The Role of Time in Music Emotion Recognition

Marcelo Caetano[1]* and Frans Wiering[2]

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas FORTH-ICS, Heraklion, Crete, Greece
[2] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
caetano@ics.forth.gr, f.wiering@uu.nl

Abstract. Music is widely perceived as expressive of emotion. Music plays such a fundamental role in society economically, culturally and in people's personal lives that the emotional impact of music on people is extremely relevant. Research on automatic recognition of emotion in music usually approaches the problem from a classification perspective, comparing "emotional labels" calculated from different representations of music with those of human annotators. Most music emotion recognition systems are just adapted genre classifiers, so the performance of music emotion recognition using this limited approach has held steady for the last few years because of several shortcomings. In this article, we discuss the importance of time, usually neglected in automatic recognition of emotion in music, and present ideas to exploit temporal information from the music and the listener's emotional ratings. We argue that only by incorporating time can we advance the present stagnant approach to music emotion recognition.

Keywords: Music, Time, Emotions, Mood, Automatic Mood Classification, Music Emotion Recognition

1 Introduction

The emotional impact of music on people and the association of music with particular emotions or 'moods' have been used in certain contexts to convey meaning, such as in movies, musicals, advertising, games, music recommendation systems, and even music therapy, music education, and music composition, among others. Empirical research on emotional expression started about one hundred years ago, mainly from a music psychology perspective [1], and has successively increased in scope up to today's computational models. Research on music and emotions usually investigates listeners' response to music by associating certain emotions to particular pieces, genres, styles, performances, etc. An emerging field is the automatic recognition of emotions (or 'mood') in music, also called music emotion recognition (MER) [7]. A typical approach to MER categorizes emotions into a number of classes and applies machine learning techniques

* This work is funded by the Marie Curie IAPP "AVID MODE" grant within the European Commissions FP7.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

to train a classifier and compare the results against human annotations [7, 22, 10]. The 'automatic mood classification' task in MIREX epitomizes the machine learning approach to MER, presenting systems whose performance range from 22 to 65 percent [3]. Researchers are currently investigating [4, 7] how to improve the performance of MER systems. Interestingly, the role of time in the automatic recognition of emotions in music is seldom discussed in MER research.

Musical experience is inherently tied to time. Studies [8, 11, 5, 18] suggest that the temporal evolution of the musical features is intrinsically linked to listeners' emotional response to music, that is, emotions expressed or aroused by music. Among the cognitive processes involved in listening to music, memory and expectations play a major role. In this article, we argue that time lies at the core of the complex link between music and emotions, and should be brought to the foreground of MER systems.

The next section presents a brief review of the classic machine learning approach to MER. Then, we discuss an important drawback of this approach, the lack of temporal information. We present the traditional representation of musical features and the model of emotions to motivate the incorporation of temporal information in the next section. Next we discuss the reationship between the temporal evolution of musical features and emotional changes. Finally, we present the conclusions and discuss future perspectives.

2 The Traditional Classification Approach

Traditionally, research into computational systems that automatically estimate the listener's emotional response to music approaches the problem from a classification standpoint, assigning "emotional labels" to pieces (or tracks) and then comparing the result against human annotations [7, 22, 10, 3]. In this case, the classifier is a system that performs a mapping from a feature space to a set of classes. When applied in MER, the features can be extracted from different representations of music, such as the audio, lyrics, the score, among others [7], and the classes are clusters of emotional labels such as "depressive" or "happy". There are several automatic classification algorithms that can be used, commonly said to belong to the machine learning paradigm of computational intelligence.

2.1 Where Does the Traditional Approach Fail?

Independently of the specific algorithm used, the investigator that chooses this approach must decide how to represent the two spaces, the musical features and the emotions. On the one hand, we should choose musical features that capture information about the expression of emotions. Some features such as tempo and loudness have been shown to bear a close relationship with the perception of emotions in music [19]. On the other hand, the model of emotion should reflect listeners' emotional response because emotions are very subjective and may change according to musical genre, cultural background, musical training and exposure, mood, physiological state, personal disposition and taste

[1]. We argue that the current approach misrepresents both music and listeners' emotional experience by neglecting the role of time.

2.2 Musical Features

Most machine learning methods described in the literature use the audio to extract the musical features [7, 22, 10, 3]. Musical features such as tempo, loudness, and timbre, among many others, are estimated from the audio by means of signal processing algorithms [12]. Typically, these features are calculated from successive frames taken from excerpts of the audio that last a few seconds [7, 22, 10, 3, 4] and then averaged, losing the temporal correlation [10]. Consequently, the whole piece (or track) is represented by a static (non time-varying) vector, intrinsically assuming that musical experience is static and that the listener's emotional response can be estimated from the audio alone. The term 'semantic gap' has been coined to refer to perceived musical information that does not seem to be contained in the acoustic patterns present in the audio, even though listeners agree about its existence [21].

However, to fully understand emotional expression in music, it is important to study the performer's and composer's intention on the one hand, and the listener's perception on the other [6]. Music happens essentially in the brain, so we need to take the cognitive mechanisms involved in processing musical information into account if we want to be able to model people's emotional response to music. Low-level audio features give rise to high-level musical features in the brain, and these, in turn, influence emotion recognition (and experience). This is where we argue that time has a major role, still neglected in most approaches found in the literature. Musical experience and the cognitive processes that regulate musical emotions are entangled with each other around the temporal dimension, so the model of emotion should account for that.

2.3 Representation of Emotions

MER research tends to use categorical descriptions of emotions where the investigator selects a set of "emotional labels" (usually mutually exclusive). The left-hand side of figure 1 illustrates these emotional labels (Hevner's adjective circle [2]) clustered in eight classes. The choice of the emotional labels is important and might even affect the results. For example, the terms associated with music usually depend on genre (pop music is much more likely than classical music to be described as "cool"). As Yang [22] points out, the categorical representation of emotions faces a granularity issue because the number of classes might be too small to span the rich range of emotions perceived by humans. Increasing the number of classes does not necessarily solve the problem because the language used to categorize emotions is ambiguous and subjective [1]. Therefore, some authors [7, 22] have proposed to adopt a parametric model from psychology research [14] known as the circumplex model of affect (CMA). The CMA consists of two independent dimensions whose axes represent continuous values

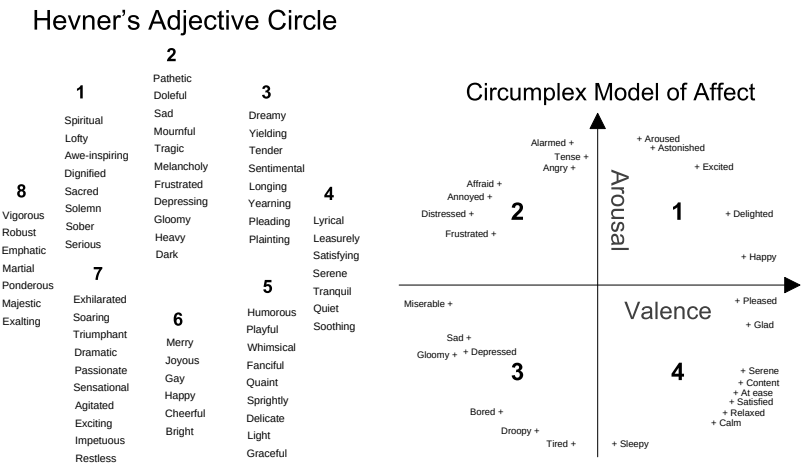

Fig. 1. Examples of models of emotion. The left-hand side shows Hevner's adjective circle [2], a categorical description. On the right, we see the circumplex model of affect [14], a parametric model.

of valence (positive or negative semantic meaning) and arousal (activity or excitation). The right-hand side of figure 1 shows the CMA and the position of some adjectives used to describe emotions associated with music in the plane. An interesting aspect of parametric representations such as the CMA lies in the continuous nature of the model and the possibility to pinpoint where specific emotions are located. Systems based on this approach train a model to compute the valence and arousal values and represent each music piece as a point in the two-dimensional emotion space [22].

One common criticism of the CMA is that the representation does not seem to be metric. That is, emotions that are very different in terms of semantic meaning (and psychological and cognitive mechanisms involved) can be close in the plane. In this article, we argue that the lack of temporal information is a much bigger problem because music happens over time and the way listeners associate emotions with music is intrinsically linked to the temporal evolution of the musical features. Also, emotions are dynamic and have distinctive temporal profiles (boredom is very different from astonishment in this respect, for example).

3 The Role of Time in the Complex Relationship Between Music and Emotions

Krumhansl [9] suggests that music is an important part of the link between emotions and cognition. More specifically, Krumhansl investigated how the dynamic aspect of musical emotion relates to the cognition of musical structure. According to Krumhansl, musical emotions change over time in intensity and quality,

and these emotional changes covary with changes in psycho-physiological measures [9]. Musical meaning and emotion depend on how the actual events in the music play against this background of expectations. David Huron [5] wrote that humans use a general principle in the cognitive system that regulates our expectations to make predictions. According to Huron, music (among other stimuli) influences this principle, modulating our emotions. Time is a very important aspect of musical cognitive processes. Music is intrinsically temporal and we need to take into account the role of human memory when experiencing music. In other words, musical experience is learned. As the music unfolds, the learned model is used to generate expectations, which are implicated in the experience of listening to music. Meyer [11] proposed that expectations play the central psychological role in musical emotions.

3.1 Temporal Evolution of Musical Features

The first important step to incorporate time into MER is to monitor the temporal evolution of musical features [18]. After the investigator chooses which features to use in a particular application, the feature vector should be calculated for every frame of the audio signal and kept as a time series (i.e., a time-varying vector of features). The temporal correlation of the features must be exploited and fed into the model of emotions to estimate listeners' response to the repetitions and the degree of "surprise" that certain elements might have [19].

Here we could make a distinction between perceptual features of musical sounds (such as pitch, timbre, and loudness) and musical parameters (such as tempo, key, and rhythm), related to the structure of the piece (and usually found in the score). Both of them contribute to listeners' perception of emotions. However, their temporal variations occur at different rates. Timbral variations, for example, and key modulations or tempo changes happen at different levels. Figure 2 illustrates these variations at the microstructural (musical sounds) and macrostructural (musical parameters) level.

3.2 Emotional Trajectories

A very simple way of recording information about the temporal variation of emotional perception of music would be to ask listeners to write down the emotional label and a time stamp as the music unfolds. The result is illustrated on the left-hand side of figure 3. However, this approach suffers from the granularity and ambiguity issues inherent of using a categorical description of emotions. Ideally, we would like to have an estimate of how much a certain emotion is present at a particular time.

Krumhansl [8] proposes to collect listener's responses continuously while the music is played, recognizing that retrospective judgments are not sensitive to unfolding processes. However, in this study [8], listeners assessed only one emotional dimension at a time. Each listener was instructed to adjust the position of a computer indicator to reflect how the amount of a specific emotion (for

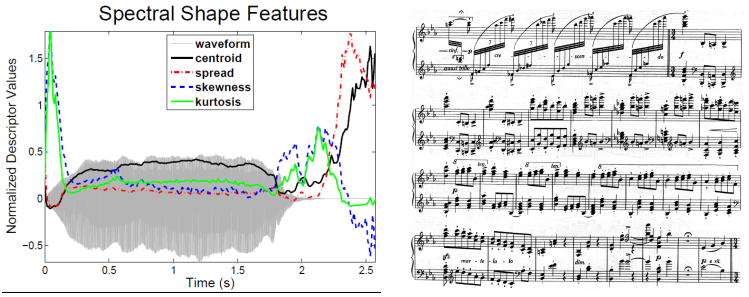

Fig. 2. Examples of the temporal variations of musical features. The left-hand side shows temporal variations of the four spectral shape features (centroid, spread, skewness, and kurtosis, perceptually correlated to timbre) during the course of a musical instrument sound (microstructural level). On the right, we see variations of musical parameters (macrostructural level) represented by the score for simplicity.

example, sadness) they perceived changed over time while listening to excerpts of pieces chosen to represent the emotions [8].

Here, we propose a similar procedure using a broader palette of emotions available to allow listeners to associate different emotions to the same piece. Recording listener's emotional ratings over time [13] would lead to an emotional trajectory like the one shown on the right of figure 3, which illustrates an emotional trajectory (time is represented by the arrow) in a conceptual emotional space, where the dimensions can be defined to suit the experimental setup. The investigator can choose to focus on specific emotions such as happiness and aggressiveness, for example. In this case, one dimension would range from happy to sad, while the other from aggressive to calm. However, we believe that Russell's CMA [14] would better fit the exploration of a broader range of emotions because the dimensions are not explicitly labeled as emotions.

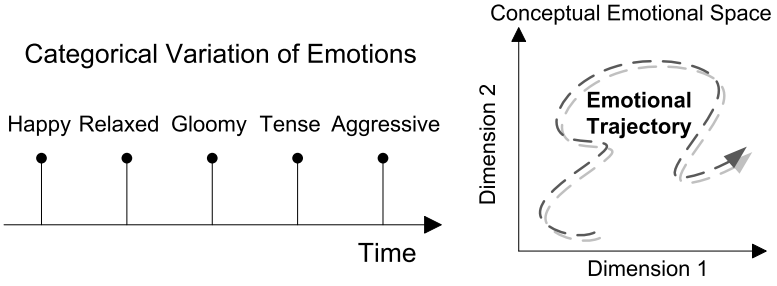

Fig. 3. Temporal variation of emotions. The left-hand side shows emotional labels recorded over time. On the right, we see a continuous conceptual emotional space with an emotional trajectory (time is represented by the arrow).

3.3 Investigating the Relationship Between the Temporal Evolution of Musical Features and the Emotional Trajectories

Finally, we should investigate the relationship between the temporal variation of musical features and the emotional trajectories. MER systems should include information about the rate of temporal change of musical features. For example, we should investigate how changes in loudness correlate with the expression of emotions. Schubert [18] studied the relationship between musical features and perceived emotion using continuous response methodology and time-series analysis. Musical features (loudness, tempo, melodic contour, texture, and spectral centroid) were differenced and used as predictors in linear regression models of valence and arousal. This study found that changes in loudness and tempo were associated positively with changes in arousal, and melodic contour varied positively with valence. When Schubert [19] discussed modeling emotion as a continuous, statistical function of musical parameters, he argued that the statistical modeling of memory is a significant step forward in understanding aesthetic responses to music. Only very recently MER systems started incorporating dynamic changes in efforts mainly by Schmidt and Kim [15–17, 20]. Therefore, this article aims at motivating the incorporation of time in MER to help break through the so-called "glass ceiling" (or "semantic gap") [21], improving the performance of computational models of musical emotion with advances in our understanding of the currently mysterious relationship between music and emotions.

4 Conclusions

Research on automatic recognition of emotion in music, still in its infancy, has focused on comparing "emotional labels" automatically calculated from different representations of music with those of human annotators. Usually the model represents the musical features as static vectors extracted from short excerpts and associates one emotion to each piece, neglecting the temporal nature of music. Studies in music psychology suggest that time is essential in emotional expression. In this article, we argue that MER systems must take musical context (what happened before) and listener expectations into account. We advocate the incorporation of time in both the representation of musical features and the model of emotions. We prompted MER researchers to represent the music as a time-varying vector of features and to investigate how the emotions evolve in time as the music develops, representing the listener's emotional response as an emotional trajectory. Finally, we discussed the relationship between the temporal evolution of the musical features and the emotional trajectories.

Future perspectives include the development of computational models that exploit the repetition of musical patterns and novel elements to predict listeners' expectations and compare them against the recorded emotional trajectories. Only by including temporal information in automatic recognition of emotions can we advance MER systems to cope with the complexity of human emotions in one of its canonical means of expression, music.

References

1. Gabrielsson, A., Lindstrom, E.: The Role of Structure in the Musical Expression of Emotions. In: Handbook of Music and Emotion: Theory, Research, Applications. Eds. Patrik N. Juslin and John Sloboda, pp. 367–400 (2011)
2. Hevner, K.: Experimental Studies of the Elements of Expression in Music. The Am. Journ. Psychology . 48 (2), pp. 246–268 (1936)
3. Hu, X., Downie, J.S., Laurier, C., Bay, M., and Ehmann, A.F.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: Proc. ISMIR (2008)
4. Huq, A., Bello, J.P., and Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journ. New Music Research. 39(4), pp. 227–244 (2010)
5. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. MIT Press, (2006)
6. Juslin, P., Timmers, R.: Expression and Communication of Emotion in Music Performance. In: Handbook of Music and Emotion: Theory, Research, Applications Eds. Patrik N. Juslin and John Sloboda, pp. 453–489 (2011)
7. Kim, Y.E., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: Music Emotion Recognition: A State of the Art Review. In: Proc. ISMIR (2010)
8. Krumhansl, C. L.: An Exploratory Study of Musical Emotions and Psychophysiology. Canadian Journ. Experimental Psychology. 51, pp. 336–352 (1997)
9. Krumhansl, C. L.: Music: A Link Between Cognition and Emotion. Current Directions in Psychological Science. 11, pp. 45–50 (2002)
10. MacDorman, K. F., Ough S., Ho C.C.: Automatic Emotion Prediction of Song Excerpts: Index Construction, Algorithm Design, and Empirical Comparison. Journ. New Music Research. 36, pp. 283–301 (2007)
11. Meyer, L.: Music, the Arts, and Ideas. University of Chicago Press, Chicago (1967)
12. Müller, M., Ellis, D.P.W., Klapuri, A., Richard, G.: Signal Processing for Music Analysis. IEEE Journal of Selected Topics in Sig. Proc. 5(6), pp. 1088–1110 (2011)
13. Nagel, F., Kopiez, R., Grewe, O., Altenmüller, E.: EMuJoy. Software for the Continuous Measurement of Emotions in Music. Behavior Research Methods, 39 (2), pp. 283–290 (2007)
14. Russell, J.A.: A Circumplex Model of Affect. Journ. Personality and Social Psychology. 39, pp. 1161–1178 (1980)
15. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In: Proc. ICMLA (2010)
16. Schmidt, E.M., Kim, Y.E.: Prediction of Time-Varying Musical Mood Distributions from Audio. In: Proc. ISMIR (2010)
17. Schmidt, E.M., Kim, Y.E.: Modeling Musical Emotion Dynamics with Conditional Random Fields. In: Proc. ISMIR (2011)
18. Schubert, E.: Modeling Perceived Emotion with Continuous Musical Features. Music Perception, 21(4), pp. 561–585 (2004)
19. Schubert, E.: Analysis of Emotional Dimensions in Music Using Time Series Techniques. Journ. Music Research, 31, pp. 65–80 (2006)
20. Vaizman, Y., Granot, R.Y., Lanckriet, G.: Modeling Dynamic Patterns for Emotional Content in Music. In: Proc. ISMIR (2011)
21. Wiggins, G. A.: Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. IEEE International Symposium on Multimedia, pp. 477-482 (2009)
22. Yang, Y., Chen, H.: Ranking-Based Emotion Recognition for Music Organization and Retrieval. IEEE Trans. Audio, Speech, Lang. Proc. 19, 4 (2011)