Anders Friberg · Erwin Schoonderwaldt · Anton Hedblad

Perceptual Ratings of Musical Parameters

1. Introduction

This study started from thoughts about what we really hear when we listen to music. Or to be more specific: can we get a better understanding of which features we are extracting on an intermediate perception level? The orientation here is not from an expert point of view that has learned to cognitively extract advanced music theoretic concepts such as implied harmony or complex rhythms. Rather, we would like to consider what we could call the musically more naive listening by laymen or during casual music listening. However, this kind of listening is not naive in a general sense since it relies on our life-long experience of sound and music exposure. One example is our ability to recognise higher-level semantic descriptions such as the emotional expression or the genre from music excerpts shorter than one second. In such a short time frame we can hear a few notes at most, thus the musical context is almost non-existing, implying that there are a number of other aspects that must be important for the overall perception.

Most models of emotion communication as well as computational models of music listening use some sort of listening model in which the incoming audio is first analysed in terms of a number of intermediate features. These features are then combined to form the final perception in terms of a high-level semantic description of the music such as genre or emotion. However, it is rare that those features are evaluated directly. Thus, the features are often considered relevant if they serve to explain the final prediction goal.

The purpose of the present study was to investigate whether mid-level features could be directly evaluated perceptually in listening experiments where subjects rate the different features on Likert scales. The following text starts with a discussion and selection of those features to be used, followed by a pilot experiment using 240 music examples and five listeners and finally a main experiment with 100 music examples and 20 listeners.

2. Which features?

It is not trivial to select a set of features since there is an abundance of proposed features in the literature. We will now consider features derived from different fields and discuss their potential relevance in terms of perception.

'Feature' is used here to denote a general attribute of a music example. Each feature can be estimated either directly from the audio signal by a signal-processing method or perceptually by listeners. We will, if necessary, call the latter 'perceptual features' to emphasize that they are based on perception and to distinguish them from their computational counterpart.

a. Music theory

Notes, rhythm, tempo, harmony etc. are the basic building blocks in music theory. Let us consider them as musical features. They all have a connection to perception but more indirectly and often from an expert point of view. For example, notes are certainly perceived as sonic events but are also coupled directly to music notation. Most studies in music perception have been using music theory as the theoretical base implicitly determining how music is organized.[1] However, when large-scale structures are analysed in a very precise way regarding harmony, grouping, and rhythm, there have been some concerns regarding the perceptual validity of such analysis. Considering the perceptual present of about five seconds forming a boundary between direct perception and higher-level cognitive processes, it is unlikely that we 'automatically' perceive large-scale structures in such a precise way without cognitive, learned processing, and memorization.[2] It is more plausible that, for example, tonality is processed in a rather sketchy manner such as that you perceive that the accompaniment is changing in pitch, or that some pitches are new compared to what was sounding before rather than the labelling used in functional harmony. This leads to a set of broad measures such as harmonic/rhythmic complexity, overall pitch or pitch direction.

b. Emotion research

In music emotion research, an important question has been the relevance of different musical features for the expression of emotions. Studies have focused on qualitative descriptions of features that were often selected in an intuitive way. For example, Kate Hevner[3] made a series of listening experiments in which music examples were manipulated with respect to different features that each could have

[1] See e. g. Diana Deutsch (ed.), *The Psychology of Music*, San Diego 1999.

[2] Eric F. Clarke, 'Rhythm and Timing in Music', in: Deutsch (ed.), *The Psychology of Music* (see note 1), pp. 473–500.

[3] Overview in Kate Hevner, 'The Affective Value of Pitch and Tempo in Music', in: *American Journal of Psychology* 49 (1937), pp. 621–630.

two values. The most important features for emotional expression were found to be mode (major/minor), tempo (high/low), pitch (high/low), rhythm (firm/flowing) and harmony (simple/complex). Alf Gabrielsson and Erik Lindström[4] made a comprehensive summary of features (or musical factors) that had been used in emotion research. They list 19 different features that showed an influence on the emotional expression in previous studies. These were essentially an extended list of the features found by Hevner. Thus, after more than 70 years of research and with a large number of subsequent studies it seems that Hevner's results to a large extent are still valid. This is remarkable considering that the cultural musical context in the 1930s was rather different. It implies that some aspects of emotion communication are rather general – a view that has been supported by multi-cultural studies.[5]

The features listed by Gabrielsson and Lindström[6] are all using simple musical categories, such as high/low pitch etc. This is not surprising given the limitations of this type of experimental setting. The stimuli need to be rather short and judgements have to be fast, encouraging the type of casual listening we discussed in the introduction. Also, most studies have been concerned with the 'perception' of emotion rather than the 'induction' of emotion.

Recently, the previous qualitative description of features was extended into more specific values and ranges in two studies.[7] The relation between features and emotional expression was investigated quantitatively either by letting musicians adjust the precise value of the features applied to different music examples or by varying each feature independently using a large number of levels each in a listening experiment with a full-factorial test design.

c. Ecological music perception

In ecological acoustics there has been an emphasis on 'everyday listening' meaning that we normally analyse sounds in our environment regarding the source properties rather than the quality of the sound itself.[8] This is quite natural considering that the human perceptual system always tries to understand and categorize sensory input. As shown in several studies we can estimate the source properties, such

4 Alf Gabrielsson and Erik Lindström, 'The Role of Structure in the Musical Expression of Emotions', in: Patrik N. Juslin and John A. Sloboda (ed.), *Handbook of Music and Emotion: Theory, Research, Applications,* New York 2010, pp. 367–400.

5 E. g. William Forde Thompson and Laura-Lee Balkwill, 'Cross-Cultural Similarities and Differences', in: Juslin and Sloboda (ed.), *Handbook of Music and Emotion* (see note 4), pp. 755 to 790.

6 Gabrielsson and Lindström, 'The Role of Structure in the Musical Expression of Emotions' (see note 4).

7 Roberto Bresin and Anders Friberg, 'Emotion Rendering in Music: Range and Characteristic Values of Seven Musical Variables', in: *Cortex* (in press); Tuomas Eerola, Anders Friberg, and Roberto Bresin, 'Emotion Perception in Music: Importance, Linearity, and Additive Effects of Seven Musical Factors' (manuscript submitted for publication).

8 E. g. William W. Gaver, 'How Do We Hear in the World?: Explorations in Ecological Acoustics', in: *Ecological Psychology* 5/4 (1993), pp. 285–313.

as identity and size, from the sound of simple objects.[9] This type of perception is evident for environmental sounds, but is the same mechanism also active in music listening? From a human voice we can estimate personal properties such as identity, distance, effort, and emotion. Similar properties, such as gender, can even be estimated from the sound of a person's footsteps.[10] Thus, from this and the emotion studies mentioned above, it is evident that to a certain extent similar features can be estimated from the sound of a musician playing an instrument.[11]

In this context it is interesting to consider 'effort' or 'energy'. The dynamic level of some instruments such as the piano seems to be perceptually estimated separately from distance (or volume). The dynamic level indirectly reflects the effort by the player. In this way one can estimate both the effort provided by the source (the pianist) and the distance to it as two separate variables. Another example is the perception of the human voice. In a study on the perception of loudness of different spoken vowels, Ladefoged and McKinney[12] found that listeners did not estimate loudness from the sound level in the room. The best correlate with perceived loudness was a combination of the subglottal pressure and the airflow corresponding directly to the physical 'work' exerted by the speaker. It was later suggested that the listeners were in fact rating the effort rather than the loudness of the voice.[13]

d. Music information retrieval

In the field of music information retrieval (MIR), the common factor of all features is that they are computational and derived from the audio signal or a symbolic (MIDI) representation. Researchers have used a large number of features derived from different research fields such as basic acoustics, psychoacoustics, or music theory.[14] They can be broadly divided in two categories: 1. Low-level short-time frame measures. These are often different spectral features such as MFCC coefficients, spectral centroid or the number of zero crossings per time unit, but also psychoacoustic measures such as roughness and loudness. 2. Mid-level features with a slightly longer analysis window. The mid-level features are often typical concepts

[9] E. g. Bruno L. Giordano and Stephen McAdams, 'Material Identification of Real Impact Sounds: Effects of Size Variation in Steel, Glass, Wood, and Plexiglass Plates', in: *Journal of the Acoustical Society of America* 119/2 (2006), pp. 1172–1181.

[10] Xiaofeng Li, Robert J. Logan, and Richard E. Pastore, 'Perception of Acoustic Source Characteristics: Walking Sounds', in: *Journal of the Acoustical Society of America* 90/6 (1991), pp. 3036–3049.

[11] See also Eric F. Clarke, *Ways of Listening: An Ecological Approach to the Perception of Musical Meaning,* Oxford 2005.

[12] Peter Ladefoged and Norris P. McKinney, 'Loudness, Sound Pressure, and Subglottal Pressure in Speech', in: *Journal of the Acoustical Society of America* 35/4 (1963), pp. 454–460.

[13] Anders Eriksson and Hartmut Traunmüller, 'Perception of Vocal Effort and Distance from the Speaker on the Basis of Vowel Utterances', in: *Perception & Psychophysics* 64/1 (2002), pp. 131–139.

[14] See e. g. Juan J. Burred and Alexander Lerch, 'Hierarchical Automatic Audio Signal Classification', in: *Journal of the Audio Engineering Society* 52/7/8 (2004), pp. 724–739.

from music theory and music perception such as beat strength, rhythmic regularity, meter, mode, harmony, and key strength. They are often modelled on so called 'ground-truth' data, i. e. a set of music examples are annotated with the 'correct' answer according to a number of judges and the feature is then optimised using these reference data. Several software packages that compute audio features such as the MIRToolbox[15] are freely available.

Extracting meaningful, perceptually based features from audio is not a simple task and will remain a major challenge in the future. Although a large variety of features has been modelled, many of them remain to be verified using perceptual data.

e. Previous work

The method of perceptual ratings has often been used for understanding how the mechanism is working, for example in terms of the underlying dimensions.[16] However, previous perceptual ratings of musical features are rare in the literature. One of the few exceptions is a study by Lage Wedin,[17] in which 15 subjects were asked to rate 13 different features concerning structure and performance aspects in 40 music examples. The overall reliability of the mean ratings across subjects was considered as being sufficient and was estimated using Ebel's intraclass correlation (r_{kk} = 0.90 to 0.98). However, some of the ratings of similar perceptual features were highly correlated, for example tempo and pulse rate (r = 0.98) or harmony (dissonant, complex–consonant, simple) and tonality (atonal–tonal) with a correlation of r = 0.95. Our conclusion from his data is that the method seems to be a good way of estimating more complex musical variations using general descriptive parameters. However, it also indicates that some subtle expected differences such as between harmony and tonality are possibly hard to distinguish using perceptual ratings.

As pointed out by Wedin, a step-wise regression from the perceptual ratings to a set of dimensions derived in another experiment was problematic, since the most prominent ratings enter a bit ad-hoc into the regression equation due to the sometimes high inter-correlations. Possibly only one of the perceptual ratings with high intercorrelation (for example tempo or pulse rate) should have been used. Unfortunately, a regression using all the musical parameters was not possible due to the insufficient number of cases (40 music examples). This is a general problem that is still relevant today, in particular considering the often even larger number of features used in computational studies.

15 Olivier Lartillot and Petri Toiviainen, 'A MATLAB Toolbox for Musical Feature Extraction from Audio', in: *Proceedings of the 10th International Conference on Digital Audio Effects 2007 (DAFx-07)*.

16 For an overview and discussion, see Stephen Handel, *Listening – An Introduction to the Perception of Auditory Events*, Cambridge/MA 1989. See also Weinzierl and Maempel in this issue.

17 Lage Wedin, 'A Multidimensional Study of Perceptual-Emotional Qualities in Music', in: *Scandinavian Journal of Psychology* 13 (1972), pp. 241–257.

Recently, using similar methods of direct perceptual ratings, Lartillot et al.[18] modelled perceived pulse clarity using perceptual ratings in combination with a selected and specifically designed set of audio features.

3. Features

The following list is a limited selection of perceptual features that we considered important mainly due to their relevance in emotion communication but also considering the ecological approach. They were selected to represent the most important and general aspects in each main variable: time (tempo, rhythm, articulation), amplitude (dynamics), pitch (melodic and harmonic aspects) and timbre. Due to experimental constraints the number was limited to nine basic features plus two or four emotion labels.

a. Tempo

'Speed' (slow–fast)

Indicates the general speed of the music disregarding any deeper analysis such as the tempo. Following the ecological approach, our hypothesis is that this general speed perception is a more relevant parameter for determining emotional character than tempo and note density.[19] In a recent study[20] it was found that note density was in many cases restricted to a small range for a certain emotional expression across different music examples. Speed may presumably be modelled using a combination of tempo and rhythm features.

b. Rhythm

'Rhythmic clarity' (flowing–firm)

Indicates how well the rhythm is accentuated disregarding the actual rhythmic pattern. Similar measures have been used in several studies. Gabrielsson[21] made a factor analysis of the perceived aspects of different rhythm patterns and found that one dimension for describing rhythm could be described as 'marked basic pattern': 'This refers to the perceptual prominence of a basic pattern, irrespective of which type the pattern is.'[22] Similar descriptions have been used by Hevner[23] ('rhythm',

[18] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and José Fornari, 'Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization', in: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Philadelphia 2008, pp. 521–526.

[19] For a discussion of tempo and note density in this context, see Alf Gabrielsson, *Studies in Rhythm*, Phil. Diss. Uppsala University 1973.

[20] Bresin and Friberg, 'Emotion Rendering in Music' (see note 7).

[21] Alf Gabrielsson, 'Adjective Ratings and Dimension Analysis of Auditory Rhythm Patterns', in: *Scandinavian Journal of Psychology* 14 (1973), pp. 244–260.

[22] Gabrielsson, *Studies in Rhythm* (see note 19), p. 9.

[23] Hevner, 'The Affective Value of Pitch and Tempo in Music' (see note 3).

flowing–firm), Wedin[24] ('rhythm', vague–outstanding; 'rhythm articulation', firm–fluent) and Lartillot et al.[25] ('pulse clarity', unclear–clear).

'Rhythmic complexity' (simple–complex)

This is a natural companion to rhythmic clarity and presumably an independent rhythmic measure, which however may co-vary in some musical examples. Thus, a typical backbeat in rock music would be described as firm and simple. However, a straight sequence of eighth notes on a hi-hat could be described as simple with a flowing character.

c. Articulation

'Articulation' (staccato–legato)

Articulation has been verified in a number of studies as relevant for emotion communication.[26]

d. Dynamics

'Dynamics' (soft–loud)

The intention is to estimate the played dynamic level disregarding listening volume. Dynamics have been used in a large number of emotion studies using a variety of labels. This measure is in line with the ecological perspective where presumably the source properties such as the energy used to excite the instrument is an important feature.

e. Tonality

'Modality' (minor–major)

Modality has been shown to be important for the emotion recognition in a large number of studies. It is remarkable considering that this is largely related to the Western music tradition and apparently learned from listening. We consider modality here as a continuous scale ranging from minor to major.

'Overall Pitch' (low–high)

The simplest possible representation of melody, i. e. its general pitch height.

'Harmonic complexity' (simple–complex)

It has been used in several emotion studies, i. e. Hevner[27] ('harmony', simple–com-

[24] Wedin, 'A Multidimensional Study of Perceptual-Emotional Qualities in Music' (see note 17).
[25] Lartillot et al., 'Multi-Feature Modeling of Pulse Clarity' (see note 18).
[26] See summary in Gabrielsson and Lindström, 'The Role of Structure in the Musical Expression of Emotions' (see note 4).
[27] Hevner, 'The Affective Value of Pitch and Tempo in Music' (see note 3).

plex) and Wedin[28] ('harmony', dissonant, complex – consonant, simple). The relation to the musical structure might reflect for example the amount of chord changes and deviations from a certain key scale structure. In this context it is possibly the most difficult feature to rate, demanding some knowledge of music theory. Thus, we would expect less consistent results with non-expert listeners.

f. Timbre

Timbre is a dimension that has been less investigated in emotion studies. However, it is likely to play an important role for the expression nowadays reflected by the focus on timbre and sound in recent popular music production. It is also highly relevant from an ecological perspective. Vinoo Alluri and Petri Toiviainen[29] studied the perception of timbre in a polyphonic context. Eight bipolar timbre scales were reduced to the three dimensions 'activity' (soft–hard, strong–weak, high energy–low energy), 'brightness' (colourless–colourful, dark–bright) and 'fullness' (empty–full) using factor analysis.

'Brightness' (dark–bright)

Brightness has been suggested in a large number of studies as possibly the most important timbre dimension. It is primarily associated with the amount of spectral energy in the treble range of the spectrum.

g. Emotion

In addition, for testing purposes, different measures mainly reflecting the emotional expression were selected:

'Happiness', 'Anger', 'Sadness', and 'Tenderness'

These are four discrete commonly used emotions that also may represent each quadrant in the energy-valence space. Used in the pilot experiment.

'Energy' (low–high)
'Valence' (negative–positive)

These are two dimensions of the commonly used dimensional model of emotion. The energy dimension is often labelled 'activity' or 'arousal'. Energy was chosen since it is also motivated from the ecological perspective. Used in the main experiment.

[28] Wedin, 'A Multidimensional Study of Perceptual-Emotional Qualities in Music' (see note 17).

[29] Vinoo Alluri and Petri Toiviainen, 'In Search of Perceptual and Acoustical Correlates of Polyphonic Timbre', in: *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)*, Jyväskylä, Finland.

4. Ringtone data base

The stimuli were constructed from a MIDI database of 242 popular ringtones used in a previous experiment.[30] The ring tones were randomly selected from a large commercial database consisting of popular music of various styles. They were in a majority of cases instrumental polyphonic versions of the original songs. The average duration of the ringtones was about 30 s. Ringtones of longer duration (>40 s) were cropped by editing the original MIDI files, making sure that the endings would sound natural. Furthermore, a one-second lead was added to all MIDI files, in order to give the used synthesizer enough time to carry out program changes at the beginning of the files. The MIDI files were converted to audio (wav) by a batch process in Matlab. To obtain a high-quality audio rendering, the files were played on an external Roland JV-1010 MIDI synthesizer via Windows Media Player (v 11) and recorded as separate wav files using a professional sound card (RME, HDSP 9632). For the preparation of the final stimuli, the wav files were normalized according to the loudness standard specification ITU-R BS. 1770. Finally, the sound files were trimmed at their beginnings and endings so that there was a constant short delay before the first sound.

5. Pilot experiment

A pilot listening experiment was conducted to obtain a first classification of the stimuli and testing of the method. The results were used for selection of a subset of stimuli for the main listening experiment. The subjects were asked to rate a subset of six structural features from the list described above and four basic emotions using continuous sliders. The features were 'speed', 'rhythmic clarity', 'articulation', 'dynamics', 'modality' and 'brightness'. The four basic emotions were 'happiness', 'sadness', 'anger' and 'tenderness'. The experimental interface was programmed in Skatta (version 1.1 beta 6), a multiplatform system for designing and running audio and visual perception tests.[31]

a. Method

The pilot listening test was performed in a low-reverberant recording studio. The stimuli were played in stereo via two monitor loudspeakers (Genelec 1031A) placed at a distance of about 2 m from the listener at head height. The sound level was calibrated using a noise signal with a spectral slope of –6 dB per octave obtained

[30] Anders Friberg and Sven Ahlbäck, 'Recognition of the Main Melody in a Polyphonic Symbolic Score Using Perceptual Knowledge', in: *Journal of New Music Research* 38/2 (2009), pp. 155–169.

[31] John Lindberg, *Skatta – A Multiplatform System for the Design and Running of Audio and Visual Perception Tests*, Master thesis, Speech, Music and Hearing, KTH Stockholm (in preparation).

by applying a low-pass filter to white noise.[32] This spectral shape corresponded roughly to the spectral content of the stimuli. The amplitude of the noise signal was normalized to the same perceptual loudness value as the stimuli. The playback volume was adjusted so that the noise level corresponded to 75 dB(A) at listener position.

Five subjects (two students and three researchers) rated all features on a continuous scale for each ring tone. For each ring tone, a screen was presented containing a play button and a horizontal slider for each of the ten ratings. The end positions of each slider were labelled according to the description of the features above. The ratings were coded from 1 to 100 for all scales. The experiment was divided in four blocks that took approximately 50 min each to complete.

b. Results

The analysis of the results will focus on different measures that could indicate the easiness and ability to rate different features. Table 1 shows the cross-correlations between feature ratings. This is an indication as to whether the subjects were able to perceive the different features in an independent way. Although many of the correlations are significant, the values are not alarmingly high, indicating that the features were to a large extent independently rated. A moderate amount of correlation would also be expected in a database of existing music. For example, music with fast tempo (speed) may also have a high dynamic level. Note that both 'modality' and 'brightness' exhibit a large degree of independence.

	Speed	Rhythmic clarity	Articulation	Dynamics	Modality
Rhythmic clarity	0.39***				
Articulation	0.45***	0.48***			
Dynamics	0.40***	0.39***	0.39***		
Modality	0.00	–0.04	–0.04	–0.05	
Brightness	0.16*	0.11	0.10	0.41***	0.35***

Table 1: Cross-correlations between feature ratings (averaged over subjects) in the pilot experiment, p-values: *<0.05; **<0.01, ***<0.001

The mean inter-subject correlation and Cronbach's alpha was used to assess the agreement among raters, see Table 2. It is usually recommended that the alpha value should be at least 0.7. However, in this case there are a few items (raters) with lower alpha values. As seen in the Table, the agreement was fairly high for 'speed',

[32] Single-pole low-pass filter, $H(z) = 1/(1-0.95z^{-1})$, cut-off frequency 359 Hz.

'articulation', and 'modality'. This was also reflected in the pairwise inter-subject correlations which all were positive and significant. The other three perceptual features, 'rhythmic clarity', 'dynamics' and 'brightness' obtained more modest agreement. The pairwise inter-subject correlations for these features were either relatively high or relatively small, indicating that different judgement strategies may have been used.

Ratings	Mean inter-subject correlation	Cronbach's alpha
Speed	0.66	0.90
Rhythmic clarity	0.22	0.58
Articulation	0.45	0.80
Dynamics	0.26	0.62
Modality	0.42	0.80
Brightness	0.30	0.69
Happiness	0.53	0.85
Sadness	0.40	0.78
Anger	0.57	0.87
Tenderness	0.57	0.85

Table 2: Mean inter-subject correlations and Cronbach's alphas for the five raters in the pilot experiment

The inter-rater agreement was good for the four emotions. Also, the emotions could be rather well estimated using linear regression with the feature ratings as the independent variables. The adjusted R^2 were 0.74 for 'happiness', 0.56 for 'sadness', 0.62 for 'anger' and 0.59 for 'tenderness'. This data would thus suggest that the emotions were appropriately selected. However, informal reports by the subjects indicated that the rating of the discrete emotions was difficult for many music examples in this database. The chosen emotions were reported less suited to describe the music in many cases. Therefore, a decision was made to use the energy-valence measure in the main experiment. This measure can be considered less biased since it allows for the categorization of a large number of discrete emotions.[33]

[33] James A. Russell, 'A Circumplex Model of Affect', in: *Journal of Personality and Social Psychology* 39/6 (1980), pp. 1161–1178.

6. Stimuli selection

The ratings from the pilot experiment were used to make an effective selection of stimuli for the main listening test. A major problem with arbitrarily chosen music collections like this is that the features may have rather skewed distributions, which will affect any statistical analysis such as multiple regression. Informal listening to the ringtones revealed that the range of several features was rather limited in this set. For example, many of the melodies were played at approximately the same medium tempo. Thus, the reduction of the set was done in order to keep the range in each feature while reducing the number of cases in the middle. The first attempt to reduce the ringtone set was done by trying to also enhance the interdependence between all features. In order to achieve this, the rating space of the combined features was quantized by dividing the rating scales of all individual features into two or three categories. The rating scale of speed was divided into three categories with the 33 and 66 percentiles of the mean rating distribution as boundaries; the other five features were subdivided into two categories with the median as boundary, yielding a subdivision of the total rating space in 96 cells. It turned out that we got a reasonable number of cases in each category. However, further analysis using histograms of each feature rating revealed that the resulting set had a smaller variation in each feature. Therefore, the final selection criterion was simply to select a fixed number of music examples from the extreme ratings (high and low) of each feature. This resulted in a number of duplicates that were removed. The reason was that a music example with an extreme value in one feature also had an extreme value in other features as well (e. g. high tempo and high dynamics). The selected number of music examples with extreme values for each feature was varied on a trial-and-error basis until exactly 100 cases remained. This constituted the final set for the main experiment.

7. Main experiment

In the main experiment, all nine of the features and the two-dimensional emotion labels energy and valence listed above were used and the music examples were restricted to the selected 100 cases.

a. Subjects

There were 20 subjects with an average age of 30 years (range 18–55) consisting of seven women and 13 men. Most of them were students at KTH with some musical experience in playing an instrument. They reported that they listened to music 15 hours a week on average (range 3–40). All of them played one or several instruments and the average number of years that they played their main instrument was 14 years (range 3–45). Six subjects reported minor hearing impairments.

According to the questionnaires, these impairments were rather small and were not considered crucial for this specific experiment. Thus, all subjects were kept in the study.

b. Procedure

The experimental conditions (room, loudspeakers, calibration) were the same as in the pilot experiment except that we used another, slightly more reverberant room. The whole test was done in one session and the subjects were free to take breaks at any time. The total duration varied from about 1–2.5 hours depending on the subject. They filled in a questionnaire regarding musical experience and were reimbursed with two cinema tickets. In order to speed up the procedure, the scale was limited to nine discrete steps each represented by a box on the screen.

c. Results

The cross-correlations of the feature ratings are shown in Table 3. Only about half of the correlations were significant and did rarely exceed 0.6 (corresponding to 36 % covariation). The only exception was 'pitch' and 'brightness' with r = 0.9.

	Speed	Rhythmic complexity	Rhythmic clarity	Articulation	Dynamics	Modality	Harmonic complexity	Pitch
Rhythmic complexity	–0.09							
Rhythmic clarity	0.51***	–0.54***						
Articulation	0.57***	–0.06	0.56***					
Dynamics	0.66***	0.00	0.53***	0.57***				
Modality	0.19	–0.17	0.01	0.20	0.03			
Harmonic complexity	–0.37***	0.51***	–0.63***	–0.49***	–0.31**	–0.22*		
Pitch	–0.03	–0.04	–0.17	–0.09	0.05	0.46***	0.21*	
Brightness	0.01	–0.05	–0.16	–0.02	0.12	0.59***	0.15	0.90***

Table 3: Cross-correlations between feature ratings (averaged over subjects) in the main experiment, N = 100, p-values: * < 0.05; ** < 0.01, *** < 0.001

The agreement among subjects in terms of mean inter-subject correlation and Cronbach's alpha is shown in Table 4. The alpha values were relatively high ranging from 0.98 to 0.83. As expected the harmonic complexity obtained the lowest value in comparison (alpha = 0.83). The higher values in the main experiment as

compared to the pilot experiment were presumably due to the increased number of subjects.

The mean inter-subject correlation showed a more clear differentiation ranging from 0.71 for speed to 0.21 for harmonic complexity. A closer inspection of the pairwise inter-subject correlations showed that for some rating scales, one subject deviated substantially from all others. The most likely explanation is that they did not understand the particular task. The numbers in parenthesis in Table 4 show the results when these cases were removed. The relatively large change in the case of modality would seem to be the result of one subject having reversed the direction of the scale. Note that it did not affect the alpha values very much even though one subject apparently had a deviant strategy. It was considered problematic to trim only some of the scales so all subjects were kept in the subsequent analysis. It only marginally affected the mean values and the regression analysis below.

Ratings	Mean inter-subject correlation	Cronbach's alpha
Speed	0.71	0.98
Rhythmic complexity	0.29 (0.33)	0.89 (0.89)
Rhythmic clarity	0.31 (0.34)	0.90 (0.90)
Articulation	0.37 (0.41)	0.93 (0.93)
Dynamics	0.41 (0.44)	0.93 (0.93)
Modality	0.38 (0.47)	0.93 (0.94)
Harmonic complexity	0.21	0.83
Pitch	0.37 (0.42)	0.93 (0.93)
Brightness	0.27	0.88
Energy	0.57	0.96
Valence	0.42 (0.47)	0.94 (0.94)

Table 4: Agreement among subjects in the main experiment expressed in terms of mean inter-subject correlation and Cronbach's alpha. Numbers in parenthesis refer to trimmed data.

In the final analysis, the feature ratings were used to predict the estimated emotion ratings energy and valence. One could argue that using the ratings in this way from the same subjects is not a proper method. However, since we are using mean values and the agreement is high, it is likely that another group of subjects would produce very similar ratings. Also, the purpose is only to check if it is possible to predict the emotions using the feature ratings rather than making a detailed analysis. More specific conclusions regarding the details of different features for example,

should be treated with caution. This type of detailed estimation is more properly done using a full-factorial design of the stimuli, such as by Eerola et al.[34]

	Beta	Semipartial sr^2	p-value
Speed	0.562	0.380	0.000***
Rhythmic complexity	0.064	0.042	0.116
Rhythmic clarity	0.157	0.083	0.002**
Articulation	0.061	0.041	0.126
Dynamics	0.335	0.199	0.000***
Modality	0.135	0.089	0.001**
Harmonic complexity	0.018	0.011	0.665
Pitch	0.013	0.005	0.841
Brightness	–0.045	0.016	0.549

Table 5: Multiple regression with energy as dependent variable, N = 100, $R^2 = 0.94$, Adj. $R^2 = 0.93$, p-values: *<0.05; **<0.01, ***<0.001

	Beta	Semipartial sr^2	p-value
Speed	0.119	0.081	0.028*
Rhythmic complexity	–0.015	0.010	0.791
Rhythmic clarity	–0.019	0.010	0.779
Articulation	0.135	0.091	0.014*
Dynamics	–0.246	0.146	0.000***
Modality	0.733	0.484	0.000***
Harmonic complexity	0.020	0.013	0.722
Pitch	–0.080	0.033	0.366
Brightness	0.304	0.107	0.004**

Table 6: Multiple regression with valence as dependent variable, N = 100, $R^2 = 0.88$, Adj. $R^2 = 0.87$, p-values: *<0.05; **<0.01, ***<0.001

As shown in Table 5 and 6, both energy and valence could be predicted to a very high degree with the resulting adjusted $R^2 = 0.93$ and 0.87 respectively. In particular, the results for valence were surprising given that it has often been difficult to model in previous studies. Both variables were predicted by a small number of feature ratings with mainly a priori expected behaviour. High energy was predicted as a combina-

34 Eerola et al., 'Emotion Perception in Music' (see note 7).

tion of high speed and loud dynamics with the smaller contribution of major tonality and rhythmic clarity while positive valence was predicted mainly by low dynamics, major modality and possibly high brightness. The semipartial correlation sr^2 shown in the third column is expressing the independent contribution of each feature rating. Note that for pitch and brightness sr^2 is much smaller than the beta coefficient compared to the other features. This indicates a possible problem of covariation, also indicated by the high correlation between these two features in Table 3.

8. Conclusion and discussion

The results indicate that the perceptual evaluation of the features selected in this study is a viable method. There was a high inter-rater agreement with Cronbach's alpha ranging from 0.83 to 0.98 in the main experiment. The relatively low correlations among feature ratings indicate that the subjects were able to rate most of the different features independently. The only clear problematic case was brightness, which had a correlation with pitch of $r = 0.9$. A preliminary regression analysis indicated that the emotion ratings could be predicted by the feature ratings with an explained variation of about 93% for energy and 87% for valence.

20 subjects is rather few considering this type of perceptual rating experiment. One might conclude that since the agreement is high the number of subjects is enough for this type of experiment. In particular, the mean values of the perceptual ratings become rather stable and insensitive to a single inconsistent rater. This was also supported by the recent study of emotional expression in which two separate groups of subjects from different countries ($n = 20$ and 26) obtained very similar results.[35] On the other hand, using a few experts as in the pilot study seems to be more problematic since it is more susceptible to rater inconsistency.

It is common to use Cronbach's alpha for estimating inter-rater agreement. However, as seen in the present study, this could be problematic. First, it is sensitive to the number of subjects. Secondly, it was found to be insensitive to a rather drastic change in one of the subject's rating strategy. For investigating differences in subjects, an analysis of pairwise correlations was more useful and yielded more differentiated results.

There were clear differences in the agreement among the listeners for each perceptual feature that they rated. This may reflect the 'naturalness' of the feature. The ratings with the highest agreement were speed followed by energy and valence, modality, dynamics, pitch, articulation etc. We see that many of the rating scales that can be motivated from an ecological point of view show the highest inter-subject agreement (speed and the emotion dimensions energy and valence), while more musically specific features seem to be more difficult to rate, with the lowest agreement obtained for the complexity measures. One notable exception was modality. The modality of a piece is rather difficult to understand from a music theoretical

[35] Eerola et al., ibid.

point of view and needs a relatively long context in terms of several chords to be determined. Still, there was a good agreement for modality indicating that it was an easy task. There is close coupling between valence and modality as seen in the regression analysis above and in previous studies. Modality seems to be the most important cue for determining valence in Western music. Thus, one could speculate that our ability to determine modality developed due to the need to improve the communication of valence in music.

In future investigations we will compare these perceptual features with existing computational models and try to develop new models based on previous work such as the MIRToolbox.[36] Given that it is possible to find reliable computational models, the problem of predicting high-level semantic descriptions such as emotions could be reduced to a two-step process each with a more limited number of prediction variables. The first step is going from an acoustic surface to perceptual features and in the second step the final semantic description is predicted. This may then lead to a more precise and perceptually grounded procedure using a more focused set of features than what is common today.

Acknowledgement

This work was supported by the Swedish Research Council, Grant Nr. 2009-4285.

[36] Lartillot and Toiviainen, 'A MATLAB Toolbox for Musical Feature Extraction from Audio' (see note 15).