

Using audio features to model the affective response to music

Marc Leman, Valery Vermeulen, Liesbeth De Voogdt, Dirk Moelants

IPEM – Dept of Musicology, Ghent University, Belgium

Marc.Leman@UGent.be

Abstract

A series of 60 natural musical excerpts of various styles and genres was analyzed at three different levels: (a) subjective judgments, (b) manual-based musical analysis, (c) acoustical-based feature analysis. The responses of the subjective judgments can be modeled as a space with three axes: valence (gay-sad), activity (tender-calm) and interest (exciting-boring). Subjects seem to agree most on the activity dimension, which can be modeled using the prominence, loudness and brightness features of the musical audio. The valence axis can be modeled using the clarity and tempo/staccato cues. For the interest dimensions it is not possible to find a model that accounts for all the subjects.

The results show that affect attribution to natural musical excerpts can be captured using a valence-activity axis, while a third axis, the interest accounts for more personal aspects. The two main dimensions can be related to meaningful acoustical cues.

The relevance of this study is discussed in view of audio-mining, interactive multimedia systems, and brain research.

1. Introduction

The affective qualities that listeners tend to attribute to music cover a broad range of emotive, expressive and motoric adjectives. Until now, the characterization of the musical properties that determine the perception of these qualities in music has been based on (i) philosophical and musicological analysis [e.g. 1-4], (ii) perceptual analysis, where listeners judge the salience of a particular property [e.g. 5-8], and (iii) analysis-by-synthesis experiments, where the properties of a stimulus are systematically varied in order to test differences in perceived musical affect [e.g. 9-13]. Computational models have been developed that extract pitch and tonality [14-17], beat and rhythm [18-19] and timbre [20] from musical audio, but few attempts have been undertaken to relate these and similar audio-extracted properties to the perception of affective qualities in music [21].

This paper addresses the problem in two studies looking for: (i) the nature of the affect attribution space, (ii) the relationships between this space and the acoustically extracted cues.

2. The Affect Attribution Space

The structure of affect perception in a large set of musical stimuli was addressed using the method of

semantic differentials [22]. 60 musical audio excerpts were presented in random order and evaluated by 100 students. Each excerpt was evaluated with 15 bipolar adjectives using a 7-point scale. Factor Analysis was then used to gain insight in the nature of these correlations in terms of underlying factors (interpreted as affect attribution space). The first factor related to the valence (gay-sad), the second to activity (tender-calm), and the third to interest (exciting-boring).

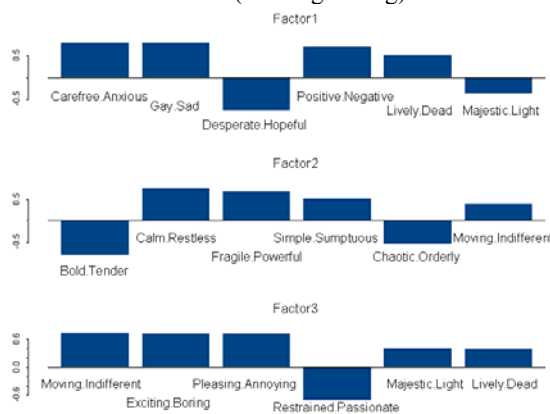


Figure 1. Factor Loadings of the Affect Attribution Space. The factors are: valence, activity, and interest. The adjectives are bi-polar and negative loadings can be turned into positive loadings by switching the adjective-pairs.

3. Manual and Automated Annotation of Syntactical Cues

3.1. Manual Cue Annotation by Experts

In a manual annotation experiment, 25 musicologists, all staff members, alumni or graduate students of Ghent University, rated the 60 fragments for 7 different features: (a) the tempo, measured by tapping along with the fragments, (b) the general loudness, measured on a 7-point soft-loud scale, (c) the articulation on a staccato-legato scale, (d) the brightness on a dull-sharp scale, (e) the absence or dominance of a melody (melodiousness), (f) the ambitus on a small-big scale, and (g) register on a low-high scale.

For features (b)-(g) the mean values as well as their standard deviations were stored, the latter giving an idea of the (dis)agreement between subjects. For (a), the

tempo chosen most often was stored together with the number of people tapping this tempo.

3.2. Cue Annotation Using Auditory Modeling

A set of cues has been automatically extracted from the audio excerpts using an auditory model. The extracted features are: (a) the loudness (soft-loud), (b) the roughness (flat-rough, related to the consonant-dissonant percept), (c) the spectral centroid (dull-sharp), (d) the number of onsets a second and (e) the inter-onset interval (both indicating temporal density), (f) the bandwidth, (g) the articulation (staccato-legato), (h) the number of channels in the auditory model that agree on the most prominent pitch (called pitch prominence 1), (i) the number of channels that agree on the same pitch (called: pitch prominence 2).

To obtain features that summarize an excerpt of music, average and standard deviation were computed from the time-domain features. The standard deviations represent the amount of change in the feature, during the excerpt. Single value descriptors are justified on the basis of rather homogeneous expressiveness in each of the 60 excerpts.

4. Correlation of Affect Attribution Space and Extracted Cues

This study proceeds in two steps. In the first step, individual responses of 8 listeners were projected onto the three-dimensional affect attribution space of valence-activity-interest. Using stepwise multiple regression analysis we then looked whether a linear combination of cues could approximate each individual valence-activity-interest projection.

Table 3 shows the result for the auditory-based extracted cues, after they were categorized into 7 groups:

- F1= Prominence (the two features describing pitch prominence and their standard deviations, each contributed positively to the prominence factor)
- F2: Articulation (with a positive contribution of the articulation feature and its standard deviation and a negative contribution of inter onset interval)
- F3: Loudness/roughness changes (determined by the standard deviations of both loudness and roughness)
- F4: Brightness (decreasing centroid and pitch agreement, increasing bandwidth)
- F5: Onsets (determined by the number of onsets and its standard deviation)

- F6: Brightness changes (determined by increasing standard deviations in spectral centroid and bandwidth)
- F7: Loudness/roughness (the mean values of the loudness and roughness parameters)

Table 1. Multiple regression analysis of categorized auditory-based cues with the affect attribution space (* for $p < .05$, ** for $p < .01$, *** for $p < .001$). First column is factor (1, 2 or 3), second column is subject identification, third column is R^2 , then follow the 7 auditory-based extracted cues

Factor	Id	R^2	F1	F2	F3	F4	F5	F6	F7
1	1	0.296	0.209	0	0	0	-0.507 ***	-0.27	0
1	2	0.147	0	0	-0.22	0	-0.35	0	0
1	3	0.227	0.17	0	0	0	-0.301 ***	0	0
1	4	0.133	0	0	0	0	-0.293 **	0	0
1	5	0.184	0.216	0	0	0	-0.354	0	-0.214
1	6	0.115	0	0	0	0	-0.245 *	0	0
1	7	0.226	0.156	0	0	0	-0.329 ***	0	0
1	8	0.229	0	0	0	-0.203	-0.37 ***	0	0
2	1	0.406	-0.376 ***	0	-0.223	0.439 ***	0	-0.198	0
2	2	0.412	-0.355 **	-0.277 *	0	0.407 ***	0.256	0	0
2	3	0.546	-0.393 ***	0	0	0.35 ***	0	0	0
2	4	0.404	-0.25 **	0	0	0.522 ***	0	-0.187	0
2	5	0.39	-0.28 **	0	0	0.427 ***	0	0	0.179
2	6	0.331	-0.353 ***	0	-0.184	0.313	0	0	0
2	7	0.437	-0.216	0	0	0.487 ***	0	-0.197	0.189
2	8	0.394	-0.369 ***	0	0	0.47 ***	0	0	0.196
3	1	0.115	0	0	0	0.159	0.216	0	0
3	2	0.162	0	0	-0.31	0.381	0	0	0
3	3	0.129	0	0	0	0.147	0.089	0	0
3	4	0.158	0	0	0	0.388 **	0	0	0
3	5	0.135	0	0	0	0	0.334 **	0	0
3	6	0.064	0	0	0	-0.174	0	0	0
3	7	0.134	0	0	0	0.179	0	0	0.27
3	8	0.277	-0.198	-0.253	-0.226	0	0	0.302 *	0

It turns out that a number of automatically extracted cues can be used to account for affect attribution. In particular valence-related adjectives, such as carefree, gay, hopeful, are related to cues that account for onsets. It was found that temporal density and a higher degree of musical consonance trigger the focus of attention and enhance the perception of positive qualities. An even clearer picture is found for activity-related adjectives such as bold, restless, and powerful. They are related to the low-level cues that account for centroid/width, and pitch prominence features. The high-level cue that is most relevant in this case, however, seems to be loudness. The higher the loudness, the more bold, restless, and powerful the music is perceived. For the Interest dimension, no significant trends in correlation with any of the 7 factors could be observed.

An attempt was undertaken to interpret the meaning of the acoustical categories. This was done using a stepwise multiple regression analysis where manually annotated cues are explained in function of audio cues. It turned out that only a few of the manually annotated cues can be effectively accounted for in terms of auditory-based extracted cues. In particular manual annotated loudness and manually annotated articulation score rather well in terms of a combination of audio cues. This is shown in Table 2, manual annotated

loudness, for example, is accounted for in terms of the auditory-based cues called brightness, loudness/roughness, and prominence. Notice that to a lesser extent, an account could be given of manually annotated tempo, and ambitus.

Table 2. Multiple regression analysis of acoustical categorized cues that account for manually extracted cues. *Tm* =mean tempo, *Tp*=percentage tapping mean tempo, *Rm*=roughness mean, *Rs*=roughness standard deviation, *Lm*=loudness mean, *Ls*= loudness std, *Am*=articulation (staccato-legato) mean, *As*= articulation std, *Bm*=brightness mean, *Bs*=brightness std, *Mm*=melody mean, *Ms*=melody std, *AAm*=ambitus mean, *AAs*=ambitus std, *RRm*=register mean, *RRs*=register std.

RowNames	R ²	F1	F2	F3	F4	F5	F6	F7
Tm	0.34	-8.46	0	-8.80	12.04*	14.16**	0	0
Tp	0.26	0	-2.14	0	2.43	3.32**	0	0
Rm	0.13	0	0	0	0.45**	0	0	0
Rs	0.15	0	0	0	0.05	0.07	0.06	0
Lm	0.61	-0.36***	0	0	0.56***	0	0	0.37***
Ls	0.44	0.05	0	0.05	-0.09***	-0.05	0	-0.08**
Am	0.66	0.58***	0.48***	0	-0.34**	-0.63***	-0.30*	0
As	0	0	0	0	0	0	0	0
Bm	0.19	0	0	0.26**	0.17	0	0	0
Bs	0.23	-0.07	0	-0.08	0.07	0	0	0.09
Mm	0.17	0.32	0	0	-0.32	0.26	0	0
Ms	0.07	0	0	0	0.09	0	0	0
AAm	0.36	0.21	0	0	0	-0.26*	-0.30**	-0.20
AAs	0	0	0	0	0	0	0	0
RRm	0.26	0.25	0	0.26*	0	0	-0.24	0
RRs	0.22	0	0	-0.07*	0.07**	0	0	0

5. Discussion

Although the auditory-based extracted cues cannot give a complete account of high-level manually extracted cues, some important high-level cues such as loudness and articulation can be accounted for with a combination of low-level acoustical cues. This allows a meaningful modeling of musical affect perception in terms of the automatically extracted auditory-based cues.

Overall, however, it turns out that the manual annotated features of roughness, brightness, and melody recognition, are not very adequately modeled with the approach adopted here. There seems to be a large difference between how listeners perceive roughness and brightness in natural musical stimuli, and the findings of the modeling, based on how they perceive it in artificial stimuli. Context-dependent interactions between different musical properties may be the main reason why apparently simple acoustical cues cannot straightforwardly account for manually annotated cues. Our analysis nevertheless shows that certain global trends can be established.

The cues, and their interpretation, are in agreement with the literature on manually extracted determinants of musical emotions [23]. The study shows that the focus

on individual affect processing is feasible and may be straightforwardly elaborated in future studies. It is indeed not unrealistic to develop acoustical models of affect perception on a purely individual basis, using an extended methodology of the one developed in this paper. Inter-subjective and individual acoustical models of musical affect perception are useful in different domains of music information processing.

6. Conclusions

Inter-subjective and individual acoustical models of musical affect perception are useful in different domains of music information processing. This study is a step towards the development of an instrumental theory of musical content analysis which is needed in view of a number of applications, in particular audio-mining, interactive multimedia, and brain research.

In audio-mining [24] most of the present day music information retrieval systems are based on low and mid-level objective/syntactical descriptors of audio, such as timbre [25,26], chroma [27,28] and modulation spectra [29]. Future systems however could also support descriptions related to affect perception. The fact that there is still very little experience with the use of these subjective qualities in music information retrieval systems is mainly because it was not so clear yet which subjective qualities are at the one hand effective for classifying music and at the other hand extractable with some reliability from the audio. Clearly, music information retrieval research will have to develop new retrieval paradigms that can incorporate the proposed broad palette of affect descriptors. This study shows the feasibility of acoustical modeling of affect perception.

Multi-modal interactive music systems are a new generation of musical instruments based on real-time and intelligent human-machine interaction [30,31]. These systems are multi-modal in the sense that they deal with multiple modalities such as audio, vision, and different kinds of haptic sensing. They have an enormous potential for the production of art, and can also be used for multi-modal scientific research and human aid programs such as education or revalidation [32]. The analysis-by-synthesis approach to affect perception has already led to sophisticated models of musical expressiveness [33]. Systems dealing with affect perception are still working on lower levels of feature extraction [34]. If these systems have to interact in an intelligent and spontaneous way with users, their communication capabilities should rely on a set of advanced musical and gestural content processing tools. There are indeed many occasions where users may want to interact in a spontaneous and even expressive way

with these systems, using descriptions of perceived qualities, or making expressive movements.

Acknowledgments

The authors wish to thank Johannes Taelman for his contribution to the acoustical analysis.

References

- [1] Pratt, C.C., *The Meaning of Music: A Study in Psychological Aesthetics* (Johnson Reprint Corp., New York), 1968.
- [2] Imberty, M., *Entendre la musique* (Hearing music) (Dunod, Paris), 1979.
- [3] Broeckx, J., *Muziek, ratio en affect* (Music, ratio and affect) (Metropolis, Antwerpen), 1981.
- [4] Kivy, P., *New essays on musical understanding* (Clarendon Press, Oxford), 2001.
- [5] Watson, K.B., "The nature and measurement of musical meanings," *Psychol. Monogr.*, 54:1-43, 1942.
- [6] Wedin, L., "A multidimensional study of perceptual-emotional qualities in music," *Swed. J. Musicology*, 13:241-257, 1972.
- [7] Nielzén, S. and Cesarec, Z., "Emotional experience of music as a function of musical structure," *Psychol. Music*, 10:7-17, 1982.
- [8] Balkwill, L.-L. and Thompson, W.F., "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," *Music Percept.*, 17:43-64, 1999.
- [9] Hevner, K., "Experimental studies of the elements of expression in music," *Am. J. Psychol.*, 48:246-268, 1936.
- [10] Rigg, M.G., "What features of a musical phrase have emotional suggestiveness?," *Pub. Soc. Sci. Res. Coun. Oklahoma A. and M. Coll.*, 1:1-38, 1939.
- [11] Thompson, W.F., and Robitaille, B., "Can composers express emotions through music?," *Empirical Stud. Arts*, 10:79-89, 1992.
- [12] Juslin, P.N., "Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgment policy," *Musicae Scientiae*, 1:225-256, 1997.
- [13] Gagnon, L. and Peretz, I., "Mode and tempo relative contributions to "happy-sad" judgments in equitone melodies," *Cognition Emotion*, 17:25-40, 2003.
- [14] Terhardt, E., "Pitch, consonance, and harmony," *J. Acoust. Soc. Am.*, 55:1061-1069, 1974.
- [15] Parncutt, R., *Harmony: A Psycho-acoustical Approach* (Springer-Verlag, Berlin), 1989.
- [16] Leman, M., "A model of retroactive tone center perception," *Music Percept.*, 12:439-471, 1995.
- [17] Leman, M., "An Auditory Model of the Role of Short-Term Memory in Probe-Tone Ratings," *Music Percept.*, 17:481-509, 2000.
- [18] Toiviainen, P., "Real-time recognition of improvisations with adaptive oscillators and a recursive Bayesian classifier," *J. New Music Res.* 30:137-147, 2001.
- [19] Large, E. and Kolen, J., "Resonance and the perception of musical meter," *Connect. Sci.* 6:177-208, 1994.
- [20] De Poli, G. and Prandoni, P., "Sonological models for timbre characterization," *J. New Music Res.* 26:170-197, 1997.
- [21] Scheirer, E.D., Watson, R.B., and Vercoe, B.L., "On the perceived complexity of short musical segments," *Proc. Int. Conf. on Music Percept. and Cognition*, Keele (CD-ROM), 2000.
- [22] Leman, M., Vermeulen, V., De Voogdt, L., Taelman, J., Moelants, D. (2003), "Acoustical and Computational Modeling of Musical Affect Perception" (submitted).
- [23] Gabrielsson, A. and Juslin, P.N., "Emotional expression in music," in *Handbook of affective sciences*, edited by R.J. Davidson, H.H. Goldsmith and K.R. Scherer (Oxford University Press, New York), 503-534, 2003.
- [24] Leman, M., Clarisse, L., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G., Martens, J. P., and Van Steelant, D., "Tendencies, perspectives, and opportunities of musical audio-mining," in *Forum Acusticum Sevilla* (CD-ROM), 2002.
- [25] Aucouturier, J.J. and Pachet, F., "Music similarity measures: What's the use?," *Proc. Int. Symp. on Music Information Retrieval*, Paris, 157-163, 2002.
- [26] Schwartz, D. and Rodet, X., "Spectral estimation and representation for sound analysis-synthesis," *Proc. Int. Computer Music Conf.*, Beijing, 1999.
- [27] Bartsch, M. and Wakefield, G., "To catch a chorus: using chroma-based representations for audio thumbnailing," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.
- [28] Dannenberg, R. and Hu, N., "Pattern discovery techniques for music audio," *Proc. Int. Symp. on Music Information Retrieval*, Paris, 63-70, 2002.
- [29] Rauber, A. and Pampalk, E., "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity," *Proc. Int. Symp. on Music Information Retrieval*, Paris, 71-80, 2002.
- [30] Rowe, R., *Machine Musicianship* (MIT Press, Cambridge), 2001.
- [31] Camurri, A., "Machine musicianship," *Music Percept.* 20:323-327, 2003.
- [32] Camurri, A., "Music content processing and multimedia: case studies and emerging applications of intelligent interactive systems," *J. New Music Res.* 28:351-363, 1999.
- [33] Bresin, R. and Friberg, A., "Emotional coloring of computer controlled music performance", *Comp. Music J.* 24:44-63, 2000.
- [34] Camurri, A., De Poli, G., Leman, M., and Volpe, G., "A multi-layered conceptual framework for expressive gesture applications," *Proc. Workshop on Current Research Directions in Comp. Music*, Barcelona, 29-34, 2001.