Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland, 11th - 15th June 2013. Geoff Luck & Olivier Brabant (Eds.)

ABSOLUTE OR RELATIVE? A NEW APPROACH TO BUILDING FEATURE VECTORS FOR EMOTION TRACKING IN MUSIC

Vaiva Imbrasaitė, Peter Robinson

Computer Laboratory, University of Cambridge, UK
Vaiva.Imbrasaite@cl.cam.ac.uk

Abstract

It is believed that violation of or conformity to expectancy when listening to music is one of the main sources of musical emotion. To address this, we test a new way of building feature vectors and representing features within the vector for the machine learning approach to continuous emotion tracking systems. Instead of looking at the absolute values for specific features, we concentrate on the average value of that feature across the whole song and the difference between that and the absolute value for a particular sample. To test this "relative" representation, we used a corpus of popular music with continuous labels on the arousal-valence space. The model consists of a Support Vector Regression classifier for each axis, with one feature vector for each second of a song. The relative representation, when compared to the standard way of building feature vectors, gives a 10% improvement on average (and up to 25% improvement for some models) on the explained variance for both the valence and arousal axes. We also show that this result is not due to having the average of a feature in the feature vector, but due to the actual relative representation.

Keywords: continuous emotion tracking, dimensional space, machine learning

1. Introduction

Over the last twenty years or so, the interest in music as a research focus has been growing, and it is attracting attention from a wide range of disciplines: philosophy, psychology, sociology, musicology, neurobiology, anthropology, and computer science. From the computer science perspective there has been an increasing amount of research concerned with automatic information extraction from music that would allow us to manage our growing digital music libraries more efficiently.

In addition to that, the birth of the Affective Computing field (Picard, 1997) together with a sparking interest in emotion research in general led people to look into the relationship between music and emotion (Juslin & Sloboda, 2001, 2010).

After the early debate about whether or not music could express or induce emotions at all, both are now generally accepted with multidisciplinary backing. Not only that, but it has been shown that emotion in music is shared between different cultures (Peretz, 2010), and therefore is universal and related to the basic emotions in people. It also has as strong an effect on the brain, as everyday emotions, activating the same or similar areas in the brain (Koelsch, Siebel, & Fritz, 2010).

Since the first paper on automatic emotion detection in music (Li & Ogihara, 2003) was published nearly 10 years ago, the field has been growing quite rapidly, but there is still a

lot to be explored and a lot of guidelines for future work to be set.

In this paper we aggregate and show several different sources of information (temporal, axis-dependency) that are present in music, but not in the basic approach to building feature vectors for machine learning approach to emotion tracking. We also test a novel feature representation technique that provides a substantial improvement to the results.

2. Background

There are several things that complicate music emotion research. One of the least recognized ones is that there are two types of musical emotions one can investigate – emotion "expressed" by the music, and emotion induced in the listener. The former is concerned with what the music sounds like and is mainly influenced by the musical features and cultural understanding of music. It is also more objective, since the listener's state and preferences have less of an effect on the perception of emotion in music. The later, on the other hand, describes the user's response to a piece of music. It clearly depends on the perceived (expressed by music) emotion, but is also heavily influenced by the individual's experiences, history, personality, preferences and social context. It is therefore much more subjective and varies more between different people.

Even though the vast majority of papers in Music Emotion Recognition (MER) do not make the distinction, there is clear evidence that the two are different. In their study, Zentner et al. (Zentner, Grandjean, & Scherer, 2008) have found a statistically significant difference between the (reported) felt and perceived emotions in people's reported emotional response to music. They have also found that certain emotions are more frequently perceived than felt in response to music (particularly the negative ones), and some are more frequently felt rather than perceived (e.g. amazement, activation, etc.).

Another issue that needs to be addressed is the granularity of the labels attached to a song. Even though there is no doubt that emotion in music can and does change over time (Schmidt & Kim, 2010a), the majority of research in MIR is aimed at classifying the whole musical piece, rather than tracking the emotion. In order to get around the dynamic nature of emotion and music, many researchers choose to look at a (usually) 30s segment of a piece, therefore making their systems less applicable in the real world. It has also been shown that emotion tracking can lead to an improvement in accuracy if classification of the whole musical piece is required (Carvalho & Chao, 2005).

The last key choice is the representation of emotion. A growing number of researchers choose to use dimensional emotion models. These models disregard the notion of basic (or complex) emotions. Instead, they describe emotions in terms of affective dimensions. The theory does not limit the number of dimensions that is used – it normally ranges between one (e.g. arousal) and three (valence, activation and power or dominance), but four and higher dimensional systems have also been proposed. The most commonly used model is Thayer's arousal-valence (AV) emotion space, where arousal describes how active/passive emotion is and valence - how positive/negative it is. In addition to being more flexible and less interpretation dependent than basic emotion (happy, sad, etc.) model, it has also been shown that classification which predicts AV values internally has higher accuracy than models that predict basic emotions directly.

The dimensional representation offers, in our opinion, the best solution – time varying MER, or continuous emotion tracking. Even though it is clearly not restricted to the dimensional approach (as has been shown by (Liu, 2006) and (Schubert, Ferguson, Farrar, Taylor, & Mcpherson, 2012)), it is inherently more difficult to use, especially in user studies.

Even within dimensional emotion tracking, there are different ways of approaching the problem. (Korhonen, Clausi, & Jernigan, 2006), (Panda & Paiva, 2011), (Schmidt & Kim, 2010a), (Schmidt, Turnbull, & Kim, 2010), and others have tried to infer the emotion label over a time window individually. Another solution is to incorporate temporal information in the feature vector either by using features extracted over varying window length for each second/sample (Schubert, 2004), or by using machine learning techniques that are adapted for

sequential learning (e.g. sequential stacking algorithm (Carvalho & Chao, 2005), Kalman filtering (Schmidt & Kim, 2010b) or conditional random fields (Schmidt & Kim, 2011). Interestingly, it has also been reported (Panda & Paiva, 2011; Schmidt et al., 2010) that taking the average of the time-varying emotion produces results that are statistically significantly better than simply performing emotion recognition on the whole piece of music.

3. Methodology

Dataset: The dataset that we have been using for our experiments is, to our knowledge, the only publicly available emotion tracking dataset of music extracts labelled on an arousal-valence dimensional space. It also focuses on perceived emotion rather than the perceived one. The data has been collected by (Speck, Schmidt, Morton, & Kim, 2011) using Mechanical Turk (MTurk, http://mturk.com), asking paid participants to label 15-second long excerpts with continuous emotion ratings on the AV space, with another 15 seconds given as a practice for each song. The songs in the dataset cover a wide range of genres – pop, various types of rock, hip-hop/rap, etc., and are drawn from the "uspop2002" (http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html) database containing Western popular songs. The dataset consists of 240 15-second clips (without the practice run) with 16.9 +/- 2.7 ratings for each clip. In addition, the dataset contains a standard set of features extracted from those musical clips: MFCCs, octave-based spectral contrast, statistical spectrum descriptors, chromagram and a set of EchoNest (http://developer.echonest.com/downloads) features.

The design of the experiments: Using the audio analysis features provided in the MTurk dataset and LIBSVM (Chang & Lin, 2001) implementation of support vector regression, we implemented a number of models for emotion tracking in music. The most basic model, based on the features provided and the LIBSVM library is also the common baseline model used in the field. We chose not to use the EchoNest features, since they have been extracted with proprietary software that does not provide clear documentation or explanation of how the features are extracted.

For the baseline method, the feature vector consists of the audio features averaged over a 1s window – the mean and standard deviation for each feature. There is only one feature vector for each second of the song (so 15 training/testing samples for each song), labelled with the average valence or arousal value computed from the labels in the dataset for that second. Two support vector regressors are trained – one for the arousal and one for the valence axes. Both regressors use RBF kernels and use 5-fold cross-validation within the training set to choose the best values for the parameters used.

Cross-validation: In all of our experiments we used 5-fold cross-validation to split the dataset into training and testing sets. This minimizes the risk of accidentally choosing a particularly bad or good set of songs and therefore making the results more reliable.

Table 1: Squared correlation of the baseline approach using different ways of splitting songs across folds.

	No constraints	Song-level split	Album-level split	Artist-level split
Arousal	0.69	0.64	0.65	0.64
Valence	0.34	0.25	0.26	0.23

We experimented with three different ways of distributing the songs between the folds (the effect on the squared correlation of the baseline method is depicted in **Table 1** and **Table 2**). The most obvious requirement is to keep all the feature vectors from a song in the same fold, to ensure that the model is not overfitting to individual songs. For the baseline method, this lowers the squared correlation coefficient (R^2) from 0.34 to 0.25 for valence and 0.69 to 0.64 for arousal, and increases the mean squared error (MSE) from 0.038 to 0.045 for valence and from 0.032 to 0.039 for arousal.

Table 2: Mean squared error of the baseline approach using different ways of splitting songs across folds.

	No constraints	Song-level split	Album-level split	Artist-level split
Arousal	0.033	0.039	0.038	0.038
Valence	0.038	0.045	0.045	0.046

Another factor worth considering is making sure that songs from the same album are all within a single fold. It has been reported and widely accepted that the so called "album effect" can artificially improve the performance as machine learning models overfit to a particular set of post-production techniques used on an album (Kim, Williamson, & Pilli, 2006). Removing the album effect made little difference to the results of the baseline method with the dataset we use. This is probably due to the fact that a large majority of songs come from unique albums – the 240 songs we am using come from 200 different albums.

The third approach we used was to make sure that all the songs from the same artist are within the same fold. Unsurprisingly, there is often statistically significant correlation between artists and mood in music (Hu & Downie, 2007), which, we expected, might lead to some overfitting. Again, it did not have a significant effect on the results, with the baseline method, which is most likely because the dataset is fairly well balanced even for the artists – the 240 songs used were recorded by 156 different artists. It could also be argued that this restriction is unnecessarily strict – in real life, a fully trained system is unlikely to receive unseen songs from an album that it was trained on, but is definitely expected to analyse unseen songs from an artist that it has seen before. For these reasons, we decided to use album-level cross-validation for all of the experiments.

Further experiments: The next step we took was to exploit some of the dependency between the valence and arousal axis (Eerola & Vuoskoski, 2010). It has been reported that including the valence label in the feature vector for arousal prediction and the arousal label for valence prediction can improve the accuracy of the model both in emotion recognition in music (Schmidt et al., 2010) and affect prediction from human behaviour (Nicolaou, Gunes, & Pantic, 2011a).

Another dependency that we decided to exploit was time. Since the emotional ratings for each second are clearly dependent on the previous ratings, in the next experiment we included audio features from a several one-second feature vectors. We experimented with varying sizes of windows – from 1s lag (just the audio features for the current second and all the audio features for the previous second) to 5s lag (current second and five previous seconds) for both the valence and the arousal axes.

Expectancy is also a very important factor to consider. There is a theory that violation of or conformity to expectancy when listening to music is a (main) source of musical emotion. It has been at least partially proven across different fields concerned with emotion in music (e.g. neuroimaging – (Koelsch et al., 2010), experimental aesthetics – (Hargreaves & North, 2010), etc.). To address that, we tried three different approaches: adding a "future" window in addition to the delay (similar to that used by (Nicolaou, Gunes, & Pantic, 2011b)), including the average over a song for each audio feature, and representing each feature as a difference between its (absolute) value at that second and the average over that song (which we will refer to as the relative representation).

4. Results

The results achieved by our basic implementation fall within the area of the results achieved within the field (R^2 of 0.65 for arousal and 0.26 for valence, and MSE of 0.038 for arousal and 0.045 for valence). Using the relative representation in the standard approach, on the other hand, showed a substantial improvement on the results (R^2 of 0.74 for arousal and 0.34 for valence, and MSE of 0.028 for arousal and 0.040 for valence).

Adding the label of the other emotion axis to the feature vector, as expected, had a positive effect on the valence prediction, but no effect on arousal prediction – results that agree with the findings in the literature (Schmidt et al., 2010). The same effect was

seen both in the standard, basic representation and in the relative representation.

Adding temporal information in the form of concatenating several seconds' worth of previous vectors (delay window) improved the performance of the basic representation models for both the valence and the arousal axes. For valence, the R^2 peaks at 2-3s window size and then plateaus or drops slightly. For arousal the optimal window size appears to be 4s (**Table 3** and **4**). For the relative representation, on the other hand, the effect is smaller or non-existent.

Table 3: R^2 of the basic (basic) and relative representations (rel) using delay windows of different size for arousal (A) and valence (V) axes.

	1s	2s	3s	4s	5s
A-basic	0.68	0.69	0.69	0.70	0.71
A-rel	0.74	0.76	0.73	0.74	0.74
V-basic	0.26	0.29	0.31	0.29	0.29
V-rel	0.31	0.31	0.31	0.32	0.31

Table 4: MSE of the basic (basic) and relative representations (rel) using delay windows of different size for arousal (A) and valence (V) axes.

	1s	2s	3s	4s	5s
A-basic	0.035	0.033	0.033	0.032	0.031
A-rel	0.028	0.026	0.029	0.028	0.028
V-basic	0.042	0.045	0.042	0.042	0.043
V-rel	0.042	0.042	0.041	0.041	0.042

Concatenating the current frame with feature vectors of the "upcoming" frames (future window) was also tested. We kept the range of future window sizes the same as for the delay window and it led to an improvement (between 0.01 and 0.02 for the R^2 value) when used on a standard feature representation for arousal at each window size. For the relative representation, adding the future window to the arousal model had no effect at all, and for valence model the results were inconsistent both in the standard and the relative representations. The addition of average was only tested on the standard representation, as the relative representation already contains average values by definition. For the basic approach, it produced a similar effect to that of the addition of the future window – inconsistent results on the valence model and small improvement on the arousal model (though smaller than the addition of future window).

5. Discussion

The results we have achieved with our models are very encouraging. The performance of the baseline method falls within the expected range reported in the literature, which suggests that the same techniques we used could be employed on other datasets. We have also managed to achieve the expected improvements by incorporating valence-arousal and temporal dependence information, in a similar way that has been achieved in the field. This confirms that there is a dependency both between different frames (temporal information) and between the two axes, and that it is beneficial to extract that information.

In order to address the expectancy, we tried several different approaches. Using a future window and adding an average over the whole song showed little, if any, improvement on the results. The major improvement on the accuracy of our predictions was introduced by the use of relative representation in the feature vectors. Interestingly, it seems that this representation makes a lot of other additions redundant – the results are not improved by adding the future window or the label of the other axis. This might be because the size of the feature vector grows too large, or because the information is somehow covered by this new representation.

Another important observation can be made from the results of these experiments – different modifications can have different levels of improvement to the valence and arousal models. This seem to imply that in order to achieve the best results, different feature representations and/or feature fusion techniques might need to be used for the two models, in

addition to potentially using or prioritizing different feature sets.

References

Carvalho, V. R., & Chao, C. (2005). Sentiment Retrieval in Popular Music Based on Sequential Learning. *Proc. ACM SIGIR*.

Chang, C., & Lin, C. (2001). LIBSVM: a library for support vector machines. *Computer*, *2*(3), 1–39.

Eerola, T., & Vuoskoski, J. K. (2010). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*, 18–49.

Hargreaves, D. J., & North, A. C. (2010). Experimental aesthetics and liking for music. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotions theory research applications* (pp. 515–547). OUP.

Hu, X., & Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. *Information Retrieval*, 67–72.

Juslin, P. N., & Sloboda, J. A. (2001). *Music and emotion: Theory and research*. (P. N. Jusling & J. A. Sloboda, Eds.) *Book* (Vol. 20, p. viii, 487 p.). OUP.

Juslin, P. N., & Sloboda, J. A. (2010). *Music and Emotion: Theory, Research, Applications*. (P. N. Juslin & J. A. Sloboda, Eds.) (p. 975). OUP.

Kim, Y. E., Williamson, D. S., & Pilli, S. (2006). Towards quantifying the album effect in artist identification. *Proceedings of ISMIR* (pp. 393–394).

Koelsch, S., Siebel, W. A., & Fritz, T. (2010). Chapter 12, Functional neuroimaging. In P. N. Juslin & J. A. Sloboda, *Handbook of music and emotion theory research application* (pp. 313–346). OUP.

Korhonen, M. D., Clausi, D. A., & Jernigan, M. E. (2006). Modeling emotional content of music using system identification. *IEEE transactions on systems man and cybernetics Part B Cybernetics a publication of the IEEE Systems Man and Cybernetics Society*, *36*(3), 588–599.

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In H. H. Hoos & D. Bainbridge (Eds.), *Proceedings ISMIR* (pp. 239–240).

Liu, D. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, *14*, 5–18.

Nicolaou, M. A., Gunes, H., & Pantic, M. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space, IEEE Transactions on Affective Computing 92–105 (2011). IEEE.

Nicolaou, M. A., Gunes, H., & Pantic, M. Output-associative RVM regression for dimensional and continuous emotion prediction. , Face and Gesture 2011 16–23 (2011).

Panda, R., & Paiva, R. P. (2011). Using Support Vector Machines for Automatic Mood Tracking in Audio Music. *130th Audio Engineering Society Convention*.

Peretz, I. (2010). Towards a neurobiology of musical emotions. In P. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion Theory research applications* (pp. 99–126). OUP.

Picard, R. W. (1997). *Affective Computing*. (R. W. Picard, Ed.)*Studies In Health Technology And Informatics* (Vol. 136, p. 292). MIT Press.

Schmidt, E. M., & Kim, Y. E. (2010a). Prediction of time-varying musical mood distributions from audio. *Information Retrieval* (pp. 465–470).

Schmidt, E. M., & Kim, Y. E. (2010b). Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. *2010 Ninth International Conference on Machine Learning and Applications*, *0*, 655–660.

Schmidt, E. M., & Kim, Y. E. (2011). Modeling musical emotion dynamics with Conditional Random Fields. *Information Retrieval*, *21*, 777–782.

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. *Proceedings of the international conference on Multimedia information retrieval* (pp. 267–274). ACM.

Schubert, E. (2004). Modeling Perceived Emotion With Continuous Musical Features. *Music Perception*, *21*(4), 561–585.

Schubert, E., Ferguson, S., Farrar, N., Taylor, D., & Mcpherson, G. E. (2012). Continuous Response to Music using Discrete Emotion Faces. *Proceedings of Computer Music Modeling and Retrieval* (pp. 3–19).

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. *Proceedings of International Symposium on Music Information Retrieval*, 549–554.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Differentiation, classification, and measurement. *Emotion*, *8*(4), 494–521.