# Predicting Emotional Prosody of Music with High-Level Acoustic Features

Jose Fornari & Tuomas Eerola

Finnish Centre of Excellence in Interdisciplinary Music Research
Department of Music
University of Jyväskylä, Finland

## BACKGROUND

The automatic prediction of emotional content in music is nowadays a growing area of interest. Several algorithms have been developed to retrieve music features and computational models using these features are continuously being created in order to predict music emotional content.

The literature mentions three main models for music emotions: 1) categorical model, originated from the work of [1], that describes music in terms of a list of basic emotions [2], 2) dimensional model, originated from the work of [3], who proposed that all emotions can be described in a Cartesian coordinate system of emotional dimensions [4], and 3) component process model, from the work of [5] that describes emotion appraised according to the situation of its occurrence and the current listener's mental (emotional) state. In a two-dimensional model, the dimensions normally addressed are: arousal (from calm to excited) and valence (from sad to happy). [6] used a two-dimensional model to measure the temporal development of music emotions appraised by listeners, for several music pieces. Then, he proposed a linear model with five acoustic descriptors to predict these dimensions in a time series analysis of each music piece. [7] applied the same listeners' mean ratings of [6], however, to develop and test a general model to describe the emotional dimensions for all music pieces. This model was created with System Identification techniques. These former models were successful in predicting arousal with a high degree of certainty, however the retrieval of valence has being particularly difficult to be predicted. This may be due to the fact that these models did not make extensive usage of high-level acoustic features: the cognitive musical features that are contextual-based and deliver one prediction for the overall music excerpt.

## AIMS

This work uses eight high-level descriptors previously developed to build a linear model for the dynamic prediction of emotional content, based on the dimensional model of music emotion. As arousal is already successfully predicted by low-level descriptors, such as RMS or loudness, here we focused on the prediction of valence.

We have previously developed high-level descriptors for the following contextual acoustic features: 1) Pulse Clarity (the sensation of pulse in music). 2) Key Clarity (the sensation of a tonal center). 3) Harmonic complexity (the sensation of complexity delivered by musical chords). 4) Articulation (from staccato to legato). 5) Repetition (presence of repeated musical patterns). 6) Mode (from minor to major tonalities). 7) Event Density (amount of distinctive and simultaneous musical events). 8) Brightness

(the sensation of musical brightness). For this development, behavioral data was also collected from listeners that rated the same features described above and their mean-rate was then correlated with the descriptors predictions.

The final model was then created using these high-level. This model prediction for valence is described and compared with the prediction of the previous models described in [6] and [7].

## METHOD

We have developed computational models for the eight high-level musical descriptors previously mentioned. The design of these descriptors was done using *Matlab*. Each one of them has a different technique and approach whose thorough explanations are too extensive to be part of this work. They use known techniques, such as spectral analysis, auto-correlation, chromagram analysis and auto-similarity matrix.

To test and improve the development of these descriptors, behavioral data was collected from thirty-three listeners that were asked to rate the same features that were predicted by those descriptors. They rated one hundred short excerpts of music (five seconds of length each) from movie sound tracks. Their mean-rating was then correlated with the descriptors predictions. After several experiments and adjustments, all descriptors presented a correlation with this ground-truth of: $r > 0.5$.

## RESULTS

In the temporal dynamics of emotion study described in [6], it was created a ground-truth with data collected from thirty-five listeners that dynamically measured the emotional ratings depicted into a two-dimensional emotion plan that was then mapped in two coordinates: arousal and valence. Listener's ratings were sampled every one second. The mean-rate of these measurements, mapped into arousal and valence, created a ground-truth that was used later by [7] and also in the present work. Here, the correlation between each high-level descriptor prediction and the valence mean-rate from this ground-truth was calculated. The valence mean-rate utilized was the one from the music piece called "Aranjuez concerto". This piece has duration of 2 minutes and 45 seconds. During its first minute, the guitar performs alone, then it is suddenly accompanied by a full orchestra whose intensity fades towards the end, till the guitar, once again, plays the main theme alone.

For this piece, the correlation coefficient presented between the high-level descriptors and the valence mean-rate are: event density: $r = -0.59$, harmonic complexity: $r = 0.43$, brightness: $r = -0.40$, pulse clarity: $r = 0.35$, repetition: $r = -0.16$, articulation: $r = -0.09$, key clarity: $r = -0.08$, mode: $r = -0.05$.

The multiple regression model was then created with all eight descriptors. This model employs a three-second time frame, which is related to the cognitive "now time" of music [8] and hop-size of one second to predict the continuous development of valence. This model presented a correlation coefficient $r = 0.65$, which leads to a coefficient of determination: $R^2 = 42\%$.

For the same ground-truth, in [6] the model used five music descriptors: 1) Tempo, 2) Spectral Centroid, 3) Loudness, 4) Melodic Contour and 5) Texture. The descriptors output differentiation was regarded as the model predictors. Using time series analysis, he built an ordinary least square (OLS) model for each music excerpt.

In [7], the model used eighteen low-level descriptors to test several models designed throughout system identification. The best generic model reported in his work was an ARX (autoregressive with extra inputs).

Table 1 shows the results comparison of $R^2$ for the measurement of valence in the Aranjuez concerto, for all models.

**Table 1.** Models comparison for the prediction of Valence. (Music subject: Aranjuez concerto).

| | **[6]** | **[7]** | **This work model** | **Event Density** |
|---|---|---|---|---|
| **Type** | OLS | ARX | Multiple Regression | One Descriptor |
| **$R^2$** | 33% | -88% | 42% | 35% |

This table shows that the model reported here performed significantly better than the previous ones for this specific music piece. The last column of table 1 shows the performance for the high-level descriptor "event density", the one that presented the highest correlation with the ground-truth. This descriptor alone presents higher results than previous models. The results shown seem to suggest that high-level descriptors can successfully help to improve the dynamic prediction of valence.

Figure 1 shows the comparison between the thirty-three listeners mean rating for valence in the Aranjuez concerto and the prediction given by the multiple regressive model using the eight high-level musical descriptors.

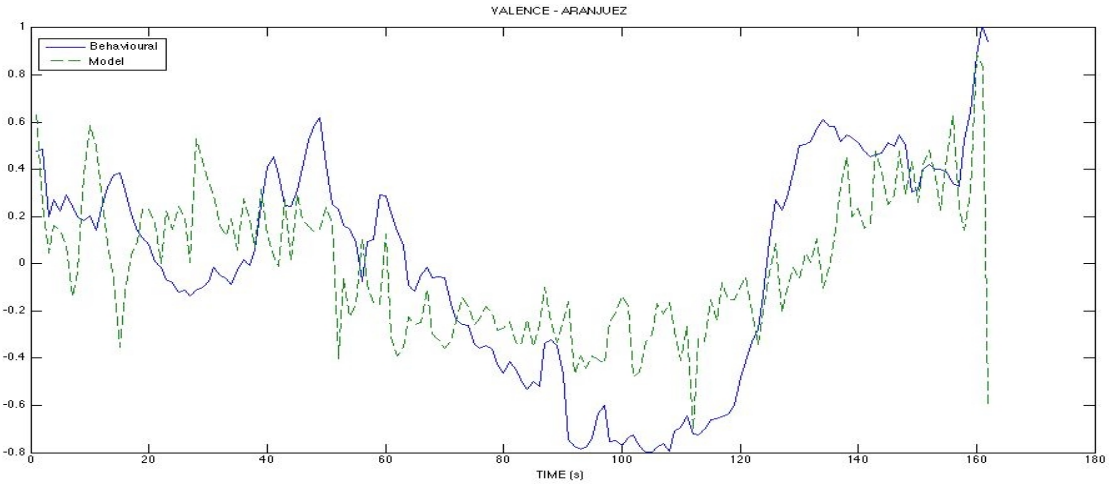


**Figure 1.** Listeners' mean-rating for valence (continuous line) and model prediction (dashed line).

## CONCLUSIONS

It was interesting to notice that the prediction of the high-level descriptor "event density" presented the highest correlation with the valence mean-rate, while the predictions of "key clarity" and "mode" correlated very poorly. This seems to imply that, at least in this particular case, musical sensation of a major or minor tonality (represented by "mode") or a tonal center ("key clarity") is not as related to valence as it might be previously inferred. What most prominent here, at least in this example, was the amount

of simultaneous musical events (given by “event density”). By “event”, it is understood here any perceivable rhythmic, melodic or harmonic pattern.

This experiment chose the music piece “Aranjuez” because it was the one that the previous models presented the lowest prediction correlation. Surely, more experiments are needed to be performed in order to pursuit more evidences on the prediction of valence by this model. Nevertheless, we believed that this work in fact presented an interesting prospect for the prediction of contextual music features with high-level descriptors. We hope that this result kindles the development of better high-level descriptors and models for the continuous measurement of contextual musical features such as valence.

## REFERENCES


1. Ekman, P.: An argument for basic emotions. Cognition & Emotion, 6 (3/4): 169–200, (1992).

2. Juslin, P. N., & Laukka, P.: Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin(129), 770-814. (2003)

3. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological Review Vol. 110, No. 1, 145- 172. (2003)

4. Laukka, P., Juslin, P. N., & Bresin, R.: A dimensional approach to vocal expression of emotion. Cognition and Emotion, 19, 633-653. (2005)

5. Scherer, K. R., & Zentner, K. R.: Emotional effects of music: production rules. In J. P. N. & J. A. Sloboda (Eds.), Music and emotion: Theory and research (pp. 361-392). Oxford: Oxford University Press (2001)

6. Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensional emotion space. Aust. J. Psychol., vol. 51, no. 3, pp. 154–165. (1999)

7. Korhonen, M., Clausi, D., Jernigan, M.: Modeling Emotional Content of Music Using System Identification. IEEE Transactions on Systems, Man and Cybernetics. Volume: 36, Issue: 3, pages: 588- 599. (2006)

8. Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M.: Correlation of Gestural Musical Audio Cues. Gesture-Based Communication in Human-Computer Interaction. 5th International Gesture Workshop, GW 2003, 40-54. (2004)