked for the relation between human and automatized feature extraction and their relation to emotion perception (rated by the same sound engineers).

feature extraction and selection. Second, we were interested in comparing the predictive power of automatically extracted and humanly perceived features for emotion perception. Figure 1 shows the different perspectives and their relations. Many studies are in line with this research, showing dependencies between perceived features and emotion perception (e.g., Friberg et al., 2014; Leman et al., 2005), extracted features and emotion perception (e.g., Gingras et al., 2014; Korhonen et al., 2006; Schubert, 2004; Yang et al., 2008), and extracted and perceived musical features (e.g., Friberg et al., 2014). There are only a few studies relating automatically extracted features to both musical feature perception and perceived emotional expression (Friberg et al., 2014; Leman et al., 2005), and one of these did not report on the critical comparison of predictive power of extracted and perceived features, which is the subject we are interested in (Friberg et al., 2014, but these authors had a conceptual model that excludes such a comparison). Leman et al. (2005) followed an approach very similar to ours, utilizing 60 musical excerpts from a broad range of musical styles. First, students (Study I) rated 30-second excerpts using bipolar adjectives that could be clustered into the two dimensions of valence and arousal. Second, 25 musicologists (Study II) rated those excerpts on acoustic features such as pitch, tempo, etc. Third, the excerpts were analyzed via MIR using an algorithm described in the publication. Fourth, regression analysis attempted to predict affective evaluations by either rated or automatically extracted acoustic features (Study III). One limitation of this regression analysis was that only eight students rated all excerpts (100 students rated the excerpts in a distributed method in Study I) and individuals varied in how well their evaluations could be accounted for. However, some of the perceived features predicted emotional ratings impressively well: the consonance-dissonance ratings of the musicologists predicted the valence dimension in five out of eight students; loudness predicted the activity dimension for all students. In addition, extracted features showed weaker effects in predicting the valence dimension. Unfortunately, when building participant-specific models with a subset of musical excerpts to predict the outcome for the remaining excerpts, the results were convincing for only a small number of participants. Nevertheless, the results of Leman et al. are promising and encouraged us to pursue this issue further. In contrast to Leman, we decided on a complete within-subject design in which twenty sound engineers evaluated the music excerpts' acoustic and musical properties as well as emotional expression. It was important to us to invite professional sound engineers because it was assumed they would at least converge on ratings of the low-level features, as this is what they have been trained in.

We also included one additional aspect. Communication about music is characterized by non-musical descriptions. Attributes like bright–dark (for timbre) or high–low (for pitch) are not inherently musical (though partly onomatopoetic) but visual-spatial features. In the recent past, research has empirically demonstrated that music is associated with non-musical attributes from different modalities such as gustatory perception (Bronner, Frieler, Bruhn, Hirt, & Piper, 2012; Reinoso Carvalho et al., 2015; Wang, Woods, & Spence, 2015). We therefore implemented in our study a few modal features from the tactile and visual domain such as temperature, roughness, and brightness in order to evaluate in an exploratory manner the predictive power of those modal perceptions for musical emotion perception.

In our study, we decided to use the MIRtoolbox (Lartillot & Toiviainen, 2007) for feature extraction for several reasons. It is a MATLAB-based toolbox that is very easy to handle, and it offers a great variety of features within one programming environment (Moffat, Ronan, & Reiss, 2015; see Yang et al., 2008, and Korhonen et al., 2006, on using at least two different programs—PsySound for low level and Marsyas for medium-level features—and Leman et al., 2005, on developing their own mathematical model to extract features). In one comparison, MIRtoolbox and Marsyas seem to outperform PsySound (Panda et al., 2015); this was attributed in part to their larger and more differentiated feature sets. Importantly, the MIRtoolbox is one of the most prominent and widely used programs to extract musical features; the relevant publication, Lartillot & Toiviainen (2007), has been cited more than 800 times (Google Scholar, September 2017). For example, in a publication in the field of neuroscience, Alluri et al. (2012) extracted low and mid-level features that were correlated with brain activations. The authors were able to identify distinct areas for the timbral, tonal, and rhythmic features, supporting the hypothesis that these features might contribute differently to the processing of emotional expression in music. In a related publication (Poikonen et al., 2016), the temporal evolution of features was analyzed and peak changes detected. Those peak changes were related to a specific signature in the brain using event-related potentials (ERP) as a method. Investigations of these kinds show the far-reaching consequences of MIR and the high potential for music psychology. In addition, researchers without in-depth MIR experience are most likely to reach for the most well-known, easily available, and easy to use toolbox, which at present is the MIRtoolbox. We therefore also decided to narrow the algorithms applied to MIRtoolbox. However, to our knowledge, at present only a few dependent investigations and one independent investigation have validated that the features acoustically extracted by MIRtoolbox had some psychological relevance for musical feature perception (Alluri & Toiviainen, 2010; Friberg et al., 2014; Poikonen et al., 2016) and emotional expression in music (Eerola, Lartillot, & Toiviainen, 2009; Friberg et al., 2014; Gingras et al., 2014).

More specifically, we applied functions of MIRtoolbox to a set of 60 musical excerpts from a broad range of mainly Western musical styles (e.g., metal, pop, classical music, techno, rap, etc.). Research on musical emotions has been predominantly applied to classical music (e.g., Bigand, Vieillard, Madurell, Marizeau, & Dacquet, 2005; Coutinho & Cangelosi, 2011; Korhonen et al., 2006; but see Leman et al., 2005; Yang et al., 2008). Given that classical music listeners are only a small subset of all music listeners, it was important to us to set up a corpus of music with a greater variety of musical styles, at the same time spanning the emotion space from low to high arousal and positive to negative valence.

Method

PARTICIPANTS

We recruited 20 professional sound engineers, aged between 20 and 60 (median = 40), six of whom were female, twelve male, and two who did not choose to specify their sex. Seven of them were students of sound engineering, twelve were full-time employed sound engineers, and one worked as a freelancer. Their professional experience ranged from 1 to 45 years, with a mean of 15.1 and median of 15.5 years. Two participants reported studying or working at present in other non-specified, non-musical domains. None of the participants reported hearing problems. All experimental procedures were ethically approved by the Ethics Council of the Max Planck Society, and were undertaken with each participant's written informed consent. Participants largely classified as musically sophisticated, with a median of 96.5 in the Goldsmith Music Sophistication Index subscale (Gold-MSI, Müllensiefen, Gringas, Musil, & Stewart, 2014; range = 71 to 118); a questionnaire that included 14 musical styles (see Musical Stimuli section) showed their broad range of musical preferences.

APPARATUS

The study was run in a group testing room with parallel data collection of one to four participants. The hardware between testing environments was matched. A Windows PC ran the procedure programmed in PsychoPy 1.82.01 (Peirce, 2007). Musical stimuli were presented via Beyerdynamic headphones (DT 770 Pro 80 Ohm).

MUSICAL STIMULI

Music samples were taken from a collection originally selected for another experiment in which participants had the task of immersing themselves in music (see Lange, Zweck, & Sinn, 2017). That pool of 56 excerpts spanned a broad range of Western musical styles: blues, country, electronica, folk, hip hop, classical music, jazz, metal, pop, reggae, rock, soul, traditional German folk (Volksmusik), together with world music. For our final sample, we added four more classical music samples that had been used in a study by Bigand et al. (2005). None of the excerpts had lyrics. The total stimuli selection covered a broad emotional space, ranging from high to low arousal as well as from positive to negative valence. The excerpts lasted 43–61 s, depending on where the phrase ended within the music. The digital WAV files had a sample rate of 44.100 Hz, 16 bits, and loudness was adjusted via the r128gain software (Belkner, r128gain.sourceforge.net; 17.02.2015), which applies a normalization algorithm taking perceptual loudness differences as well as amplitude differences into account. It is based on the standard procedure of the European Broadcast Union (EBU Technical Committee, 2011) to match the perceived loudness of audio recordings presented subsequently. After normalization of the loudness, it was necessary to adjust some of the excerpts again (e.g., a slow and soft piano piece from Satie would sound exceptionally loud when brought up to the same level as a rock song). We amplified 28 excerpts by -/+3 to 5 dB. For a complete list of music selection and editing, see Appendix. The final volume was self-chosen based on individual preferences and remained the same throughout the data collection, except in the case of two participants who wanted to reduce the volume after the first half of the study. Note that these between-participant differences in loudness will contribute to variance, which can be accounted for in the final linear mixed effect model account.

SELECTION AND EXTRACTION OF MIR FEATURES

With the MIRtoolbox we extracted a large battery of features. Table 1 gives an overview of the functions and a simplified verbal description of the content of the toolbox. Table 2 provides a list of all extracted features. These features, or a subset thereof, are commonly used in music research (Alluri & Toiviainen, 2010; Coutinho & Cangelosi, 2011; Coutinho & Dibben, 2013; Eerola, 2011; Friberg et al., 2014; Leman et al., 2005), neuroscience (Alluri et al., 2012; Poikonen et al., 2016), and computer science (Casey et al., 2008; Eerola et al., 2009; Hwang et al., 2013; Koelstra, et al., 2012; Li & Ogihara, 2006; Mion & De Poli, 2008; Panda et al., 2015; Tzanetakis & Cook, 2002; Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013; Yang et al., 2008). It should be pointed out that most of these are low-level features when defined in terms of the audio waveform (e.g., *mirrms, mirzerocross*) or the spectrum (e.g., *mircentroid, mirmfcc*), but some are based on a low-level perceptual model (e.g., *mirroughness, mirpitch*), and some are based on more cognitive models (e.g., *mirkeyclarity, mirtempo*).

For the automatized feature extraction, we used windowing for our audio samples (50 ms for low-level, 2–3 s for medium-level audio features, relying mostly on default values to mimic the naïve approach of a non-expert user) with 50% overlap. For most feature types, we used the arithmetic mean and the sample standard deviation of the sequence of windowed features. Using averages of windowed features is a common approach, based on modeling results that show no improvement when dynamic changes are captured instead for classification (for a discussion in the context of timbre features, see Alluri & Toiviainen, 2010). However, one has to keep in mind that this procedure already presupposes a similar averaging process in the mind of a listener. For instance, the peak-end rule shows that other mechanisms than averaging might be in effect (Kahnemann, 2000).

QUESTIONNAIRE ON MUSIC PERCEPTION

The questionnaire consisted of 22 items split into four parts. They were presented electronically on a PC screen, one page for one part. On the first page, six basic low level features were rated: pitch from low to high (*pitch_h*), modality from minor to major (*tonal_maj*), speed from slow to fast (*speed_sl*), articulation from accentuated/staccato to fluent/legato (*artic_flow*), loudness from loud to soft (*loud_soft*), and consonance from consonant to dissonant (*con_diss*). The second page asked for variability in the temporal evolution of harmony (*harm_vari*), melody (*mel_vari*), rhythm (*rhyth_vari*), and dynamics/intensity (*dyn_vari*). On the third page the sound was to be described using a brief semantic differential: dark to bright (*dark_bright*), rough to smooth (*rough_smooth*), homophonic to polyphonic (*homo_poly*), cold to warm (*cold_warm*), narrow to wide spectrum (*wideSp*), and dense to sparse spectrum (*sparseSp*). See Table 3 for assumed analogies between subjective evaluation and feature extraction. Finally, on the last page the task was to rate the emotional expression of the sounds from *not at all* to *very much*: happy (*happy*), sad (*sad*), tender (*tender*), fearful (*fearf*), peaceful (*peacef*), and angry (*angry*).

To render the output comparable, all items were evaluated on a seven-point rating scale. The musical features were described on bipolar scales, the emotional ones on unipolar scales. The seven bullet points for the ratings were ordered horizontally from left to right, labeled with digits from one to seven, with the verbal labels located above. Upon mouse click, the color changed from gray (which also served as the background color of the screen) to red. Participants were free to choose the order in which they answered for each page of the questionnaire. Changing back to a prior page was not possible. In the written instructions, some of the items that might be ambiguous even for sound engineers were explained by examples (e.g., for the dimension rough–smooth). We explicitly asked the participants to attend to the noisiness of the signal because roughness might also have been interpreted in terms of dissonance.

We selected those items for the questionnaire based on a careful reading of the literature. The basic features of the first page, some of the dynamic features of the second page, as well as the emotional categories are commonly used in studies (e.g., Balkwill & Thompson, 1999; Bresin & Friberg, 2011; Dalla Bella et al., 2001; Eerola et al., 2013; Illie & Thompson, 2006; Juslin, 1997; Juslin & Lindström, 2011; Leman et al., 2005; Scherer & Oshinsky, 1977). Of those studies, the following features are most prominent: tempo, pitch (mean or range), articulation, timbre, mode, rhythmic complexity, dynamics, sound level, and melodic complexity. In addition to the emotional categories happy, sad, fearful, and angry, which are most prevalent in musical emotion research cited above, we included tender and peaceful to add some more

TABLE 1. *Selection of Automatically Extracted Features by the MIRtoolbox*

Musical dimension	MIRtoolbox function	Technical definition	Perceptual interpretation
Dynamics	*mirrms*	Root mean square of the amplitude	Loudness
	mirlowenergy	Percentage of frames with less-than-average energy	Loudness contrast ("musical dynamics")
Timbre	*mirentropy*	Shannon entropy of spectrum (viewed as random distribution)	Noisiness, spectral density (wide/narrow)
	mirbrightness	Amount of energy above a fixed frequency, default = 1500 Hz	High-frequency content, brightness impression.
	mirzerocross	Number of times the signal crosses the x-axis	Noisiness conflated with high-frequency content
	mirrolloff	The frequency such that a certain fraction of the total energy is contained below that frequency; the ratios .85 and .95 were both calculated	High-frequency content, brightness
	mirmfcc	Mel-frequency cepstral coefficients (13 sub-bands)	Description of spectral shape (e.g., spectrum of spectral envelope). No direct interpretation available.
	mirroughness	Based on dissonance curve after Plomp & Levelt; mean for all pairs of spectral peaks	Dissonance
	mircentroid	Centroid (center of mass, frequency expectation value) of the spectral distribution	Brightness
	mirflatness	Ratio of geometric and arithmetic mean of spectrum	Peakedness (prominent frequencies) and/ or smoothness (noisiness) of spectrum
	mirspread	Standard deviation of the spectrum	Measure of spectral density (wide or narrow), Fullness
	mirskewness	Skewness of spectrum	Asymmetry of the spectrum
	mirregularity	Degree of variation between successive peaks in spectrum	Homogeneity of spectral peaks, rough measure of polyphony
	mirflux	The Euclidean distance between spectrums of successive frames; median & mean	Sudden changes in spectral content indicative of (percussive) onsets
	mirflux subband	First decomposes the input waveform using a 10-channel filterbank of octave-scaled second-order elliptical filters, with frequency cut of the first (low-pass) filter at 50 Hz	As *mirflux* but separately for spectral sub-bands (here: sub-band 1 to sub-band 10)
	mirkurtosis	Excess kurtosis, of the spectrum	—
	mirnovelty	Novelty index; here: based on *mirspectrum*, a frame-based analysis of the spectrum; computes spectral self-similarity; probability of spectral transitions	Spectral changes; might correspond to perceived musical contrast
	mirattacktime	Temporal duration of attack phase (seconds); based on the *mironset* function	Articulation (staccato, legato), presence of percussion instruments
Pitch	*mirpitch*	Extracts pitch estimates using peaks of the spectral autocorrelation function	Mean of extracted pitch estimates, might correspond to brightness
Tonality	*mirkeyclarity*	Unambiguousness (clarity) of the estimation of tonal centers	Key clarity can be viewed as a measure for chromaticism
	mirmode	Estimates the modality, i.e., major vs. minor, returned as a numerical value between -1 and +1	Overall modality
	mirhcdf	The Harmonic Change Detection Function (HCDF) is the flux of the tonal centroid	Rate of harmonic changes
Tempo	*mirtempo*	Tempo based on periodicities from the onset detection curve	Tempo
Rhythm	*mirpulseclarity*	Rhythmic clarity, indicating the strength of the beats estimated by the *mirtempo* function	Beat strength, beat induction
	mirfluctuation	Rhythmic periodicity along auditory channels; here: mean of maximum across frames	Rhythmicity, beat induction

TABLE 2. *List of Extracted Features*

MIRtoolbox function	Extracted variables	Number
Mirrms	*rms_mean, rms_std*	2
mirlowenergy	*low_energy_mean*	1
mirentropy	*spec_entropy_mean, spec_entropy_std*	2
mirbrightness	*brightness_mean, brightness_std*	2
mirzerocross	*zerocross_mean, zerocross_std*	2
mirrolloff	*rolloff85_mean, rolloff95_mean*	2
mirmfcc	*mfccx_mean, mfccx_std (x = 1 to 13)*	26
mirroughness	*roughness_mean*	1
mircentroid	*centroid_mean, centroid_std*	2
mirflatness	*flatness_mean, flatness_std*	2
mirspread	*spread_mean, spread_std*	2
mirskewness	*skewness_mean, skewness_std*	2
mirregularity	*regularity_mean*	1
mirflux	*flux_mean, flux_std, flux_med_mean, flux_med_std, subbandx_mean, subbandx_std (x = 1 to 10)*	24
mirkurtosis	*kurtosis_mean, kurtosis_std*	2
mirnovelty	*spectral_novelty_mean*	1
mirattacktime	*attacktime_mean, attacktime_std*	2
mirpitch	*pitch_mean, pitch_std*	2
mirkeyclarity	*keyclarity_mean*	1
mirmode	*mode_mean, mode_std*	2
mirhcdf	*harmonic_change_mean*	1
mirtempo	*mirtempo_mean, mirtempo_std*	2
mirpulseclarity	*pulse_clarity_mean*	1
mirfluctuation	*fluctuation_max_mean*	1

Note: std = standard deviation, spec = spectral, flux_med = median of flux; rolloff85/rolloff95 = ratio see Table 1, fluctuation_max = maximum of fluctuation.

positive emotions (see also Bresin & Friberg, 2011; Eerola et al., 2013; Friberg et al., 2014; Juslin, 1997; Juslin & Lindström, 2011; Juslin et al., 2002). In addition, we implemented some cross-modal adjectives to capture semantic connotations of emotion categories that might have a counterpart in acoustics (e.g., dark–bright, rough–smooth, cold–warm; see, for example, Alluri & Toiviainen, 2010, for a similar approach).

We decided on the seven-point rating scale for increased consistency between items, though other solutions are conceivable (e.g., Eerola et al., 2013). For example, mode can be interpreted as a nominal characteristic of music. Providing a rating scale enables the possibility to capture the full range of diversity.

PROCEDURE

The evaluation of 60 musical excerpts took place in one session, ranging from 2.5 to 4.0 hours. Participants received written instructions before the session. The session started with three music examples on the basis of which participants chose their preferred volume for listening, followed by the general assessment of their musical preferences and the Gold-MSI. Then the evaluation part started. In each trial, participants first listened to the complete musical excerpt, followed by an evaluation on the 22 items. Participants initiated each trial by key-press and breaks could be taken between trials at any time. After 30 trials the participant contacted the experimenter to clarify whether he or she preferred to continue, or would rather postpone to

TABLE 3. *Key MIR Features Corresponding to Perceptual Ratings*

Evaluated feature	Verbal anchors (perceptual variable)	Perceptual variable	Acoustic variable	Spearman's |*rho*|
Pitch	low–high	*pitch_h*	*pitch_mean*	0.40***
Loudness	loud–soft	*loud_soft*	*rms_mean*	0.64***
Spectrum range	small–wide	*wideSp*	*spread_mean*	0.32*
Spectrum fullness	full–sparse	*sparseSp*	*spec_entropy_mean*	0.45***
Tonality	minor–major	*tonal_maj*	*mode_mean*	0.58***
Speed/Tempo	fast–slow	*speed_sl*	*mirtempo_mean*	*ns*
Articulation	staccato–legato	*artic_flow*	*attacktime_mean*	*ns*
Consonance	consonance–dissonance	*con_diss*	*roughness_mean*	0.30**
Harmonic variability	low–high	*harm_vari*	*harmonic_change_mean*	0.34**
Melodic variability	low–high	*mel_vari*	–	–
Rhythmic variability	low–high	*rhyth_vari*	*pulse_clarity_mean*	*ns*
Dynamic variability	low–high	*dyn_vari*	*low_energy_mean, rms_std*	0.26*, *ns*
Homophony	homophonic–polyphonic	*homo_poly*	–	–
Brightness	dark–bright	*dark_bright*	*brightness_mean, centroid_mean*	*ns*, 0.32*
Roughness	rough–smooth	*rough_smooth*	*roughness_mean*	0.48***
Temperature	cold–warm	*cold_warm*	–	–

Note. List of perceptual features rated by the participants, their verbal anchors for the seven-point rating scale, their variable names to decode Figure 3, the assumed dominant MIR feature, and the correlation coefficient between acoustic and perceptual feature. There is no directly related feature for the ratings of melodic variability, homophony and temperature. *** = $p < .001$, ** = $p < .01$, * = $p < .05$, *ns* = no significant correlation.

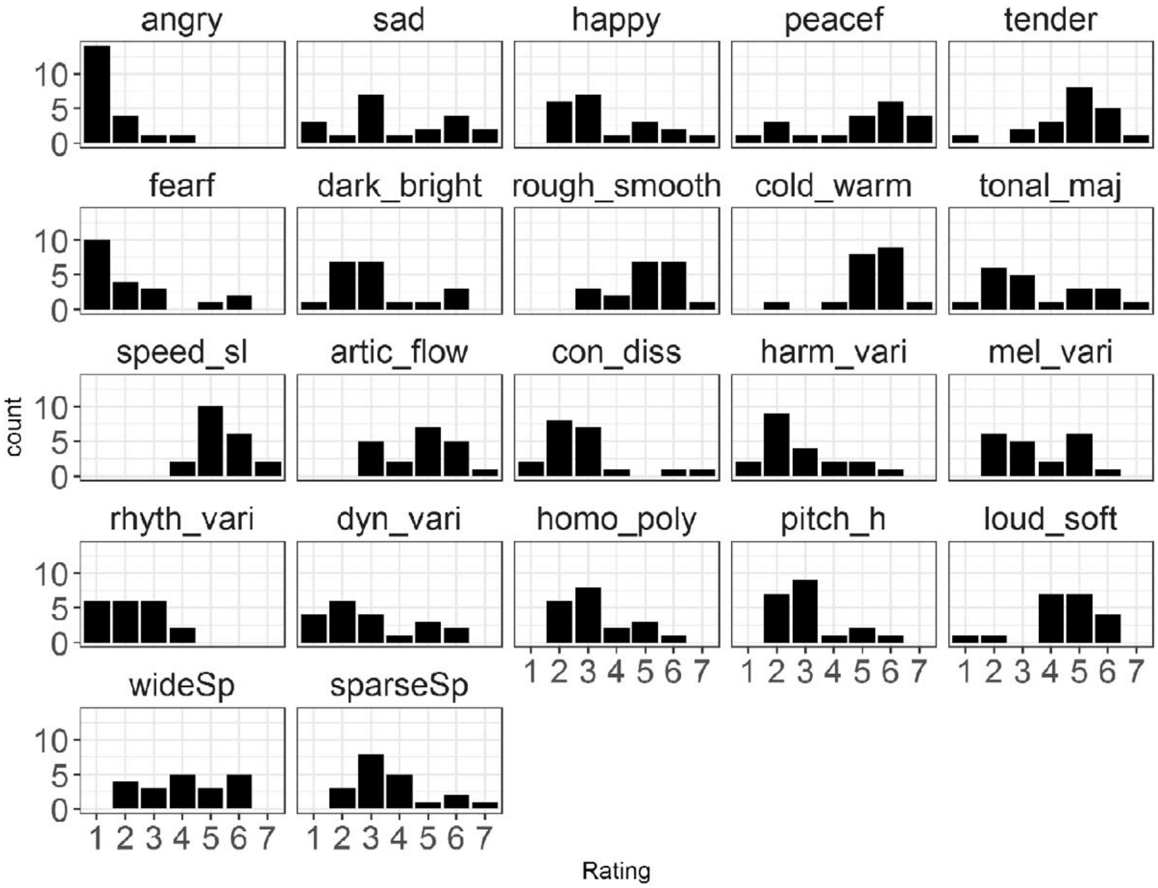

FIGURE 2. Histograms of the 22 perceptual ratings of the excerpt from Albatross (Fleetwood Mac). The rated value from the seven-point scale is shown on the *x*-axis. Variable names are explained in Table 3 of the main text. Ratings often did not show a clear unimodal or normal distribution (e.g., pitch, loudness, articulation, variability of melody, happy, peaceful, etc.).

a different day. This was implemented to offer the opportunity of adjusting time scheduling based on the participants' needs. All of the participants decided to stay for the next 30 trials, but took a break of self-chosen duration before continuing. The serial order of musical excerpts was randomized for each participant.

Results

We will first evaluate subjective ratings, then feature extraction and feature selection. Finally, we will compare the predictive power for subjective ratings and extracted features on emotion perception. We decided on this serial order of evaluations to mimic the perspective of a researcher without experience in MIR who wants to apply feature extraction to gain more knowledge about human musical processing. Our exploratory approach starts with testing data requirements and ranges to evaluate relations between musical features and emotional expressions. In our analysis, we regard $p = .05$ as significance level or t >= 2 in the *lmer* account. However, our study is largely a descriptive and a model building study, and *p* values are not used to test hypotheses (please see also the interactive visualization of some of the results at https://jazzomat.hfm-weimar.de/MIER/).

SUBJECTIVE EVALUATIONS

Statistical analysis typically requires identifying an appropriate distribution to characterize the data. A common assumption is that variables are distributed normally and that means and standard deviations capture the essence of this distribution. Hence, only means and standard deviations are reported. We question this general assumption of normally distributed features. In fact, separately examining the distributions of our 22 subjective evaluations for each music excerpt showed frequently bimodal and widespread distributions (see Figure 2 for all perceptual ratings for one musical piece).

TABLE 4. *Reliability of Subjective Ratings*

Variable	Type	MMP	MMR	*M*	*SD*	α
pitch_h	spectral	0.75	15.00	4.27	1.34	0.38
loud_soft	spectral	0.87	17.33	3.55	1.22	0.36
wideSp	spectral	0.70	14.00	4.33	1.53	0.11
sparseSp	spectral	0.72	14.33	3.85	1.41	0.11
tonal_maj	musical	0.72	14.33	3.79	1.75	0.31
speed_sl	musical	0.82	16.33	4.00	1.60	0.65
artic_flow	musical	0.68	13.67	3.75	1.66	0.34
con_diss	musical	0.72	14.33	3.17	1.45	0.32
harm_vari	musical	0.63	12.67	3.52	1.64	0.20
mel_vari	musical	0.67	13.33	3.87	1.67	0.26
rhyth_vari	musical	0.67	13.33	3.11	1.58	0.19
dyn_vari	musical	0.65	13.00	3.36	1.58	0.20
homo_poly	musical	0.65	13.00	3.87	1.58	0.14
dark_bright	modal	0.80	16.00	4.38	1.50	0.37
rough_smooth	modal	0.78	15.67	3.98	1.58	0.32
cold_warm	modal	0.73	14.67	4.41	1.40	0.21
angry	emotion	0.40	8.00	2.43	1.67	0.15
sad	emotion	0.58	11.67	2.92	1.80	0.20
happy	emotion	0.67	13.33	3.70	1.86	0.32
peacef	emotion	0.48	9.67	3.41	1.92	0.15
tender	emotion	0.60	12.00	2.91	1.80	0.21
fearf	emotion	0.47	9.33	2.30	1.58	0.09

Note: MMP: Proportion of significant Dip Tests ($p < .05$) across all excerpts, MMR: Ratio of observed to expected number of significant Dip Tests, *M*: Mean across all excerpts, *SD*: Standard deviation across all excerpts, α: Krippendorff's α.

To capture the multimodality in our data set in a quantitative way, we used Hartigans' Dip Test for multimodality (Hartigan & Hartigan, 1985). Table 4 lists the percentage of significant Dip Tests on the $\alpha = .05$ level for all judgments over all excerpts. Instead of using Bonferroni correction for multiple tests, we calculated the ratio of the proportion of estimated to expected significant Dip tests (multimodal ratio, MMR). From Table 4, it can be seen that the MMR ranges between 8.00 and 17.33, with a mean of 13.41; hence, all rating variables show much more multimodality than expected. There are many possible reasons for multimodality. First of all, it may just be an artifact of the measurement process (i.e., different understandings of the rating scales utilized). Intriguingly, it might also be due to perceptions that actually differ. Without independent external gauging, these two cases cannot be differentiated, which hints at a general measurement problem of using simple rating scales in psychology.

For the bipolar items, multimodality often occurred because participants judged the item in one direction or the other, avoiding the middle of the scale. This also means, on the positive side, that our participants had a specific opinion and did not often rely on the "neither-nor" judgment. However, as a consequence, means are often close to the scale midpoint and the standard deviations were relatively high (see Table 4). One might argue that some participants may not have understood the correct direction of the scale. But if so, only few participants should disagree with the majority of the others, and this was not the case. Outliers are also unlikely to explain multimodality. Hence, means are only a crude approximation of the full distributions for the rating variables.

One potential solution when analyzing data like this is using linear mixed models. Systematic disagreement between participants and items are specifically accounted for. Multimodally distributed variables pose no problem as long as residuals of the models are normally distributed. This makes it possible to uncover a more general relation between predictors and emotion rating.

Besides multimodality, there is also the problem of inter-rater reliability. Hence we checked whether the ratings were consistent among participants. We choose Krippendorff's α (Krippendorff, 1970) for inter-rater reliability as a complementary method to our multimodality index, as it is a generalization of several other different inter-rater reliability measures (e.g., Fleiss's Kappa) and can be applied to data on all measurement levels; it can also handle missing data. Basically, it compares the number of all matches in ratings observed and expected (i.e., chance matches), weighted by the difference of value pairs for numerical variable. We also calculated the mean Cronbach's α (as the mean over stimuli of mean pairwise correlations of rating variables).[2] It turned out that the two values correlation of $r = .99$ indicates that they are practically identical, but identically low. In fact, Table 4 shows rather low to moderate agreement between participants. The mean Krippendorff's α_K across all rating variables was .253 (mean Cronbach's α_C is .296). For emotion-related variables, the values were slightly lower ($\alpha_K = .186$) than for all other variables ($\alpha_K = .279$), but a Wilcoxon test was not significant ($p = .15$). Similar low α_Ks for listener agreement have been reported recently in a study by Schedl et al. (2016), using very homogeneous classical musical material and a large student sample. This indicates that our observed disagreement in judgments was not likely to have been due to our heterogeneous music selection or the size of our sample. One underlying reason for heterogeneity might be individual

[2] We did not calculate Cronbach's αs with stimuli as items, which seems to be done in several previous studies, and which would have given much higher values (>.9) due to the large number of items ($N = 60$). We think that such a procedure would violate the basic assumptions underlying Cronbach's α derivation, i.e., items are measuring the same construct. This can also be seen by the fact that for $N = 60$ items, mean correlations of small as $r = .14$ result in Cronbach's α larger than .90.

differences. It has been demonstrated that at least emotional judgments can be affected by the current positive or negative mood as well as whether one has a neurotic or extroverted personality (Vuoskoski & Eerola, 2011).

To check for such a potential systematic influence, we applied a cluster analysis using K-means with three clusters. This type of analysis aims to reveal subgroups of participants with similar rating behavior based on the overall mean of ratings across all items. Indeed, we were able to obtain three clusters of approximately equal size. The clusters were particularly driven by the emotion ratings (see Figure 3). One group showed rather low ratings on perceived six emotions ($n = 7$, $M = 2.58$, $SD = 0.94$), the next medium ($n = 8$, $M = 3.0$, $SD = 0.82$), and the last high ($n = 5$, $M = 3.37$, $SD = 0.88$). Those groups were thus labeled "reserved," "moderate," and "enthusiastic," respectively. Across all other (non-emotional) ratings the groups showed comparable results, with M(reserved) = 3.81, M(moderate) = 3.85, M(enthusiastic) = 3.81. The clusters imply that the relational structure across ratings of different music excerpts remained unchanged (e.g., the raters agreed that excerpts x and y were happier than z). Baseline differences on emotion ratings between groups could have been the result of rating style (from more reserved to more enthusiastic), differences in traits, or mood and personality (Vuoskoski & Eerola, 2011).

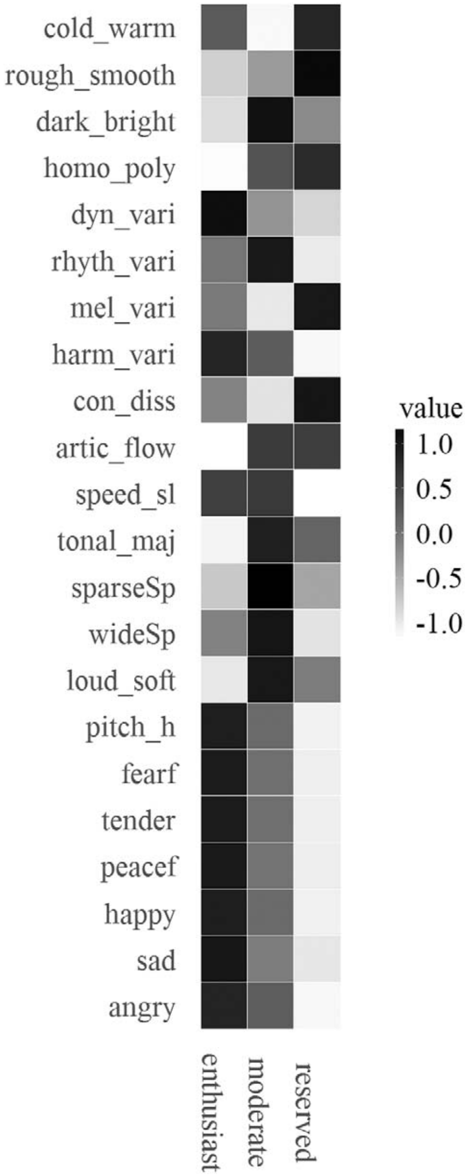

FIGURE 3. Heat map of scaled ratings with respect to groups of emotion rating tendencies. For an explanation of variable names, see Table 3. The grey shading in each square corresponds to mean value of z-transformed ratings.

TESTING MUSICAL TEMPO

The MIRtoolbox is very commonly used in music research due to its convenient and easy application to WAV files. It has been promoted for music research for many years, and is an excellent tool for researchers who lack the ability to get deeply into signal analysis. As noted above, extraction of the "tempo" feature only recently became reliable as indicated by the MIREX competition. However, it is one of the key features for emotion communication in music (e.g., Juslin, 2000). Therefore, we decided to evaluate this feature more specifically. To this end, we (the second author, KF) manually annotated all 60 musical pieces with a tempo using the tapping capabilities of Sonic Visualiser (Cannam, Landone, & Sandler, 2010). The first observation is that no meaningful tempo can be ascribed to some stimuli because the beat induction is too low or completely absent (one extreme example is *Poème symphonique* by György Ligeti). In total, seven of the 60 stimuli showed no meaningful tempo (three from world music, three from classical music, and one jazz piece). In contrast, tempo detection algorithms (nearly) always produce a tempo value for any stimulus, which can in some circumstances be misleading. The detection of tempo absence, rubato, and unmetered music is not an easy task, but should nevertheless be the first step of any tempo detection algorithm (Ahlbäck, 2004). The second observation is that many musical pieces admit more than one possible meaningful tempo. This is due to simultaneous rhythmic layers that suggest different tempos—often in a half-time or double-time relationship—which have roughly equal perceptual salience. In these cases, the annotator chose the tempo that could be considered subjectively and analytically most plausible or conventionally defined. As it turns out, manual annotated tempo and the extracted MIR tempo did not correlate, $r(51) = -.05, p = .73$. This is mostly due to half and double time errors. However, the tempo detection algorithm implemented in MIRtoolbox is definitely no longer state of the art (Böck, Krebs, & Widmer, 2015), which calls for a cross-check with more modern algorithms. Furthermore, this result also reflects the diverse nature of our sound samples, since tempo detection algorithms typically work best and are designed for pop music with a prominent drum groove.

ACOUSTIC AND PERCEPTUAL FEATURES

One interesting question in the music psychological community is whether algorithms can model human perception sufficiently well to be used as a surrogate. In Table 3 we listed all perceptual features rated by the participants and related them to (presumably) appropriate MIR features. For simplification, we will continue to use the term *acoustic features* for features extracted from the acoustic signal via MIR and *perceptual features* for ratings depicting the participants' related perception. We report correlations between extracted and perceived measures in the last column of Table 3. On the positive side, most correlations were significant. Some of them, such as loudness, tonality, roughness, and fullness of the spectrum resulted in high correlation coefficients. Of those, at least loudness and tonality are established key features for predicting the emotional expression of music. Hence, overall, results are satisfactory. However, algorithms hardly match human impressions for features of higher levels, like variability of harmony, rhythm, and loudness. In addition, the MIR tempo estimate is especially problematic with $|r| < .001$, $p = .99$. However, a strong correlation occurred when we correlated perceptual speed with our manually annotated tempo, $|r| = .83, p < .001$. Hence, in case of tempo, musical knowledge and experience is essential to capture tempo measurements (e.g., annotated tempo) that match perception. Interestingly, low correlations such as between *mirtempo_mean* and *speed_sl* are likely not due to the insufficient reliability of the perceptual ratings. For instance, though reliability for tonality was not high ($\alpha_K = .31$), mean perceived tonality correlated highly with the extracted one ($|r| = .58$).

ACOUSTIC FEATURE SELECTION: REDUCTION OF REDUNDANCY IN THE MIR DATA SET

Within the MIR context a huge variety of different algorithms were developed that are able to extract features from audio data. These have been used in a large variety of contexts and tasks in the MIR community with differing success and progress. In the context of research on music perception, however, there is no consensus on which features and algorithms to choose out of the plethora of possibilities. As in the MIR field, one common approach is to use a large set of features (e.g., 25 features in Alluri et al., 2012; 54 in Friberg et al., 2014; 17 in Leman et al., 2005; 114 in Yang et al., 2008) and to select a suitable subset for the task at hand using statistical methods.

One serious and common problem is that features can be highly correlated (Figure 4), making it difficult not only to use them in statistical models but also to understand the underlying characteristics of the music, thus hindering interpretability. Surely, correlations are inevitable in part due to the mathematical dependencies between algorithms to detect features and in part to actual correlations in the musical domain between certain sound characteristics. For instance, zero crossing rate is related to brightness as well as to noisiness, and musically, fast music is often played with higher volume. In addition, in certain heavy metal styles, the use of distorted guitar sounds and a preference for high tempo will result in high roughness values stemming from different sources. These underlying relations cannot be avoided. Hence, it is important to increase the awareness that these kinds of problems can occur. In our data set of 86 features, 12% of the pairwise correlations showed $|r| > .60$ with $p < .001$, and about 1% correlations were higher than $|r| > .90$, e.g., *centroid_mean* and *skewness_mean*, or *brightness_mean* and *zerocross_mean,* or *flatness_mean* and *kurtosis_mean*. Unfortunately, no systematic study of the correlations of audio features in music is known to the authors. One approach could be to test audio features with well-defined and fully controllable sounds such as noise, click tracks, and sine tones to detect systematic confounds.

Then, how to select appropriate features? We tested and compared four approaches: reduction of correlations by stepwise variable elimination, feature clustering, manual selection for interpretability, and random features as a baseline control. The feature clustering and interpretability accounts, though somewhat subjective, are clearly not arbitrary but rather justified. They are complemented and contrasted by the more objective method of correlation reduction. Additionally, we attempted to implement a PCA of acoustical features for comparison. A PCA is not a feature selection procedure in the narrow sense, but it is a standard feature reduction method. However, due to the high number of highly correlating features, the PCA did not find any factor solution and had to be discarded as a feature reduction tool in our case.

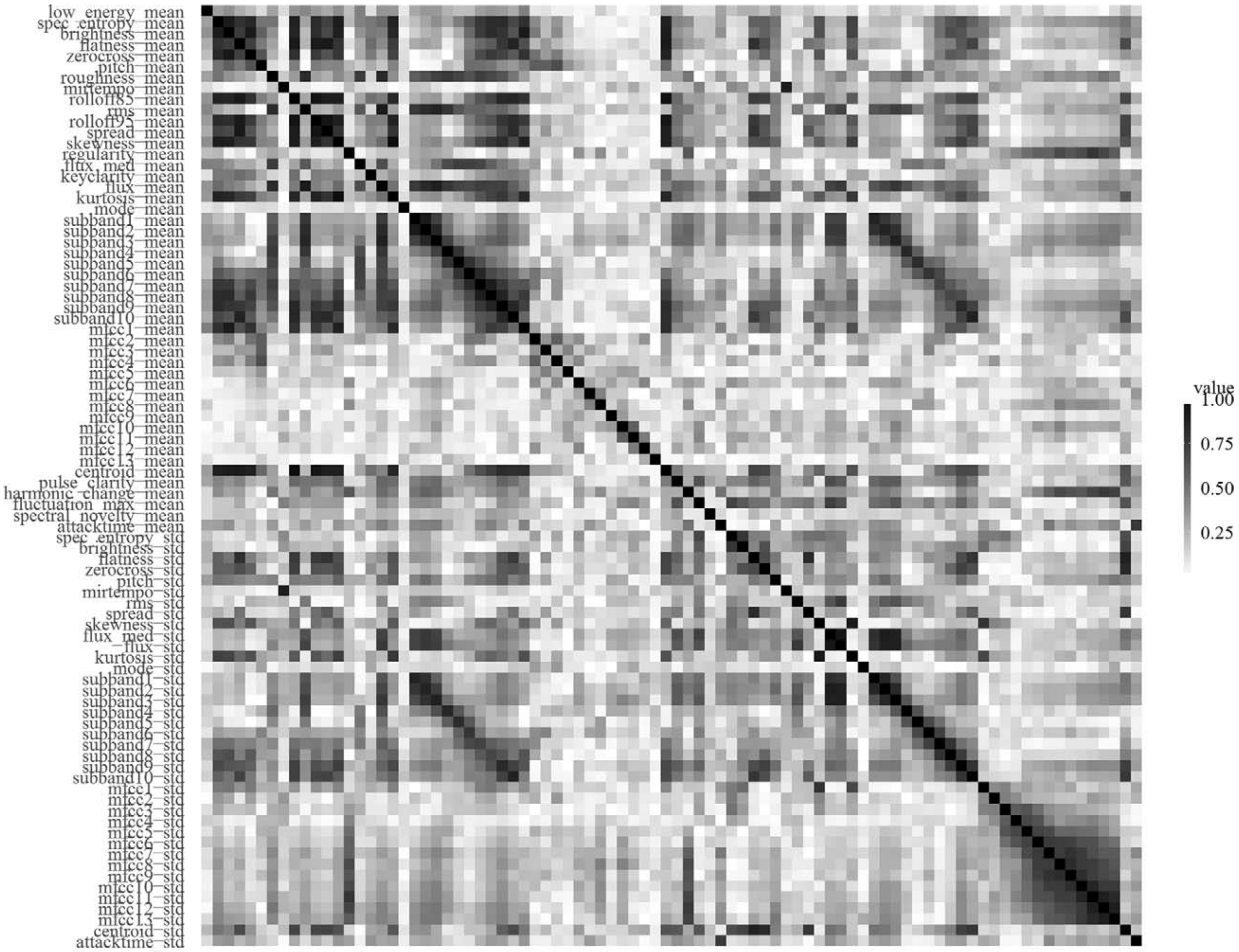

FIGURE 4. Plot of absolute values of Pearson correlation between 86 MIR features.

Correlation reduction by variable elimination (CR). Figure 4 depicts the complete correlation matrix (absolute values): the darker the fields, the higher the correlations. In the upper left part mean values are listed, in the lower left part *SD*s. The two chunks of high correlations are interrupted by the 13 mean MFCCs which, as expected (due to their construction), do not correlate highly. Further exceptions of mostly orthogonal features are *mirtempo_mean*, *regularity_mean*, and *mode_mean*. Other than that, most of the means and most of the *SD*s correlate highly with each other. We devised a features selection procedure by reducing the overall sum of absolute correlations. One way to do this is to remove successively the variable with the highest absolute mean correlation (independently of positive/negative values, since the sign of a correlation is irrelevant for predictive power), until no pairwise correlations higher than a certain threshold were left. The threshold here was chosen as $|r| = .30$, which roughly corresponds to the threshold of significance for correlations of $N = 60$ sample points. Using all 60 excerpts and 86 extracted features, this procedure resulted in 10 key features. Since random variation affects the algorithms, we cross-checked our selection by applying a bootstrap procedure, randomly selecting 50 music excerpts from our pool and repeating the selection 100 times. The most frequently selected features were thus: *low_energy_mean*, *mode_mean*, *mfcc5_mean, mfcc8_mean, mfcc13_mean, spectral_novelty_mean, rms_std, mode_std, subband8_std, mfcc2_std*. It is not surprising that three *mfccs* are included here, as they do not correlate with other variables in the first place (see Figure 4).

Feature clustering (FC). In a second method, we used hierarchical clustering on the correlation matrix of features and step-wise selection of features from strongly connected feature groups. To this end, we manually selected one variable for each redundancy cluster that

1) best represents the cluster, and 2) is most easily interpretable within the cluster, and repeated this procedure with the remaining variables until saturation was reached (i.e., until clusters were mostly singletons or no clear preference for best variables could be found). Of course, this selection is rather subjective as it is shaped by researchers' individual preferences. We suggest these 12 selected features**:** *low_energy_mean, pitch_mean, mirtempo_mean, regularity_mean, keyclarity_mean, mode_mean, pulse_clarity_mean, harmonic_change_mean, spectral_novelty_mean, attacktime_mean, mode_std, attacktime_std*. Two features overlapped with the results of CR: *low_energy_mean*, and *mode_mean*, indicating that they are distinct.

Interpretability (INT). Our most subjective selection procedure is based on the interpretability of the features. After careful examination, we finally chose the following 11 features: *tempo* (annotated), *mode_mean, rms_mean, pitch_mean,* for the key musical characteristics of tempo, tonality, loudness, pitch, and *attacktime_std, pulse_clarity_mean, keyclarity_mean, mode_std, low_energy_mean, regularity_mean, spectral_novelty_mean* for variability on these dimensions including articulation and rhythm (see Table 1 and 2).

ACOUSTIC FEATURE SELECTION: EMOTION MODELING

Our main goal was to test whether emotion ratings can be modeled using different predictor sets, notable extracted audio features, or perceived features. For a general solution that should be applicable in many different contexts, it is inevitably necessary to model mean values of emotion ratings. However, as discussed above, multimodality and the large variance of ratings—as well as low inter-rater reliability—make this approach rather doubtful from the outset. Nevertheless, using mean ratings as a proxy for random distribution is a common method in music psychology, so we decided to explore how successful this approach might be with our data. To this end, we fitted linear models for each of the six emotion variables using our sets of selected acoustic features as predictors and evaluated the model quality. For the modeling account, we added *tempo* (annotated) to all feature sets but INT, which already included *tempo*. We calculated five-fold cross-validations for root-mean-square prediction errors (RMSPE) for each target variable and feature set. In addition, we calculated adjusted R^2 and *p* values as well as Akaikes Information Criterion (AIC) for the full (non-cross-validated) model as standard evaluation parameters for model quality. Since our feature sets contained the manually annotated tempo, which is not available for seven stimuli, we used a reduced set of 53 musical pieces. As a baseline, we added a set of 10 randomly chosen features, the *Random features (RF): zerocross_std, pitch_mean, mfcc3_mean, mfcc9_mean, mfcc13_std, subband7_mean, subband2_std, subband10_std, centroid_std, skewness_std.* Results are depicted in Figure 5.

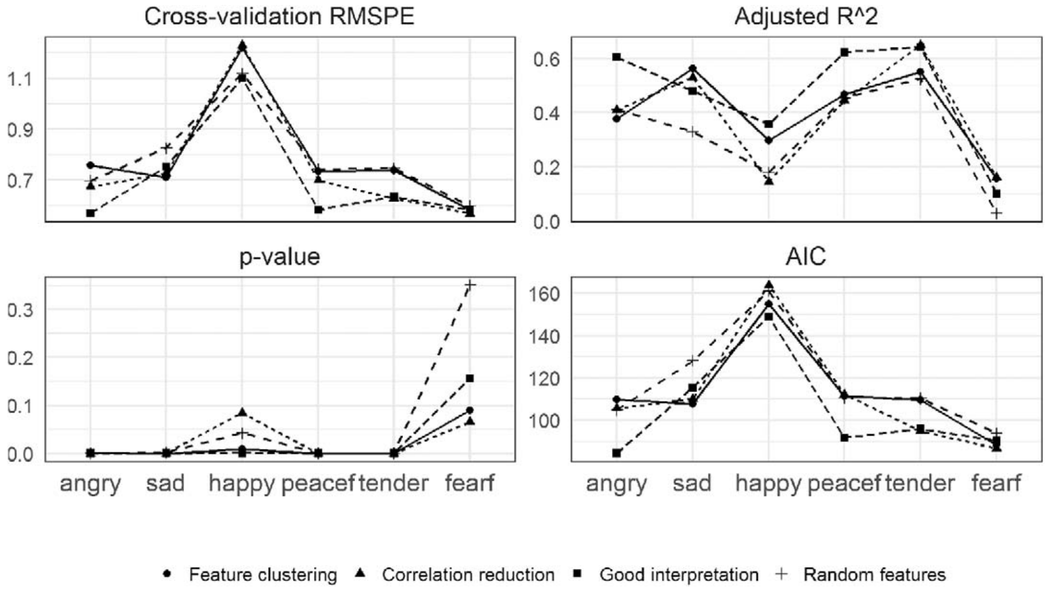

FIGURE 5. Five-fold cross-validation of linear models for six emotion ratings using four different subsets of MIR features: correlation reduction ($n = 10$ plus *tempo* (annotated)), feature clustering ($n = 12$ plus *tempo* (annotated)), good interpretation ($n = 11$), and random features ($n = 10$ plus *tempo* (annotated)). Models differ only marginally with respect to prediction quality (here: root-mean-square prediction error, RMSPE), explained variance (adjusted R^2), significance (*p* value), or quality of model fit (AIC).

It can clearly be seen that all four feature sets show quite comparable performance. Interestingly, the best model is the set chosen for good interpretability (mean AIC = 104.5). As in the case of fully random feature sets, the variable *happy* is the hardest to predict, with RMSPEs of about 1.5; only two feature sets resulted in significant models for *happy* (FC and INT with $p < .01$). Though the variable *fearf* shows low RMSPE, no model actually was significant (all p's > .01), which may be due to the fact that *fearf* has very low variance (Table 4) so that a rather low RMSPE is already achieved by the null model. In addition, using means of mean ratings further reduces the variance, since the variance of the means is smaller than the mean of variances. For instance, the standard deviation of the mean values of *angry* ratings across all stimuli is 0.73, but the average of standard deviations for all *angry* ratings is 1.49, about twice as large. This also explains the results for the random feature set. It cannot predict the *happy* variable ($p > .05$) (for which the variance is slightly larger as for the other emotion variables), but it can satisfactorily predict *angry*, *sad, peacef*, and *tender*. This is partially due to the fact that the features are correlated with each other and the mean emotion ratings. Considering these results, the actual set of features seemed to be rather arbitrary for modeling emotion perception. But one has to bear in mind that we used a common set of features to predict different emotion ratings. This was the easiest approach but it might be suboptimal, as different emotions might call for different sets of optimal features (Eerola et al., 2013). However, this does not alleviate the general problem of rater disagreement and multimodality of ratings.

TABLE 5. *Pearson Correlations Between Key Acoustic Features*

	mode_ mean	*rms_ mean*	*pitch_ mean*	*attacktime_ mean*
tempo	− .03	.24	.33*	−.13
mode_mean		.01	−.07	.25
rms_mean			.13	−.42**
pitch_mean				−.05

Note. The correlation coefficient r is based on $n = 53$ musical pieces. ** = $p < .01$, * = $p < .05$.

TABLE 6. *Pearson Correlations Between Key Perceptual Features*

	tonal_maj	*loud_soft*	*pitch_h*	*artic_flow*
speed_sl	−.19	.43	−.22	.41
tonal_maj		−.04	.27	−.13
loud_soft			−.13	.24
pitch_h				−.17

Note. The correlation coefficient r is based on $n = 1060$ ratings. No p values are reported due to statistical dependency of ratings.

COMPARISON OF ACOUSTIC AND PERCEPTUAL FEATURES PREDICTING EMOTION PERCEPTION

We started our results section by pointing out that subjective ratings showed low Krippendorff's α in agreement with some other studies (e.g., Schedl et al., 2016). One method of taking systematic variances of individuals and/or musical pieces into account is using linear mixed-effect models. However, our statistically driven feature selection did not result in a clear choice of features to be included as predictors in the models. We therefore went back to what is known from the literature and decided for tempo, mode, loudness, pitch, timbre (here: brightness and roughness), and articulation as the most important cues for emotion communication (e.g., Gabrielsson & Lindström, 2001). We matched those perceptual ratings with the extracted features of annotated *tempo, mode_mean, rms_mean, pitch_mean, brightness_mean, roughness_mean,* and *attacktime_mean*. However, uncorrelated predictors are preferable for linear mixed models, because otherwise results are difficult to interpret. That is, because of high pairwise correlations ($|r| >= .58$ to .84) with other variables we excluded the timbre variables of brightness and roughness from the perceptual and the acoustic models. Pairwise correlations for the acoustic and perceptual features chosen are listed in Table 5 and 6.

We decided to fit linear mixed models, even though emotion ratings were discrete values between 1 and 7, but we assumed a continuous underlying distribution. Inspection of plotted residuals did not show a large deviation from normality (besides the consequences of discrete values). We included random intercepts for participants and musical pieces. We calculated model fits with the *lme4* package for R and p values by including the Satterthwaite approximation from the package *lmerTest*. We started with an implementation of all fixed effects and dropped fixed effects as long as the remaining model had an improved fit (or the same fit with fewer degrees of freedom).

As can be seen in Tables 7 and 8, acoustically measured as well as perceived tempo, tonality, and loudness were the dominant significant predictors for perceived emotions, corresponding with what is known from the literature (e.g., Gabrielsson & Lindström, 2001). Brightness (*pitch_mean, pitch_h*) and articulation (*attacktime_mean, artic_flow)* played only a minor role for both acoustic and rated features. Hence, overall there was quite a lot of agreement about which musical features can predict emotional content. However, in

TABLE 7. *Model Fits for Six Emotion Ratings Predicted by Acoustic Features*

	tempo (annotated)	*mode_mean*	*rms_mean*	*pitch_mean*	*attacktime_mean*
happy	b = 0.41 *t* = 3.35**	b = 0.33 *t* = 2.71**	b = −0.37 *t* = −2.78**		b = 0.27 *t* = 2.01
sad	b = −0.59 *t* = −6.67***	b = −0.19 *t* = −2.18*			
angry		b = −0.15 *t* = −1.96	b = 0.44 *t* = 5.80***	b = 0.24 *t* = 3.13**	
peacef	b = −0.26 *t* = −3.29**		b = −0.38 *t* = −4.66***	b = −0.19 *t* = −2.50*	b = 0.17 *t* = 2.09*
tender	b = −0.43 *t* = −5.35***		b = −0.45 *t* = −5.57***		
fearf	b = −0.17 *t* = −2.35*				

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$; we also report tendencies for fixed effects when $p < .10$

TABLE 8. *Model Fits for Six Emotion Ratings Predicted by Perceptual Features*

	speed_sl	*tonal_maj*	*loud_soft_*	*pitch_h*	*artic_flow*
happy		b = 0.25 *t* = 8.23***		b = 0.18 *t* = 4.32***	
sad	b = 0.23 *t* = 6.08***	b = −0.20 *t* = −7.00***			
angry	b = −0.13 *t* = −3.20**	b = −0.15 *t* = −4.90***	b = −0.24 *t* = −5.29***		b = −0.06 *t* = −1.96
peacef	b = 0.17 *t* = 3.95***	b = 0.07 *t* = 2.18*	b = 0.18 *t* = 3.51***		
tender	b = 0.17 *t* = 4.24***		b = 0.24 *t* = 5.24***		
fearf	b = 0.06 *t* = 1.86	b = −0.09 *t* = −3.44***			

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$

detail, the models differed by the exact relations between predictors and the predicted emotion. For instance, *tonal_maj* and *pitch_h* had high beta weights to predict *happy* from perceptual features. In contrast, in the acoustic model for *happy*, *pitch_mean* had no effect; instead, *rms_mean* and *tempo (annotated)* had rather high beta weights together with *mode_mean*. Basically, when comparing models for specific emotions in detail, there was not much correspondence between acoustic and perceptual models. The only exception were the models for *sad* and *tender*. Whereas *sad* was predicted by tempo and tonality, *tender* was predicted by tempo and loudness.

We then compared the model fits of acoustic and perceptual features to evaluate which predictors (acoustic or perceptual features) might be more useful to predict emotion perception. As we have already shown, perceptual and acoustic features may often be correlated but never perfectly so. That is, they capture different variance and it is not clear which kind of features are

TABLE 9. *Comparison Between Acoustic and Perceptual Predictors*

	df_a - df_p	BIC$_a$ - BIC$_p$	Better fit (smaller BIC) for:
happy	2	64.5	perceptual
sad	0	44.5	perceptual
angry	1	18.2	perceptual
peacef	− 1	− 18.4	acoustic
tender	0	− 6.6	acoustic
fearf	1	3.9	perceptual

Note. We report the difference in the Bayesian Information Criteria (BIC) between the final models with $_a$ = acoustic predictors, and $_p$ = perceptual predictors. A smaller BIC between two models indicates a better model fit. We also included the difference in degrees of freedom (*df*), i.e., the difference in the number of parameters.

more relevant for emotion communication. Table 9 reports the difference of model fits by the Bayesian Information Criterion (BIC). For the variables *happy*, *sad*, *angry*, and *fearf* the perceptual ratings resulted in better model fits than the acoustic features. For the variables *peacef* and *tender,* the acoustic models outperformed the perceptual models. That is, perceptual models showed a benefit over acoustic models in four of six comparisons. Subjective ratings outperformed the more objective extracted acoustic features in predicting the emotional content of music.

To further understand the benefit of the linear mixed effect account, we calculated explained variance (R^2) and root-mean-square prediction error (RMSPE) using the packages *MuMIn* and *merTools* in R. As can be seen in Table 10, those measures were very similar overall and did not support the interpretation of our mixed effect models. The conditional R^2s in Table 10 clearly show that there was systematic variance and accounting for it improved the model fits in comparison to the marginal R^2s. To conclude, linear mixed models do improve model fits but not as strongly as one would expect.

MODAL CONNOTATIONS OF MUSIC AND ITS RELATION TO EMOTION PERCEPTION

Some studies make the claim that musical attributes are related to extra-musical connotations (e.g., Bronner et al., 2012). We wanted to explore this issue and included three modal ratings in our study: *cold_warm*, *dark_bright*, *rough_smooth* attributes that originally stem from temperature, vision, and tactile perception. To this end, we again fitted linear mixed models, with the emotion ratings as dependent variable and the three modal ratings as potential predictors. All models included random intercepts for participants and musical pieces. Table 11 shows the significant fixed effects. Indeed, the connotation of temperature and tactile perception were involved in predicting most emotion ratings; brightness contributed less often. We can now ask which attributes have better predictive power: the musical or the "extra-musical" ones? Interestingly, in four out of six cases the modal models outperformed the perceptual ones (see Table 12). These results were not simply explained by a larger number of parameters, which was the case in one of those four cases. The perceptual fits were better in two negatively connoted emotions—*sad* and *angry*—but then included also more parameters. Comparisons of model fits converged partially with comparisons of R^2 and RMSPE in Table 10.

Discussion

Music information retrieval is a promising research area that is rapidly developing, with the central task of extracting musical features from audio files. When it is successful, its advantages are obvious: objectivity of measurements, fast extraction of a high number of features from a large amount of music. The impact on research in music psychology could be tremendous.

Our study intended to critically examine the use of MIR in the context of emotion communication by musical cues. For instance, it has been demonstrated in a vast number of studies (in both experimental and correlative designs) that tempo (Balkwill & Thompson, 1999; Eerola et al., 2013; Hevner, 1935; Juslin, 2000; Juslin & Lindström, 2011; Rigg, 1937; Scherer & Oshinsky, 1977) as well as mode (Eerola et al., 2013; Friberg et al., 2014;

TABLE 10. *Comparison Between Acoustic, Perceptual, and Modal Predictors*

	R^2m_a	R^2m_p	R^2m_m	R^2c_a	R^2c_p	R^2c_m	RMSPE$_a$	RMSPE$_p$	RMSPE$_m$
happy	0.14	0.09	0.11	0.47	0.43	0.44	1.33	1.30	1.29
sad	0.12	0.11	0.01	0.44	0.41	0.43	1.31	1.29	1.31
angry	0.11	0.09	0.05	0.30	0.29	0.25	1.38	1.36	1.37
peacef	0.12	0.05	0.07	0.38	0.34	0.36	1.48	1.48	1.46
tender	0.15	0.07	0.06	0.42	0.38	0.41	1.32	1.32	1.29
fearf	0.01	0.02	0.03	0.30	0.30	0.31	1.23	1.23	1.22

Note: R^2m= marginal R square associated with the fixed effects, R^2c= conditional R square taking fixed and random effects into account; $_a$ = acoustic predictors; $_p$ = perceptual predictors; $_m$ = modal predictors; RMSPE = prediction error.

TABLE 11. *Model Fits for Six Emotion Ratings Predicted by Modal Attributes*

	cold_warm	*dark_bright*	*rough_smooth*
happy	b = 0.13 $t = 3.57^{***}$	b = 0.31 $t = 8.66^{***}$	b = 0.10 $t = 2.89^{**}$
sad		b = −0.09 $t = -2.38^{*}$	
angry	b = −0.08 $t = -2.04^{*}$		b = −0.20 $t = -5.72^{***}$
peacef	b = 0.07 $t = 1.81$		b = 0.29 $t = 7.92^{***}$
tender	b = 0.27 $t = 7.46^{***}$		b = 0.27 $t = 2.56^{*}$
fearf		b = −0.15 $t = -4.63^{***}$	b = −0.6 $t = -1.98^{*}$

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$; we also report tendencies for fixed effects when $p < .10$

TABLE 12. *Comparison Between Modal and Perceptual Predictors*

	df_m - df_p	BIC_m - BIC_p	Better fit (smaller BIC) for:
happy	1	−8	modal
sad	−1	67.3	perceptual
angry	−2	14	perceptual
peacef	−1	−41	modal
tender	0	−24.7	modal
fearf	0	−13.5	modal

Note. We report the differences in the Bayesian Information Criteria (BIC) between the final models with $_m$ = modal predictors, and $_p$ = perceptual predictors. A smaller BIC between two models indicates a better model fit.

Hevner, 1935; Juslin & Lindström, 2011; Rigg, 1937) are highly associated with emotional content. Does that mean extraction of tempo, mode, and other features by MIR is an alternative to expert ratings?

We collected subjective evaluations of a broad range of musical excerpts by professional sound engineers. In a thorough check-up we evaluated the data quality of such perceptual ratings and of a large number of acoustic features, extracted by the MIRtoolbox (Lartillot & Eerola, 2007), one of the most widely used acoustical feature extraction tools. We then compared the predictive power of perceptual and acoustic variables. Our results are both promising and critical.

First, subjective ratings of spectral, musical, modal, and emotional attributes of music are far from consistent. Even deploying a highly homogeneous sample of experts whose profession is dependent on such evaluation skills does not necessarily result in uniform judgments. Then, interpretation of statistics based on subjective evaluations have to be treated with caution, because they are not unimodally distributed, and the more so for bipolar rating scales. Nevertheless, reliability measures can be of medium size. Our experts agreed more on speed, pitch, brightness, and loudness than on emotional attributes (see α_K in Table 4). That is, the chances are that listeners perceive features on a lower level of abstraction similarly, but on a higher level not. This conforms to what has been reported in the literature. For instance, Friberg et al. (2014) showed high inter-rater agreement (Cronbach's $\alpha > .80$) for a broad range of features such as speed and modality, but also rhythmic and harmonic complexity. Consistency on the emotional ratings such as energy and valence was lower (Cronbach's $\alpha < .57$). But very high consistency for emotional ratings can be found in other studies (e.g., Cronbach's $\alpha > .90$ in Eerola et al., 2013; Eerola & Vuoskoski, 2011; note that Eerola & Vuoskoski and Friberg et al. to some extent used the same material: film music). The broad range within emotional ratings in our study indicates large individual differences among subjects. As a consequence, predicting emotional content from perceived musical attributes is difficult. However, there seem to be systematic rating strategies, as it was possible to cluster emotion ratings from reserved to enthusiastic raters. Indeed, statistical modeling taking individual variance into account might level out such problems to some extent (e.g., linear mixed models with random intercepts for participants). Looking into the literature, methodological reasons for inter-rater inconsistencies are not conclusive. For instance, Friberg et al. (2014) showed high consistency (using Cronbach's α) between raters when the material was homogeneous (e.g., ringtones, film music), but Schedl et al. (2016) did not. Friberg and our study used a rather small number of participants ($n = 20$), but Schedl et al. (2016) used a large one ($n = 241$). Furthermore, studies usually differ on which rating scales were applied and the measurement of inter-rater agreement (e.g., Cronbach's α or Krippendorff's α). All these methodological differences make a comparison difficult. All in all, the inconclusiveness of findings from studies differing in the selection of stimuli, sampling of participants, and applied measurement indicates an intrinsic complexity of the task, posing challenges for any modeling effort.

Second, some of the perceived musical attributes correlated highly with their analogue acoustic features: loudness, tonality, and roughness. This is promising, as it shows that acoustic feature extraction has perceptual reality. It is particularly gratifying because loudness and tonality are among the key set of features for musical expressions of emotions (e.g., Gabrielsson & Lindström, 2010). In Leman et al. (2005), loudness and articulation in particular showed a good match between perceived and acoustic cues.

Third, we compared perceptual and acoustic features for emotion prediction using linear mixed models. Both kinds of attributes show that loudness, tonality, and tempo are important predictors for a range of emotions; this is in agreement with the literature (e.g., Gabrielsson & Lindström, 2001, 2010; Hevner, 1935, 1937; Scherer & Oshinsky, 1977). But the fourth known important predictor—pitch—did not show a convincing effect across all of our emotional categories. Instead, it resulted in a fixed effect only for the acoustic model. Interestingly, models diverged regarding the effective predictors. Acoustic and perceptual models agreed only in how they predicted sad and tender contents. For all other emotions, the range of fixed effects as well as their impact (relative order of beta weights) differed.

Fourth, when comparing model fits between acoustic and perceptual features, the perceptual judgments resulted in better predictions of emotional content than acoustic features for happy, sad, angry, and fearful expressions, but not for peaceful and tender. Acoustic models outperformed perceptual models for peaceful and tender expressions. But when looking into the explained variance and prediction errors of the model fits, the clear-cut benefit of perceptual models does not hold. We thereby cannot convincingly argue that we replicated Leman et al. (2005), who demonstrated a benefit for perceptual ratings in comparison to extracted acoustic features, using their own tools for feature extraction.

Fifth, when comparing linear mixed model fits (Table 10) to linear models of rating means (see Figure 5), linear mixed models do not seem to strongly improve model fits. However, using means as predictors for individual ratings (as in linear models) will add a prediction error that is about the size of the standard deviation of ratings on top of the prediction error for the mean (see also interactive visualization at https://jazzomat.hfm-weimar.de/MIER/, particularly the models tap). Since the linear mixed model RMSPEs (Table 10) are slightly but significantly lower than the standard deviations of emotions ratings (Table 4), one can conclude that linear models of means and linear mixed models for individual ratings perform comparably for modeling individual ratings. However, the overall gain in prediction accuracy of emotion ratings using acoustical, perceptual, or modal features (compared to a null-model) is rather small, about 0.2–0.4 on a seven-point Likert scale (when comparing the RMSPEs of the models in Table 10 with the *SD*s of the ratings in Table 4), but this is nonetheless significant. One possible interpretation is that these features are indeed able to capture basic components of emotional expression (e.g., the often reported features of tempo, intensity, tonality, and dissonance) but not more. We suggest that details of emotional expressions in music are contained and produced by (possibly strongly culture-dependent) semiotic layers that are on top of those representing "core affect" (Russell & Barrett, 1999). These finer details are hard to catch with rather general low-level acoustical features, but also not with perceptual features and cross-modal metaphors as used here, though these perform similarly well to the acoustical features. All in all, we think that the features investigated herein are not well-suited to model with high precision individual emotion ratings of complex real-world, non-functional music. Still, they can be used for rough measurements of general trends on the group level, with better performance for "emotionally well-defined" music such as film music or simpler stimuli (such as monophonic melodies or ringtones; e.g., Friberg et al., 2014).

Sixth, we evaluated the extracted features and showed high correlations between them. This is problematic because when those features correlate, it is difficult to understand what they are actually measuring. For instance, the psychological intention might be to measure brightness and noisiness, but *brightness_mean* correlates with *zerocross_mean* with $|r| > .90$, which is highly plausible from an acoustic point of view: brightness is driven by high frequency content which has a higher zero crossing rate than lower frequencies. The researcher might wish to investigate two independent qualities, but this is not the case. We argue that the awareness of hidden relations has to increase, and for obvious relations researchers should decide in favor of one out of a variety of related measures (e.g., to decrease the probability of Type I errors). In addition, other correlations might be less evident and more research is needed to understand, whether these are due to the way features are extracted or whether they are inherent to music.

Seventh, no universally applicable selection procedure has thus far been suggested to decrease the number of acoustic features with high pairwise correlations. A method such as partial-least square regression, which was developed for regression with highly correlated predictors, is not usable as a general feature selection method since it is tied to an actual data set for prediction, though it might be in certain application scenarios and with an alternative approach. In order to find a maximal predictive minimal set of features, we compared four selection procedures: eliminating high pairwise correlations (CR), feature clustering (FC), interpretability (INT), and a random control (RF) as baseline. There were almost no differences and models showed low RMSPEs (except for *happy*) when fitting linear regressions to predict the six emotion ratings separately. Specifically, a random selection of features did not impair model fit in comparison to the carefully selected sets. This is merely a statistical artifact due to the variance reduction of averaging low-variance variables, as well as highly correlated features. This is a rather unsatisfactory result as it implies arbitrariness for feature selection. It might partially explain why there is a rather low convergence of statistical importance of features for emotion communication. For example, in Leman et al. (2005) the following features were important to predict valence: onsets/sec, interonset interval, articulation. To predict activity, the important features were spectral centroid (SD), and channels that agree on the same pitch (e.g., dissonance-consonance measurement). Using a different implementation of extraction algorithms, Yang et al. (2008) reported for valence spectral dissonance (roughness of all spectrum components), tonality, sum of beat histogram, chord, and sum of pitch histogram as the top five features. For arousal, it was flux (SD and M), tonality, multiplicity, and spectral roll-off.

Hence, the best procedure for feature selection seems to be to strictly follow a logical account and use a low number of features of interest that are easily interpretable. Again, a low number decreases the probability of Type I errors (see also Schubert, 2004). Some authors suggest an “analytical account” factor to reduce the data set of features (e.g., Alluri et al., 2012; Leman et al., 2005), but this approach can lead to solutions that are data-dependent and hard to interpret; it might not even succeed due to very high correlations of some features, making a features selection step desirable or even necessary. Importantly, this result reveals that we need much more investigation to evaluate the features and their perceptual relevance before we will be able to properly interpret results and compare them across studies.

Eighth, we evaluated the function *mirtempo* in more detail. Specifically, we compared the outcome with an annotated tempo obtained by one experienced musicologist. Tempo is a difficult variable to extract as it depends not only on beat extraction (which can be challenging for rubato parts) but also on the choice of meter, which depends greatly on musical experience. One might even argue that tempo is actually a multidimensional variable and not a scalar value. Unsurprisingly, then, *mirtempo* did not properly capture the musical tempo. Researchers are strongly encouraged to take upon themselves the arduous task of annotating each musical piece individually by hand. An alternative might be to evaluate other, more state-of-the-art algorithms, but this was beyond the scope of our study.

Ninth, modal features might be an alternative when deciding on what perceptual attributes might be of interest. We showed that modal features predicted emotional content of music in four out of six models, though we included only three modal ratings in our study. How are emotional cues then communicated? Modal attributes—in our study, pertaining to the tactile and visual domains—might be more clearly related to emotions than musical attributes. Music seems to be associated with modal characteristics that directly capture the emotional content. It might be that the evaluations of perceived musical characteristics do not activate emotional associations as strongly as modal features do. This is, of course, a tentative interpretation that needs further direct studies.

We also see other necessary evaluative questions that we did not address in our investigation but that came up in the course of this study. For example, we found only one independent evaluation of the MIRtoolbox (Kumar, Kumar, & Bhattacharya, 2015), in which feature extraction was tested for its reliability. Three functions were analyzed: *mirchromagram*, *mirkeystrength*, and *mirpitch*. The study generally revealed very accurate measures, but also showed that the autocorrelation option in the *mirpitch* function results in unacceptable estimates for frequencies lower than 150 Hz. In addition, the chromagram for a trumpet was less accurate than for strings.

Another issue is the question of parameter settings. For example, data are analyzed within a moving window (“frame”). In the MIRtoolbox, the default setting is 50 ms with 50% overlap of successive windows (hop factor) for low level features. A bigger size and smaller hop factor is applied for *mirtempo* (3 s, 10%) and *mirpulseclarity* (5 s, 10%). However, there is no consensus in the literature on the best window sizes. Some authors report having applied default settings (e.g., Friberg et al., 2014); others did not specify it in ms (e.g., Coutinho & Cangelosi, 2011; Kumar et al., 2015; Leman et al., 2005; Li & Ogihara, 2006), others applied 25 ms and 50% overlap for short-term features (Alluri & Toviainen, 2010; Alluri et al., 2012; Poikonen et al., 2016), and 3 s with 33% overlap for long-term features (Alluri et al., 2012). Other sizes are 23 ms (Elliot, Hamilton, & Theunissen, 2013) or 46 ms (Eerola et al., 2009; Hwang et al., 2013), justified by the structure of the cochlea (Hwang et al., 2013). To our knowledge, no systematic evaluation of window size and its relation to perception has thus far been reported. However, we found that the results of functions using different window sizes did not necessarily correlate highly. This raises the concern that window size as a parameter is important and should not be underestimated in affecting the outcome. In fact, Alluri and Toviainen (2010) reported that increasing the window size most often resulted in a decrease but sometimes also an increase of explained variance in their regression models. They argue that this points to different time scales of feature processing, which would need further investigation. Finally, many features offer a variety of parameters besides window size (the rule of thumb is: the higher level of the feature, the more free parameter choices) which, strictly speaking, can be individually or jointly optimized, adding further degrees of complexity to feature extraction. However, in principle MIR offers the possibility to systematically explore parameter spaces informed by established psychoacoustical and cognitive models with a testable tool. For example, for higher-level acoustic features the most reasonable choice is to use window sizes of about the size of the subjective presence (Fraisse, 1982), which is about 3 s. Low-level features are usually extracted with a window size of up to 50 ms to match properties of auditory integration time (London, 2004).

Like the window size in feature extraction, there is a question of which kind of subjective evaluation is more appropriate: a global one after the stimulus has ended, or a dynamic, continuous one, concurrent to the stimulus presentation (see also Schubert, 2001). Both methods have disadvantages. For instance, music is a highly complex stimulus that evolves in time. Variation on several features is inherent to music. Hence, a global evaluation of musical attributes is a simplification of the highly dynamic process of perception. In addition, evaluation is known to change depending on the context and the serial position within a sequence of events (e.g., Holland & Lockhead, 1968). Global judgments are prone to perceptual misinterpretations (e.g., Susini, McAdams, & Smith, 2002) and heuristics; peaks and endings may be weighted differently than other parts (Kahnemann, 2000). However, a continuous evaluation of a dynamic stimulus is problematic as well. Evaluation is required concurrently to perception, including switching back and forth between different cognitive processes. This is challenging not only for participants. There is also a time delay between perceiving a feature and evaluating it, which might not be constant across the rating episode (Metallinou & Narayanan, 2013). Inter- and intra-individual differences in this delay make an exact mapping difficult. From a statistical point of view, the sequential dependency between ratings adds another difficulty. In our study, we decided on a simplistic approach of comparing a post hoc discrete evaluation with an average of windowed feature extraction. We also decided for excerpts of up to 61 s length in order to map rather naturalistic listening experiences rather than creating extremely short excerpts that might reduce the complexity of the percept to a minimum. One needs to keep in mind that, given the fuzziness of measurements, their relations are unlikely to be perfect. But though both measures might be more or less rough, we were indeed able to uncover relations (e.g., correlation up to $|r| = .66$), supporting the validity of our account.

Lastly, when applying MIR to describe the musical characteristics of a sample in order to draw general conclusions about music or musical processing, the selection of music relies on the experimenter. As such it is inherently subjective. That is, MIR ideally provides a more objective description of musical characteristics, but this does not ensure generalizability. Similarly, our results do not generalize to all studies carried out with the MIRtoolbox. For example, we raised concerns about the *mirtempo* function. Arguably, the function might work reliably with music that is relatively easily structured with a clear and regular beat. Again, a systematic evaluation of MIRtoolbox for music perception should resolve the question of how the music sample impacts results.

One sample-related caveat to keep in mind is that music preferences might shape evaluations, even for basic features like loudness and pitch. For instance, a metal fan might rate the loudness of metal pieces differently than a pop fan. In addition, in samples consisting of very heterogeneous musical styles, compression will likely play a role for music perception. A more rigorous account would be to select pieces from within a style and with similar compression (e.g., Gingras et al., 2014) in order to learn and perhaps experimentally test for certain relations between acoustic and perceptual features.

There are some details in our study design that need to be discussed; for example, the treatment of sound intensity. We deliberately chose music from very different styles that also deployed different recording and compression methods. As too big loudness differences would potentially have confounded judgments, we first applied a normalization algorithm, and readjusted the excerpts due to overcorrection of the normalization algorithm (e.g., for sparse and soft music or for very dense and loud excerpts). The normalization was not applied with the aim of equalizing loudness, as loudness was still one important feature implemented in the analysis. The final adjustment was necessary to retain the character of the musical excerpts. One critical point might be that volume was then self-chosen by participants: that will affect results because there is no linear relationship between amplitude and loudness (Glasberg & Moore, 2002) and sound intensity may have an important effect on subjective emotion ratings. We accepted this shortcoming because we wanted our participants to feel as comfortable as possible with the volume. Feeling uneasy or angry due to a inadequate volume setting would have induced emotions and in doing so likely would have confounded the emotion ratings. An additional shortcoming was that two participants indeed decided to change the volume halfway through the study. We decided to keep those participants in the data set, even though their evaluation was then expected to be noisier than for the other participants. In general, even though there were differences in volume and thus sound intensity between participants and also within two participants, our core interest was in comparing acoustic and perceptual features, which were in the end based on one and the same stimulus.

Another critical point might be the number of post-stimulus evaluations and the total duration needed to answer them. The individual mean summed duration to answer all items ranged from 65 to 150 s. The serial order of items was fixed to increase familiarity with the procedure and thereby facilitate the task. It was important to us to apply a within-subject design, with all participants listening to all sound excerpts and rating all sections of evaluations. Therefore, memory errors will have contributed to the evaluations as noise, and even more so the later the position of the rating was in the series (e.g., for the emotional ratings). The alternatives would be to give up the within-subject design, which adds between-participants variance, or to give participants the possibility to re-run the audio file during the evaluation phase. The latter would not ensure that participants used the re-run possibility, and in the end, we accepted this shortcoming. One might argue that we were nevertheless able to reproduce key findings in the literature.

Generally, MIR offers an exciting and promising development for music research. In particular, the MIRtoolbox is an excellent and easy-to-use analysis program. MIR supports a way to objectively describe a stimulus sample and poses interesting research questions related to the processing of musical characteristics. Interdisciplinary exchange between informatics, acoustics, musicology, and psychology is desirable for further developments and specifically for a broader evaluation and a better understanding of the available tools.

Author Note

We thank Freya Materna, Claudia Lehr, Jonas Schändlinger, and Jens Fünderich for their help with data collection, and Dirk Lange, Christoph Claßen, and Hans-Joachim Maempel for establishing contact to sound engineers, as well as our participants. In addition, we thank two anonymous reviewers and the reviewer Anders Friberg for their excellent, elaborate, and thorough reviews and lively discussions throughout the review process.

Correspondence concerning this article should be addressed to Elke B. Lange and Klaus Frieler, Max Planck Institute for Empirical Aesthetics, Grueneburgweg 14, 60322 Frankfurt a.M., Germany. E-mail: elke.lange@aesthetics.mpg.de

References

Ahlbäck, S. (2004). *Melody beyond notes: A study of melody cognition* (Unpublished doctoral dissertation). University of Göteborg, Sweden. Retrieved from: http://www.uddatoner.com/mer/MBN_ladda_ner/files/MelodyBeyondNotes.pdf

Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception, 27*, 223–241. DOI: 10.1525/Mp.2009.27.3.223

Alluri, V., Toiviainen, P., Jaaskelainen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *Neuroimage, 59*, 3677–3689. DOI: 10.1016/j.neuroimage.2011.11.019

Balkwill, L. L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception, 17*, 43–64. DOI: 10.2307/40285811

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 and Emotion, 19*, 1113–1139. DOI: 10.1080/02699930500204250

Böck, S., Krebs, F., & Widmer, G. (2015). Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In M. Müller & F. Wiering (Eds.), *Proceedings of the 16th International Society for Music Information Retrieval Conference* (pp. 625–631), Malaga, Spain. Retrieved from: http://www.mirlab.org/conference_papers/International_Conference/ISMIR%202015/website/articles_splitted/196_Paper.pdf

Bowling, D. L., Sundararajan, J., Han, S., & Purves, D. (2012). Expression of emotion in Eastern and Western music mirrors vocalization. *PLoS One, 7*(3), e31942. DOI: 1371/journal.pone.0031942

Bresin, R., & Friberg, A. (2011). Emotion rendering in music: Range and characteristic values of seven musical variables. *Cortex, 47*, 1068–1081. DOI: 10.1016/j.cortex.2011.05.009

Bronner, K., Frieler, K., Bruhn, H., Hirt, R., & Piper, D. (2012). What is the sound of a citrus? Research on the correspondences between the perception of sound and flavour. In E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Eds.), *Proceedings of the 12th International Conference of Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)* (pp. 142–148). Thessaloniki, Greece.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In A. del Bimbo, S.-F. Chang & A. Smeulders (Eds.), *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1467-1468). Firenze, Italy. DOI: 10.1145/1873951.1874248

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE, 96*, 668–696. DOI: 10.1109/JPROC.2008.916370

Cespedes-Guevara, J., & Eerola, T. (2018). Music communicates affects, not basic emotions - A constructionist account of attribution of emotional meanings *to music. Frontiers in Psychology, 9*, 215. DOI: 10.3389/fpsyg.2018.00215

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion, 11*, 921–937. DOI: 10.1037/a0024700

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition and Emotion, 27*, 658–684. DOI: 10.1080/02699931.2012.732559

Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition, 80*(3), B1–B10. DOI: 10.1016/S0010-0277(00)00136-0

EBU Technical Committee (2011). *Loudness recommendation EBU R128*. European Broadcasting Union, https://tech.ebu.ch/loudness

Eerola, T. (2011). Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres. *Journal of New Music Research, 40*, 349–366. DOI: 10.1080/09298215.2011.602195

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology, 4,* 487. DOI: 10.3389/fpsyg.2013.00487

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In K. Hirata, G. Tzanetakis & K. Yohii (Eds.), *Proceedings of the 10th International Society for Music Information Retrieval Conference* (pp. 621–626). Kobe, Japan. Retrieved from: http://ismir2009.ismir.net/proceedings/PS4-8.pdf

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music, 39*, 18–49. DOI: 10.1177/0305735610362821

Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *Journal of the Acoustical Society of America*, *133*, 389–404. DOI: 10.1121/1.4770244

Elowsson, A., & Friberg, A. (2015). Modeling the perception of tempo. *Journal of the Acoustical Society of America, 137*, 3163–3177. DOI: 10.1121/1.4919306

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.). *Psychology of music* (pp. 149–180). New York: Academic Press.

Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., & Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *Journal of the Acoustical Society of America, 136*, 1951–1963. DOI: 10.1121/1.4892767

Gabrielsson, A., & Lindström, E. (2001). The role of structure in the musical expression of emotions. In K. N. Juslin & J. A. Sloboda (Eds.). *Music and emotion* (pp. 377–400). Oxford, UK: Oxford University Press.

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. In K. N. Juslin & J. A. Sloboda (Eds.). *Handbook of music and emotion: Theory, research, applications* (pp. 367–400). Oxford, UK: Oxford University Press.

Gingras, B., Marin, M. M., & Fitch, T. (2013). Beyond intensity: Spectral features effectively predict music-induced subjective arousal. *The Quarterly Journal of Experimental Psychology, 67*, 1428–1446. DOI: 10.1080/17470218.2013.863954

Glasberg, B. R., & Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, *50*, 331–342.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics, 13*, 70–84.

Hevner, K. (1935). The affective character of the major and minor modes in music. *American Journal of Psychology, 47*, 103–118. DOI: 10.2307/1416710

Hevner, K. (1937). The affective value of pitch and tempo in music. *American Journal of Psychology, 49*, 621-630. DOI: 0.2307/1416385

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception and Psychophysics, 3*, 409-414.

Hwang, F. C., Wang, J., Chung, P. C., & Yang, C. F. (2013). Detecting emotional expression of music with feature selection approach. *Proceedings of the 1st International Conference on Orange Technologies* (pp. 282–286). Tainan, Taiwan. DOI: 10.1109/ICOT.2013.6521213

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception, 23*, 319–329. DOI: 10.1525/mp.2006.23.4.319

Juslin, P. N. (1997). Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy. *Musicae Scientiae, 1*, 225–256.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1797–1812. DOI: 10.1037//0096-1523.26.6.1797

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews, 10*, 235–266. DOI: 10.1016/j.plrev.2013.05.008

Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae, 5*(1_suppl), 63–122.

Juslin, P. N., & Lindström, E. (2011). Musical expression of emotions: Modelling listeners' judgments of composed and performed features. *Music Analysis, 29*, 334–364. DOI: 10.1111/j.1468-2249.2011.00323.

Kahneman, D. (2000). Evaluation by moments, past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, values and frames* (pp. 693–716). Cambridge, UK: Cambridge University Press.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J-S-, Yazsani, A., Ebrahimi, T., et al. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing, 3*, 18–31. DOI: 10.1109/T-AFFC.2011.15

Korhonen, M. D., Clausi, D. A., & Jernigan, M. E. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 36*, 588–599. DOI: 10.1109/Tsmcb.2005.862491

Kreutz, G., & Lotze, M. (2007). Neuroscience of music and emotion. In W. Gruhn, & F. H. Rauscher (Eds.), *Neurosciences in music pedagogy* (pp. 143–167). New York: Nova Science Publishers Inc.

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30*, 61–70.

Kumar, N., Kumar, R., & Bhattacharya, S. (2015). Testing reliability of Mirtoolbox. In P. Karthigaikumar, C. Arulmurugan & T. Manoj Kumar (Eds.), *Proceedings of the 2nd International Conference on Electronics and Communication Systems* (pp. 704-709). Coimbatore, India. DOI: 10.1109/ECS.2015.7125004.

Lange, E. B., Zweck, F., & Sinn, P. (2017). Microsaccade-rate indexes absorption by music listening. *Consciousness and Cognition, 55*, 59–78. DOI: 10.1016/j.concog.2017.07.009.

Lartillot, O. (2014). *MIRtoolbox 1.6.1 user's manual.* Aalborg, Denmark: Aalborg University. Retrieved from https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In S. Marchand (Ed.), *Proceedings of the 10th International Conference on Digital Audio Effects* (pp. 237–244). Bordeaux, France. Retrieved from http://cms2.unige.ch/fapse/neuroemo/pdf/ArticleLartillot2007Bordeaux.pdf

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of musical affect using a combination of acoustic structural cues. *Journal of New Music Research, 34*, 39–67. DOI: 10.1080/09298210500123978

Li, T. & Ogihara, M. (2006). Towards intelligent music information retrieval. *IEEE Transactions on Multimedia, 8*, 564–574. DOI: 10.1109/TMM.2006.870730

London, J. (2004). *Hearing in time: Psychological aspects of musical meter.* New York: Oxford University Press.

Lu, L., Liu, D., & Zhang, H.-J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing, 14*, 5–18. DOI: 10.1109/TSA.2005.860344

Metallinou, A., & Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In R. Chellappa, X. Chen, Q. Ji, M. Pantic, S. Sclaroff & L. Yin (Eds.), *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (pp. 1–8), Shanghai, China. DOI: 10.1109/FG.2013.6553804

Mion, L., & De Poli, G. (2008). Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing, 16*, 458–466.

Moffat, D., Ronan, D., & Reiss, J.D. (2015). *An evaluation of audio feature extraction toolboxes.* In P. Svensson & U. Kristiansen (Eds.), *Proceedings of the 18th International Conference on Digital Audio Effects* (pp. 277–283, Trondheim, Norway. Retrieved from: http://www.ntnu.edu/documents/1001201110/1266017954/DAFx-15_submission_43_v2.pdf

Müllensiefen, D., Gingras, B., Musil, J., & Stewart L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE, 9*(2), e89642. DOI: 10.1371/journal.pone.0089642

Panda, R., Rocha, B., & Paiva, R. P. (2015). Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence, 29*, 313–334. DOI: 10.1080/08839514.2015.1016389

Peirce, J.W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. DOI: 10.1016/j.jneumeth.2006.11.017

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition, 68*, 111–141. DOI: 10.1016/S0010-0277(98)00043-2

Poikonen, H., Alluri, V., Brattico, E., Lartillot, O., Tervaniemi, M., & Huotilainen, M. (2016). Event-related brain responses while listening to entire pieces of music. *Neuroscience, 312*, 58–73. DOI: 10.1016/j.neuroscience.2015.10.061

Reinoso Carvalho, F., Van Ee, R., Rychtarikova, M., Toulhafi, A., Steenhaut, K., Persoone, D., & Spencer, C. (2015). Using sound-taste correspondences to enhance the subjective value of tasting experiences. *Frontiers in Psychologie, 6*, 1309. DOI: 10.3389/fp-syg.2015.01309

Rigg, M. (1937). Musical expression: an investigation of the theories of Erich Sorantin. *Journal of Experimental Psychology, 21*, 442–455. DOI: 10.1037/h0056388

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called *emotion*: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*, 805–819. DOI: 10.1037/0022-3514.76.5.805

Saari, P., Eerola, T., & Lartillot, O. (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio Speech and Language Processing, 19*, 1802–1812. DOI: 10.1109/Tasl.2010.2101596

Schedl, M., Eghbal-Zadeh, H., Gómez, E., & Tkalčič, M. (2016). An analysis of agreement in classical music perception and its relationship to listener characteristics. In M. I. Mandel, J. Devaney, D. Turnbull & G. Tzanetakis (Eds.), *Proceedings of the 17th International Society for Music Information Retrieval Conference* (pp. 578–583). New York, USA. Retreived from http://m.mr-pc.org/ismir16/website/articles/260_Paper.pdf

Schellenberg, E. G., Krysciak, A. M., & Campbell, R. J. (2000). Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception, 18*, 155–171. DOI: 10.2307/40285907

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research, 33*, 239–251. DOI: 10.1080/0929821042000317822

Scherer, K.R., & Oshinsky, J.S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331–346. DOI: 10.1007/BF00992539

Schubert, E. (2001). Continuous self-report methods. In K. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion* (pp. 223–253). Oxford, UK: Oxford University Press.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception, 21*, 561–585. DOI: 10.1525/mp.2004.21.4.561

Susini, P., McAdams, S., & Smith, B. K. (2002). Global and continuous loudness estimation of time-varying levels. *Acta Acustica United with Acustica, 88,* 536-548.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*, 293–302. DOI: 10.1109/Tsa.2002.800560

Vuoskoski, J. K., & Eerola, T. (2011). The role of mood and personality in the perception of emotions represented by music. *Cortex, 47*, 1099–1106. DOI: 10.1016/j.cortex.2011.04.011

Wang, Q.(J.), Woods, A. T., & Spence, C. (2015). "What's your taste in music?" A comparison of the effectiveness of various soundscapes in evoing specific tastes. *i-Perception, 6*(6), 1–23. DOI: 10.1177/2041669515622001

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology, 4,* 292. DOI: 10.3389/fpsyg.2013.00292

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 16*, 448–457. DOI: 10.1109/TASL.2007.911513

Appendix

List of Musical Stimuli

Musician/Composer	Title	Musical Style	SP OffSet in dB
Neurosis	Casting Of The Ages	Metal	3
John Fahey	Steamboat Gwine Round de Bend	Folk	0
Fleetwood Mac	Albatross	Blues	0
Billy Strange	If I Were Free	Country	0
Eckbank Zithermusi	Simmerl	German folk music	−5
Four Tet	She Moves She	Electronica	−3
DJ CAM	Mad Blunted Jazz	Hip hop	0
Miles Davis	So What	Jazz	0
Euge Groove	Movin On	Pop	0
Ansel Collins & Dalton Browne	West Of The Sun	Reggae	0
Jeff Beck	Serene	Rock	0
King Curtis	The Stranger *(No Strings)*	Soul	0
Zakir Hussain	Zakir	World music	−5
György Ligeti	Poème Symphonique for 100 Metronomes	Classical music	−5
Johannes Brahms	Violin concerto in D major, Op. 77; II Adagio	Classical music	0
Big Bill Broonzy	Hey Hey	Blues	0
Chet Atkins	Caravan	Country	0
Jimmy Bryant and Speedy West	Frettin' Fingers	Country	0
Blasmusik Oberstaufen	Castaldo-Marsch	German folk music	3
Roman Flügel	Gehts noch (Bsherry RMX)	Electronica	3
Duke Ellington	Take The A Train	Jazz	0
Sonny Clark Trio	Junka	Jazz	−3
The Octopus Project	Truck	Pop	0
Boards of Canada	Dayvan Cowboy	Rock	0
Joe Satriani	Satch Boogie	Rock	3
Cliff Nobles	The Horse	Soul	0
Commodores	Machine Gun	Soul	0
Äl Jawala	Unzer Toirele	World music	0
Duofel	No Caminho das Pedras	World music	−3
Ludwig van Beethoven	Symphony No. 7 in A major, Op. 92, I.Poco sostenuto - Vivace	Classical music	5
Henry Mancini	Reflection	Blues	0
Koflgschroa	Sofia	German folk music	0
Trentemøller	Nightwalker	Electronica	0
DJ Krush	Kemuri Untouchable Mix	Hip hop	3
Bohren & der Club of Gore	Destroying Angels	Jazz	−3
Thelonious Monk	Round Midnight	Jazz	0
Erik Satie	Trois Gymnopédies No 1	Classical music	−5
Death	Voice Of The Soul	Metal	0
Kenny G	Songbird	Pop	0
Joe Gibbs & The Professionals	Third World	Reggae	−3
Funkadelic	Maggot Brain	Rock	0
Adrian Younge	Shot Me In The Heart (Instrumental)	Soul	0
Anouar Brahem	Vague E la nave va	World music	−3
Dhafer Youssef	Ascetic mood	World music	−3
Robert Wagner	Tristan and Isolde, Act 3	Classical music	0
Stephan Micus	Blossoms in the Wind	World music	0
John Fahey	Wine and Roses	Folk	0
Justice	Stress	Electronica	3
Mouse on Mars	Hi Fienilin	Electronica	3
DJ Shadow	Stem Long Stem	Hip hop	3
Flying Lotus	Riot	Hip hop	5
Tied & Tickled Trio and Billy Hart	Lonely Woman_Exit La Place Demon_The Electronic Family	Jazz	−3

(continued)

Appendix **(continued)**

Musician/Composer	Title	Musical Style	SP OffSet in dB
Dmitri Schostakowitsch	Piano trio No. 2, Op. 67, Allegretto	Classical music	3
Voodoo Glow Skulls	Los Hombres No Lloran	Reggae	0
Kyuss	Jumbo Blimp Jumbo	Rock	3
Russian Circles	Lebaron	Metal	5
James Brown	Devils Den Live	Soul	3
Krzysztof Penderecki	Threnody for the Victims of Hiroshima	Classical music	0
Apocalyptica	Hyperventilation	Metal	3
Franz Liszt	Symphonic Poems: Tasso. Lamento e Trionfo	Classical music	5