EMOTIONALLY-CONTROLLED MUSIC SYNTHESIS

ANTÓNIO PEDRO OLIVEIRA[1], AND AMÍLCAR CARDOSO[2]

[1] *Centre for Informatics and Systems (CISUC), University of Coimbra, Coimbra, Portugal*
apsimoes@student.dei.uc.pt
amilcar@dei.uc.pt

Abstract. We are implementing a system that automatically produces music with an appropriate affective content. One part of this system consists in the synthesis of music that is emotionally-controlled by experimentally-derived weighted mappings between emotions and audio features. 80 different listeners were asked to label online 2 affective dimensions (arousal and valence) of 96 musical samples with values selected from the integer interval between 0 and 10. With the help of this data, mappings were established between emotions and audio music features (e.g., spectral dissonance, spectral similarity and spectral sharpness) and between these and symbolic music features. Several non-linear regression models, using both audio and symbolic music features, were tested in the classification of music emotion. Classification results of the models obtained in this work outperformed results of previous linear regression models that used only symbolic music features.

Resumo. Estamos a implementar um sistema que produz automaticamente musica com um conteúdo afectivo desejado. Uma parte deste sistema consiste na sintese de música que é controlada emocionalmente por mapeamentos pesados, obtidos experimentalmente, entre emoções e características musicais audio. 80 ouvintes etiquetaram online 2 dimensões afectivas (activação e valência) de 96 musicas com valores selecionados do intervalo inteiro entre 0 e 10. Com a ajuda destes dados, foram estabelecidos mapeamentos entre emoções e características musicais audio (e.g., dissonância espectral, semelhança espectral e brilho espectral) e entre estas e características musicais simbólicas. Foram testados vários modelos de regressão não-linear, que usaram tanto características musicais audio como simbólicas, na classificação das emoções da música. Os resultados de classificação dos modelos obtidos neste trabalho foram melhores que os resultados obtidos com os anteriores modelos de regressão linear que usaram apenas características musicais simbólicas.

Keywords: Affective content of sound, audio features, music synthesis, regression model, emotions and music.

INTRODUCTION

Music has been widely accepted as one of the languages of emotional expression. The possibility to produce music with an appropriate affective content can be helpful to adapt music to our affective interest. However, only recently scientists have tried to quantify and explain how music expresses certain emotions. As a result of this, mappings are being established between affective dimensions and music features [5, 11, 13]. Most psychology researchers agree that affect has at least two distinct qualities [15]: valence (degree of satisfaction) and arousal (degree of activation), so we are considering these 2 dimensions.

We intend to model the relation between emotions and music features with regression models that can control the affective content of synthesized music (Section 2). This is being done with the help of a review of research works on this area (Section 1). Our work extends the modeling of emotions from a symbolic domain [13] to an audio domain, by extracting audio features (Section 3). We also extend the regression model from a linear to a non-linear domain (Section 4). Section 5 analyses the results, and finally, section 6 makes some final remarks.

1 RELATED WORK

This work entails an interdisciplinary research involving Music Psychology and Music Information Retrieval. This section makes a review of some of the most relevant contributions for our work from these areas.

We are undertaking the modeling of affective content of music as a regression problem, so special attention is devoted to works that face this challenge with similar techniques. Mosst [11] used quantitative techniques to model emotional perception in music. Several listeners made time-varying emotion annotations. Musical features were extracted: loudness, spectral centroid, onset density, articulation and mode features. Then, multiple linear regression was used to relate emotional annotations to extracted musical features. With similar objectives, Friberg et al. [4] designed a model to predict the expressive intention during music performance. Musical features like average and variability values of sound level, tempo, articulation, attack velocity and spectral content were extracted. Listening experiments served to build linear regression models to predict intended emotion based on musical features. Similarly, Korhonen [6] modeled people perception of emotion in music with ARX (Auto-Regression with eXtra inputs) and State-Space models. These models tested the output (valence and arousal) using 20 subsets of musical features as input. Scheirer [16] also developed a framework to analyze and model aspects of music perception with the help of audio musical features.

Juslin and Lindstrom [5, 8] and Shanley [17] obtained weights of importance of musical features in the emotional expression. Carvalho and Chao [2] used a sequential stack classifier and a feature set with timbral texture features. This classifier outperformed other classifiers like decision trees, logistic regression and conditional random fields. 2-label classification obtained a success of 86%, 5-label classification achieved a success of 36%.

2 COMPUTATIONAL MODEL

The system described in this paper (Figure 1) intends to model the relation between emotions and music features with regression models that can control the affective content of synthesized music. The system uses a database of MIDI music labeled with symbolic features and audio features of the corresponding synthesized audio.

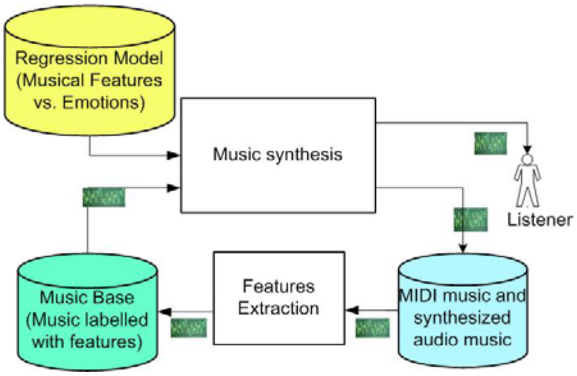
Regression Model (Musical Features vs. Emotions)
Music synthesis
Listener
Music Base (Music labelled with features)
Features Extraction
MIDI music and synthesized audio music

Figure 1: Diagram of the computational model

For the experiments with listeners, we used a set of 96 musical pieces. These pieces were of western tonal music (film music), last, approximately, from 20 seconds to 1 minute, and were used for both training and validating the classifier. 80 different listeners were asked to label online each affective dimension of the musical pieces with values selected from the integer interval between 0 and 10 [0; 10][1]. The obtained affective labels were used to make the regression models. This was done separately for the valence and arousal.

3 FEATURES EXTRACTION

The extraction of features is being done with the adaptation of available algorithms of third party software working in audio (MIR toolbox [7] and PsySound [1]) and symbolic domain (JSymbolic [9] and MIDI toolbox [3]). The importance of symbolic features in the emotional expression was analysed in previous papers [12, 13, 14]. This work intends to understand the importance of audio features in the emotional expression, as well as to understand their relation with symbolic features.

The correlation coefficient between valence/arousal and the best 6 audio features is ~61%/~75%. Separately the correlations are: spectral sharpness (~42%/~36%), spectral dissonance (~28%/~49%), loudness (~41%/~28%), spectral similarity (~-26%/~-58%) timbral width (~32%/~29%) and tonal dissonance (~-25%/~-29%). These coefficients are not very high; however, when analyzed with symbolic features, like average note duration and note density, they contribute to an increase in the correlation coefficient of these features with the affective dimensions. From previous research [13, 14] and from this work, we can infer that timbre/sound is an important musical feature to control/influence the emotional expression. Table 1 and table 2 present the correlation between several audio features [1, 7] and two affective dimensions, respectively, valence and arousal.

Corr.	Audio features
42%	Spectral Sharpness (A) Texture
28%	Spectral Dissonance (S) Texture
37%	Spectral Sharpness (Z) Texture
32%	Timbral Width (Spectral Flatness) Texture
-22%	Volume (size of the sound) Texture
-25%	Tonal Dissonance (S) Texture
-4%	Spectral Dissonance (H&K) Texture
-11%	Tonal Dissonance (H&K) Texture
41%	Loudness Texture
-26%	Spectral Similarity Texture
21%	Brightness (>1500Hz) Texture
17%	Brightness (>4000Hz) Texture
6%	Brightness (>400Hz) Texture
5%	Inharmonicity Texture
13%	Harmonic Mode Texture
23%	Energy Texture
-4%	ADSR Envelope Texture
12%	Register Texture

Table 1: Correlation coefficients between audio features and valence.

[1] http://student.dei.uc.pt/%7Eapsimoes/PhD/Music/smc08/index.html

Corr.	Audio features
36%	Spectral Sharpness (A) Texture
49%	Spectral Dissonance (S) Texture
34%	Spectral Sharpness (Z) Texture
29%	Timbral Width (Spectral Flatness) Texture
-26%	Volume (size of the sound) Texture
-29%	Tonal Dissonance (S) Texture
17%	Spectral Dissonance (H&K) Texture
-6%	Tonal Dissonance (H&K) Texture
28%	Loudness Texture
-58%	Spectral Similarity Texture
18%	Brightness (>1500Hz) Texture
21%	Brightness (>4000Hz) Texture
-3%	Brightness (>400Hz) Texture
6%	Inharmonicity Texture
8%	Harmonic Mode Texture
29%	Energy Texture
-13%	ADSR Envelope Texture
2%	Register Texture

Table 2: Correlation coefficients between audio features and arousal.

We also tried to bridge the gap between the symbolic domain and audio domain, in order to know the influence of specific symbolic features on audio features. Table 3 presents correlation coefficients between audio features with a high importance on affective discrimination (spectral similarity, spectral dissonance and spectral sharpness) and symbolic features.

Corr.	Audio and symbolic features
61%	Spectral similarity and average duration accent
50%	Spectral similarity and average note duration
44%	Spectral similarity and average time between attacks
43%	Spectral similarity and variability of time between attacks
42%	Spectral similarity and strength of strongest rhythmic pulse
46%	Spectral dissonance and variability of note prevalence of unpitched instruments
45%	Spectral dissonance and percussion prevalence
43%	Spectral dissonance and number of unpitched instruments
42%	Spectral dissonance and bass drum prevalence
41%	Spectral dissonance and melodic complexity
41%	Spectral sharpness and harpsichord fraction
40%	Spectral sharpness and number of unpitched instruments
35%	Spectral sharpness and variability of note prevalence of unpitched instruments
33%	Spectral sharpness and climax position

Table 3: Correlation coefficients between relevant audio and symbolic features.

4 REGRESSION MODEL

Regression models are mathematical models useful to establish weighted relations between dependent and independent variables. In our work these models have been useful is establishing relations between emotions and musical features. Previous work only compared the classification performance of the modeling of symbolic features with linear regression models [12, 13]. This work, presents classification results obtained by using non-linear regression models and by using both symbolic and audio features. We have used Strijov's algorithm [18] to search for the best non-linear regression model. In each generation of this algorithm, we have used the correlation coefficient between the emotional data of the listeners and the emotional output given by the model as a fitness function.

For the following results we have used the set of 96 musical pieces as a train and test set. For valence the non-linear regression models improved the classification performance in 9% (from 75% of the linear model to 84% of the best non-linear model). For arousal the non-linear regression models improved the classification performance in 6% (from 84% of the linear model to 90% of the best non-linear model). Next, we present the results for valence (table 4) and arousal (table 5) of the best non-linear regression models and the best previous linear model. To validate these results we have tested these models in 8 random sub-groups of 12 musical pieces, by using 8-fold cross validation of the classification results. We also trained and validated non-linear regression models only with the most affective discriminant audio features: spectral sharpness (A), spectral dissonance (S), loudness, spectral similarity, timbral width and tonal dissonance (S). Results of these models are in table 4 for valence and table 5 for arousal.

Corr.	Regression model
84%	w1*X1 + w2*X4 - parabola(X3) - w3*(x1+sin(X3)) + parabola(sin(X6)) + w4*X6
83%	w1*X1 + w2*X4 + $\sqrt{X3}$ + w3*sin(sin(X5)) + sin(X3) + parabola(sin(X2)) + w4*X6
83%	w1*X1 + w2*X4 + parabola(X3) + w3*(sin(sin(X5)) + sin(sin(X1))) + parabola(sin(X6)) + w4*X6
...	...
75%	w1*X1 + w2*X2 + w3*X3 + w4*X4 + w5*X5 + w6*X6 (linear model)
61%	sin(X4) + w3*X3 + w4*X4 + parabola(X5) + X4 (audio features)

Table 4: Correlation coefficients of the best regression models for valence.

Corr.	Model
90%	X1 + w1*parabola(X1) + w2*(X2 + X1 + w3*X4 + w4*X5 + w5*sin(X1) + w6*X3)
89%	parabola(X1) + w1*X4 + w2*(w3*sin(X1) + parabola(X2) - w4*sin(X1)) + X2 +w5*X5 + w6*X6
89%	(w1*X2 + parabola(w2*X3) + parabola(X1) + X5 + w3*X5 + w4*X4) * parabola(w5*X3) + w6*X1
...	...
84%	w1*X1 + w2*X2 + w3*X3 + w4*X4 + w5*X5 + w6*X6 (linear model)
75%	w1*X5 + sin(X4) + w2*(w3*X4 + w4*X4) + w5*(sin(X2) + w6*X2) (audio features)

Table 5: Correlation coefficients of the best regression models for arousal.

5 DISCUSSION

Brightness (number of sinusoids with high frequencies), spectral dissonance (sinusoids within a critical band that originate sensory dissonance), energy, spectral similarity and spectral flatness are audio features of music samples that can be changed in order to influence, positively or negatively, the valence and arousal of music. These audio features have a relation with symbolic features: on the one hand, spectral dissonance seems to be related to the use of percussion instruments; on the other hand, spectral similarity seems to be related to the use of notes with high durations and the large use of silence between notes.

In the modeling of the valence/arousal of music, the best hybrid (use of audio and symbolic features) non-linear regression model achieved a correlation of 84%/90%, the best symbolic linear regression model achieved a correlation of 75%/84% and the best audio non-linear regression model achieved a correlation of 61%/75%.

6 CONCLUSION

In this work, we tested the importance of the use of both symbolic and audio features in the classification of music emotions, as well as the importance of the use of non-linear regression models instead of linear regression models. Classification results of non-linear regression models, using both audio and symbolic music features, outperformed results of previous linear regression models that used only symbolic music features [12, 13]. The use of non-linear regression models, instead of linear regression models, as well as the use of both audio and symbolic features seems to be an important factor to increase the efficacy in the classification of music emotion [10]. From previous research [13, 14] and from this work, we can also infer that timbre/sound is an important musical feature that can be used to control/influence the emotional expression in music.

REFERENCES

[1] Cabrera, D. (1999). “Psysound: A computer program for psychoacoustical analysis.” *Australian Acoustical Society Conference*, vol. 24, pp. 47–54.

[2] Carvalho, V.R. and Chao, C. (2005). “Sentiment retrieval in popular music based on sequential learning”. *Conference on Research and Development in Information Retrieval*, vol. 28.

[3] Eerola, T. and Toiviainen, P. (2004). “Mir in matlab: The midi toolbox.” *Int. Conf. on Music Information Retrieval*.

[4] Friberg, A., Schoonderwaldt, E., Juslin, P. and Bresin, R. “Automatic real-time extraction of musical expression”. *International Computer Music Conference*, pp. 365–367.

[5] Juslin, P. and Lindström, E. (2003). “Musical expression of emotions: Modeling composed and performed features”. *ESCOM Conference.*

[6] Korhonen, M. (2004). *Modeling continuous emotional appraisals of music using system identification.* Master’s thesis, University of Waterloo.

[7] Lartillot, O. and Toiviainen, P. (2007). “MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio”. *International Conference on Music Information Retrieval,* pp. 237-244.

[8] Lindstrom, E. (2004) *A Dynamic View of Melodic Organization and Performance*. PhD thesis, Acta Universitatis Upsaliensis Uppsala.

[9] McKay, C. and Fujinaga, I. (2006). “Jsymbolic: A feature extractor for midi files.” *International Computer Music Conference*.

[10] McKay, C. and Fujinaga, I. (2008). “Combining features extracted from audio, symbolic and cultural sources.” *International Conf. on Music Information Retrieval*.

[11] Mosst, M. (2006) *Quantitative modeling of emotion perception in music*. Master’s thesis, University Of Southern California.

[12] Oliveira, A., Cardoso, A. (2008). ”Towards bi-dimensional classification of symbolic music by affective content” *International Computer Music Conference*.

[13] Oliveira, A., Cardoso, A. (2008). ”Modeling Affective Content of Music: A Knowledge Base Approach” *Sound and Music Computing Conference*.

[14] Oliveira, A., Cardoso, A. (2008). ”Affective-Driven Music Production: Selection and Transformation of Music” *International Conference on Digital Arts*.

[15] Russell, J.A. (1989). ”Measures of emotion.” Emotion: Theory, research, and experience, vol. 4, pp. 83-111.

[16] Scheirer, E. (2000). *Music-Listening Systems*. PhD thesis, Massachusetts Institute of Technology.

[17] Shanley, P. (2004). *Music and emotion: The creation of a continuous response network in order to evaluate the extent of specific musical element expressiveness and the make-up of music’s affective personality*.

[18] Strijov, V. (2007). “Search for a parametric regression model in an inductive-generated set.” *Journal of Computational Technologies*, vol. 1, pp. 93-102