A COMPARISON OF PERCEPTUAL RATINGS AND COMPUTED AUDIO FEATURES

Anders Friberg
Speech, music and hearing, CSC
KTH (Royal Institute of Technology)
afriberg@kth.se

Anton Hedblad
Speech, music and hearing, CSC
KTH (Royal Institute of Technology)
ahedblad@kth.se

ABSTRACT

The backbone of most music information retrieval systems is the features extracted from audio. There is an abundance of features suggested in previous studies ranging from low-level spectral properties to high-level semantic descriptions. These features often attempt to model different perceptual aspects. However, few studies have verified if the extracted features correspond to the assumed perceptual concepts. To investigate this we selected a set of features (or musical factors) from previous psychology studies. Subjects rated nine features and two emotion scales using a set of ringtone examples. Related audio features were extracted using existing toolboxes and compared with the perceptual ratings. The results indicate that there was a high agreement among the judges for most of the perceptual scales. The emotion ratings energy and valence could be well estimated by the perceptual features using multiple regression with adj. R^2 = 0.93 and 0.87, respectively. The corresponding audio features could only to a certain degree predict the corresponding perceptual features indicating a need for further development.

1. INTRODUCTION

The extraction of features is a fundamental part of most computational models starting with the audio signal. Therefore there exists a large number of features suggested in the literature, see e.g. [1]. They can be broadly divided in two categories: (1) Low-level features often based on short-time 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 from music theory and music perception such as beat strength, rhythmic regularity, meter, mode, harmony, and key strength. They are often verified by using ground-truth data with examples annotated by experts. In addition, a third level consists of semantic descriptions such as emotional expression or genre, see Figure 1. The distinction between mid and low-level features is in reality rather vague and was made in order to point to the differences in complexity and aims.

For modeling the higher-level concepts such as emotion description or genre it is not certain that the mid-level features derived from classic music theory (or low-level features) is the best choice. In emotion research a number of more rather imprecise overall estimations has been successfully used for a long time. Examples are pitch (high/low), dynamics (high/low) or harmonic complexity (high/low), see e.g. [2,3]. This may indicate that human music perception is retrieving something other than traditional music theoretic concepts such as the harmonic progression. This is not surprising since it demands substantial training to recognize an harmonic progression but it also points to the need for finding what we really hear when we listen to music.

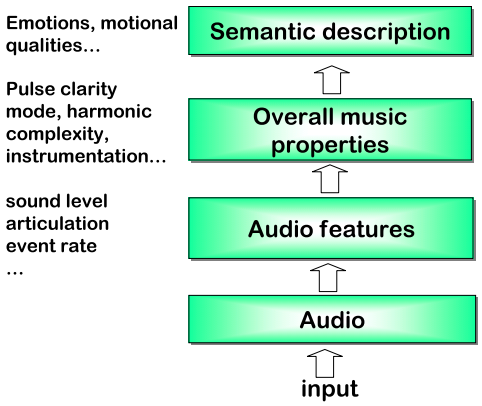

Figure 1. The different layers of music features and descriptions.

The present study is part of a series of studies in which we investigate features derived from different fields such as emotion research and ecological perception, define their perceptual values and develop computational models. We will call these perceptual features to emphasize that they are based on perception and to distinguish them from their computational counterpart.

In this paper we will report on the estimation of nine perceptual features and two emotion descriptions in a listening experiment and compare the ratings with combinations of existing audio features derived from available toolboxes.

2. RINGTONE DATABASE

The original set of music examples were 242 popular ringtones in MIDI format used in a previous experiment [4]. The ringtones 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 popular songs. The average duration of the ringtones was about 30 s. The MIDI files were converted to audio using a Roland JV-1010 MIDI synthesizer. The resulting wav files were normalized according to the loudness standard specification ITU-R BS. 1770.

In a previous pilot experiment 5 listeners, with moderate to expert music knowledge, estimated the 9 features below for all music examples, see also [5]. The purpose was both to reduce the set so that it could be rated in one listening experiment and to enhance the spread of each feature within the set. For example, it was found that many examples had a similar tempo. The number of examples was reduced to 100 by selecting the extreme cases of each perceptual rating. This slightly increased the range and spread of each variable. This constituted the final set used in this study.

3. PERCEPTUAL RATINGS

3.1 Perceptual features

This particular selection of perceptual features was motivated by their relevance in emotion research but also from the ecological perspective, see also [5]. Several of these features were used by Wedin [6] in similar experiment. Due to experimental constraints the number was limited to nine basic feature scales plus two emotion scales.

Speed (slow-fast)

The general speed of the music disregarding any deeper analysis such as the musical tempo.

Rhythmic clarity (flowing-firm)

Indication of how well the rhythm is accentuated disregarding the rhythm pattern (c.f. pulse clarity, [7]).

Rhythmic complexity (simple-complex)

This is a natural companion to rhythmic clarity and presumably an independent rhythmic measure.

Articulation (staccato-legato)

Articulation is here only related to the duration of tones in terms of *staccato* or *legato*.

Dynamics (soft-loud)

The intention was to estimate the played dynamic level disregarding listening volume. Note that the stimuli were normalized using an equal loudness model.

Modality (minor-major)

Contrary to music theory we treat modality as a continuous scale ranging from minor to major.

Overall Pitch (low-high)

The overall pitch height of the music.

Harmonic complexity (simple-complex)

A measure of how complex the harmonic progression is. It might reflect for example the amount of chord changes and deviations from a certain key scale structure. This is presumably a difficult feature to rate demanding some knowledge of music theory.

Brightness (dark-bright)

Brightness is possibly the most common description of timbre.

Energy (low-high)

Valence (negative-positive)

These are the two dimensions of the commonly used dimensional model of emotion (e.g [8]). However, the energy dimension is in previous studies often labeled activity or arousal.

3.2 Listening experiment

A listening experiment was conducted with 20 subjects rating the features and emotion descriptions on continuous scales for each of the 100 music examples (details given in [5]).

Feature	*Mean inter-subject corr.*	*Cronbach's alpha*
Speed	0.71	0.98
Rhythmic complex.	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 complex.	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 1. Agreement among the 20 subjects in terms of mean inter-subject correlation and Cronbach's alpha. A value of one indicates perfect agreement in both cases.

Could the subjects reliably estimate the perceptual features? This was estimated by the mean correlation between all subject pairs, see Table 1. In addition, for comparison with previous studies (e.g. [9]) Cronbach's alpha was also computed. The Cronbach's alpha indicated a good agreement for all ratings while the inter-subject correlation showed a more differentiated picture with lower agreement for the more complex tasks like harmonic complexity.

	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 2. Cross-correlations between rated features averaged over subjects. N=100, p-values: * < 0.05; ** < 0.01, ***<0.001.

A closer inspection of the inter-subject correlations revealed that for some features there was one subject that clearly deviated from the rest of the group. Numbers in parenthesis refer to trimmed data when these subjects were omitted. However, the original data was used in the subsequent analysis. We interpret these results as an indication that all the measures could be rated by the subjects. Although the more complex measures like harmonic complexity obtained lower agreement the mean value across subject may still be a useful estimate.

The interdependence of the different rating scales was investigated using cross-correlations shown in Table 2. As seen in the table, there were relatively few alarmingly high values. 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, which is discussed below.

It is difficult to determine the reason for the high cross-correlations in the ratings at this point since there are two different possibilities. Either there is a covariation in the music examples, or alternatively, it could be the listeners that were not able to isolate each feature as intended.

Finally, the extent to which the perceptual features could predict the emotion ratings was tested. A separate multiple regression analysis was applied for each of the emotion ratings energy and valence with all the nine perceptual features as independent variables. The energy rating could be predicted with an adj. $R^2 = 0.93$ (meaning that 93% of the variation could be predicted) with four significant perceptual features. The strongest contribution was by speed followed by dynamics, while modality and rhythmic clarity contributed with a small amount. The valence rating was predicted with an adj. $R^2 = 0.87$. The strongest contribution was by modality followed by dynamics (negative), brightness, articulation, and speed. These results were unexpectedly strong given the small number of perceptual features. However, since both the feature ratings and the emotion ratings were obtained from the same subjects this is just a preliminary observation that needs to be further validated in a future study.

3.3 COMPUTED FEATURES

Computational audio features were selected from existing toolboxes that were publicly available. Two hosts were used for computing the audio features: MIRToolbox v. 1.3.1 [10] and Sonic Annotator[1] v. 0.5. MIRToolbox is implemented in MATLAB and Sonic Annotator is a host program which can run VAMP plugins.

A list of all extracted features is shown in Table 3. Audio features were selected that we *a priori* would expect to predict a perceptual rating. Within these toolboxes we could only find *a priori* selected audio features for a subset of six perceptual ratings, namely speed, rhythmic clarity, articulation, brightness, and energy. In Table 4 below, the corresponding selected audio features are marked in grey color.

Abbreviation	**Meaning**	**Parameters**
EX - VAMP Example plugins		
EX_Onsets	Percussion Onsets	Default
EX_Tempo	Tempo	Default
MT - MIRToolbox		
MT_ASR	Average Silence Ratio	Default
MT_Bright_1.5k	Brightness	Default
MT_Bright_1k	Brightness	Cutoff: 1000 Hz
MT_Bright_3k	Brightness	Cutoff: 3000 Hz
MT_Event	Event Density	Default
MT_Mode_Best	Modality	Model: Best
MT_Mode_Sum	Modality	Model: Sum
MT_Pulse_Clarity_1	Pulse Clarity	Model: 1
MT_ Pulse_Clarity_2	Pulse Clarity	Model: 2
MT_SC	Spectral Centroid	Default
MT_SF	Spectral Flux	Default
MT_Tempo_Auto	Tempo	Model: Autocorr
MT_Tempo_Both	Tempo	Model: Autocorr & Spectrum
MT_Tempo_Spect	Tempo	Model: Spectrum
MT_ZCR	Zero Crossing Rate	Default
MZ -VAMP plugins ported from the Mazurka project.		
MZ_SF_Onsets	Spectral Flux Onsets	Default
MZ_SRF_Onsets	Spectral Reflux Onsets	Default
QM - VAMP plugins from Queen Mary.		
QM_Mode	Modality	Default
QM_Onsets	Onset detection	Default
QM_Tempo	Tempo	Default

Table 3. Overview of all computed audio features.

[1] http://www.omras2.org/SonicAnnotator

	Speed	Rhythmic complex.	Rhythmic clarity	Articulation	Dynamics	Modality	Harmonic complex.	Pitch	Bright-ness	Energy	Valence
MT_Event	0.65***	0.08	0.33***	0.52***	0.47***	-0.01	-0.27**	-0.08	-0.01	0.57***	-0.01
MT_Pulse_cla1	0.61***	-0.22*	**0.73*** **	0.69***	0.56***	0.09	-0.40***	-0.13	-0.07	0.67***	0.03
MT_Pulse_cla2	-0.08	-0.34***	0.16	0.04	-0.12	0.04	-0.11	-0.01	-0.01	-0.07	0.06
MT_ASR	0.21*	-0.03	0.44***	0.62***	0.28**	-0.03	-0.26**	-0.09	-0.13	0.33***	-0.04
MT_Bright_1k	0.26**	-0.04	0.33***	0.18	0.53***	-0.03	-0.19	0.15	0.20*	0.34***	-0.13
MT_Bright_1.5k	0.31**	-0.06	0.42***	0.28**	0.55***	-0.05	-0.22*	0.08	0.16	0.38***	-0.13
MT_Bright_3k	0.37***	-0.07	0.52***	0.40***	0.47***	-0.08	-0.26**	-0.02	0.04	0.41***	-0.15
MT_Mode_best	0.04	-0.09	-0.11	-0.11	-0.1	0.67***	-0.01	0.41***	0.51***	0	0.69***
MT_Mode_sum	-0.04	0.09	-0.05	0.04	0.11	-0.47***	0.15	-0.19	-0.25*	-0.03	-0.43***
MT_SC	0.31**	-0.12	0.45***	0.34***	0.34***	-0.1	-0.23*	-0.03	0.03	0.31**	-0.15
MT_SF	**0.72*** **	-0.03	0.66***	0.67***	0.66***	-0.03	-0.39***	-0.15	-0.08	**0.75*** **	-0.07
MT_Tempo_both	-0.11	0.17	-0.1	-0.06	0.03	-0.21*	0.15	-0.09	-0.08	-0.09	-0.22*
MT_Tempo_auto	-0.08	0.02	-0.01	0	-0.02	-0.11	0.13	-0.03	0.02	-0.08	-0.13
MT_Tempo_spect	0.02	0.12	0.04	0.08	0.07	-0.17	0.08	-0.05	-0.05	0.03	-0.16
MT_ZCR	0.43***	0.04	0.27**	0.17	0.53***	-0.02	0.01	0.14	0.15	0.45***	-0.14
QM_Onsets	**0.73*** **	0.24*	0.15	0.38***	0.50***	0	-0.13	-0.06	0	0.62***	-0.01
EX_Onsets	0.55***	0.08	0.36***	0.52***	0.34***	-0.09	-0.24*	-0.13	-0.06	0.45***	-0.06
EX_Tempo	0.15	-0.12	0	-0.05	0.08	-0.05	-0.01	-0.03	-0.04	0.06	-0.1
MZ_SF_Onsets	0.61***	0.17	0.04	0.27**	0.41***	0.06	-0.17	-0.02	0.06	0.51***	0.05
MZ_SRF_Onsets	0.64***	0.15	0.24*	0.32***	0.40***	-0.06	-0.16	-0.05	-0.02	0.55***	-0.01
QM_Mode	0	0.09	0.1	0.08	0.08	-0.58***	0.02	-0.26*	-0.39***	0.02	-0.55***
QM_Tempo	0.09	-0.21*	0.04	0.01	-0.03	0.18	-0.05	0.04	0.05	0.02	0.08

Table 4. Correlations between all perceptual ratings and computed features. Dark grey areas indicate those audio features that *a priori* were selected for predicting the perceptual ratings. N=100, p-values: * < 0.05; ** < 0.01, ***<0.001.

Each feature was computed using the default settings and in certain cases using different available models. For each sound example one feature value was obtained. All the onset measures were converted to onsets per second by counting the number of onsets and dividing by the total length of each music example. For a more detailed description, see [11].

4. COMPARISON

4.1 Correlations

The correlation between all the perceptual ratings and the computed features are shown in Table 4. There is a large number of features that correlates significantly as indicated by the stars in the table. This may serve as an initial screening where we can sort out all non-significant relations. Then the size of the correlations should be considered. According to Williams [12] a correlation coefficient between 0.4-0.7 should be considered a substantial relationship and coefficients between 0.7-0.9 should be considered a marked relationship. Following this rather *ad hoc* rule-of-thumb we note that there were only four features with a marked relationship, three of them included in the list of *a priori* selected features. These were speed and one onset model (QM_Onsets, r=0.73), speed and spectral flux (MT_SF, r=0.72), rhythmic complexity and the pulse clarity model 1 (MT_Pulse_cla1, r=0.73), and energy and spectral flux (MT_SF, r=0.75). Many of the expected relations do in fact correlate with rather high values but there are also a number of correlations that are more difficult to interpret.

As seen in Table 4, Speed is significantly correlating with many audio features. All the onset features have rather high correlations but note that none of the tempo features were significant. This result indicates that the perceived speed has little to do with the musical tempo. The results verify that the number of onsets per second is the most appropriate equivalent for perceptual speed. This was recently also verified by Bresin and Friberg [13] and Madison and Paulin [14].

Rhythmic clarity is highly correlated with pulse clarity model 1 which confirms that it is a similar measure. The pulse clarity model was developed using similar perceptual ratings [7]. Note that the second pulse clarity model is not significant and instead correlates somewhat with rhythmic complexity.

The spectral flux is an interesting case as it is correlating with almost all perceptual ratings. The high correlation with speed is not surprising since it is a measurement of spectral changes over time.

The rating of dynamics is also puzzling. As mentioned, all sound examples were normalized for equal loudness. Thus, one would possibly expect rather small variations in the ratings. Since dynamics is associated with spectral changes, the correlation with spectral features is natural. However, the strong correlations with temporal features are more difficult to interpret.

The rating of brightness had rather low correlation with any audio feature. One would have expected better correlation with the spectral features. The largest correlation is with the function for modality, using the method choos-

ing the best major and minor key. The correlation is positive, meaning major songs sound brighter. This can be due to the uncontrolled stimuli; songs in the stimuli with major key might be brighter. Another possibility is that people perceive major keys as brighter than minor, even with the same timbre. In addition the rated brightness correlated strongly with rated pitch (r=0.9). All this indicates that the brightness rating did not work the way we intended. Rather than rating the spectral quality of the sound the subjects seem to have rated a more complex quality possibly related to pitch and mode.

4.2 Regression analysis

To find out how well the perceptual features could be predicted we performed separate multiple regression analyses with each perceptual feature as the dependent variable and all the audio features as independent variables. Since the number of independent variables (22) were too high in relation to number of cases (100) we applied a step-wise multiple regression. However, this procedure is questionable and the results should be considered as preliminary and without consideration of details. The multiple regression coefficient R^2 determines how well the regression model fits the actual data. A summary of the result is shown in in Table 5. Also shown is the number of variables that were selected by the step-wise procedure in each analysis.

Dependent variable	Adjusted R^2	Number of variables
Speed	0.76	8
Rhythmic complexity	0.14	2
Rhythmic clarity	0.52	1
Articulation	0.62	5
Dynamics	0.67	6
Modality	0.54	5
Harmonic complexity	0.23	5
Pitch	0.16	1
Brightness	0.29	2
Energy	0.68	5
Valence	0.50	2

Table 5. Summary of the step-wise regression analysis. Features in grey were predicted *a priori*.

All the regressions were significant but as seen in the table, the amount of explained variance (R^2) was rather modest. The regression results were in general similar to the correlations in Table 3. For example, speed could be rather well predicted as expected and the analysis included eight variables.

5. CONCLUSIONS AND DISCUSSION

The initial results of the perceptual ratings indicate that there was a rather good agreement among the listeners and that they could reliably assess the different musical aspects. The only scale that seemed to be problematic was the rating of brightness, also indicated by the high correlation between brightness and pitch. The emotion ratings could be well estimated by the perceptual features using multiple regression with adj. R^2 = 0.93 and 0.87, respectively.

The computed audio features correlated often with the perceptual ratings that were *a priori* expected. However, the audio features could only to a rather limited extent predict the perceptual ratings. Using multiple regression the best prediction was of speed with an adjusted R^2 = 0.76.

The selection of music examples is likely to have a strong effect on the results. It sets the variation of each feature and thus indirectly influences the judgment. It also influences the accuracy of the computed features. In addition, the current examples, which were converted from MIDI, had a rather limited timbral variation since they were all produced using the same synthesizer. Thus a future goal is to replicate this experiment using a different music set

The present selection of audio features only included a small subset of all previously suggested algorithms. Certainly, a broader selection of audio features would yield better results. Nevertheless, we think that these results point to the need for further development of audio features that are more specifically designed for these perceptual features. The only exception here was pulse clarity. It is likely that a small selection of such audio features would efficiently predict also higher-level semantic descriptions as indicated in Figure 1.

ACKNOWLEDGEMENTS

We would like to thank Erwin Schoonderwaldt who prepared the stimuli and ran the pilot experiment. This work was supported by the Swedish Research Council, Grant Nr. 2009-4285.

6. REFERENCES

[1] J. J. Burred, and A. Lerch, “Hierarchical Automatic Audio Signal Classification,” in Journal of the Audio Engineering Society, 52(7/8), 2004, pp. 724-738.

[2] K. Hevner, “The affective value of pitch and tempo in music,” in American Journal of Psychology, 49, 1937, pp. 621-30.

[3] A. Friberg, “Digital audio emotions — An overview of computer analysis and synthesis of emotions in music,” In Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland 2008, pp. 1-6.

[4] A. Friberg, E. Schoonderwaldt, & P. N. Juslin, “CUEX: An algorithm for extracting expressive tone variables from audio recordings,” in Acoustica united with Acta Acoustica, 93(3), 2005, pp. 411-420.

[5] A. Friberg, E. Schoonderwaldt, and A. Hedblad, “Perceptual ratings of musical parameters,” In H. von Loesch and S. Weinzierl (eds.) Gemessene Interpretation - Computergestützte Aufführungs-

analyse im Kreuzverhör der Disziplinen, Mainz: Schott 2011 (Klang und Begriff 4). (forthcoming)

[6] L. Wedin, “A Multidimensional Study of Perceptual-Emotional Qualities in Music,” in Scand. J. Psychol., 1972, 13, pp. 241-257.

[7] O. Lartillot, T. Eerola, P. Toiviainen, and F. Fornari, “Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization,” In Proceedings of the International Conference on Music Information Retrieval (ISMIR 2008), 2008, pp. 521-526.

[8] J. A. Russell, “A circumplex model of affect,” in Journal of Personality and Social Psychology, 1980, 39, pp. 1161 - 1178.

[9] V. Alluri, and P. Toiviainen, P. “In Search of Perceptual and Acoustical Correlates of Polyphonic Timbre,” Proc. of the Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM), Jyväskylä, Finland, 2009.

[10] O. Lartillot, and P. Toiviainen, “A MATLAB toolbox for musical feature extraction from audio,” in Proc. Of the 10th Int. Conference on Digital Audio Effects, 2007, (DAFx-07).

[11] A. Hedblad, Evaluation of Musical Feature Extraction Tools Using Perceptual Ratings. Master thesis, KTH, 2011, (forthcoming).

[12] F. Williams, Reasoning With Statistics. Holt, Rinehart and Winston, New York, 1968.

[13] R. Bresin, and A. Friberg, “Emotion rendering in music: range and characteristic values of seven musical variables,” Cortex, 2011, in press.

[14] G. Madison, and J. Paulin, “Relation between tempo and perceived speed,” in J. Acoust. Soc. Am., 128(5), 2010.