MUSIC EMOTION CLASSIFICATION: DATASET ACQUISITION AND COMPARATIVE ANALYSIS

Renato Panda,

CISUC, Department of Informatics
Engineering, University of Coimbra
Coimbra, Portugal
panda@dei.uc.pt

Rui Pedro Paiva,

CISUC, Department of Informatics
Engineering, University of Coimbra
Coimbra, Portugal
ruipedro@dei.uc.pt

ABSTRACT

In this paper we present an approach to emotion classification in audio music. The process is conducted with a dataset of 903 clips and mood labels, collected from Allmusic[1] database, organized in five clusters similar to the dataset used in the MIREX[2] Mood Classification Task. Three different audio frameworks – Marsyas, MIR Toolbox and Psysound, were used to extract several features. These audio features and annotations are used with supervised learning techniques to train and test various classifiers based on support vector machines. To access the importance of each feature several different combinations of features, obtained with feature selection algorithms or manually selected were tested. The performance of the solution was measured with 20 repetitions of 10-fold cross validation, achieving a F-measure of 47.2% with precision of 46.8% and recall of 47.6%.

1. INTRODUCTION

Over the years music has always been connected to us, in many areas from entertainment to war or religion, serving a myriad of purposes both socially and individually. Given the major importance of music in all human societies throughout history and particularly in the digital society, music plays a relevant role in the world economy.

During the last decades, the technological innovations brought by the digital era created a complete shift in the way music is presented to and consumed by users. Factors like the widespread access to the Internet, bandwidth increasing in domestic accesses or the generalized use of compact high quality audio formats, such as mp3, have given a great contribution to that change. The frenetic growth in music supply and demand uncovered the need for powerful methods for automatically retrieving useful and relevant songs in a given context from such huge databases. In fact, any large music database, or, generically speaking, any multimedia database, is only really useful if users can find what they are seeking in an efficient manner. Furthermore, it is also important that the organization of such a database can be performed as objectively and efficiently as possible.

Digital music repositories need, then, more advanced, flexible and user-friendly search mechanisms, adapted to the requirements of individual users. In fact, “music’s preeminent functions are social and psychological”, and so “the most useful retrieval indexes are those that facilitate searching in conformity with such social and psychological functions. Typically, such indexes will focus on stylistic, mood, and similarity information” [1]. This is supported by studies on music information behavior that have identified music mood[3] as an important criterion for music retrieval and organization [2].

Besides the music industry, the range of applications of mood detection in music is wide and varied, e.g., game development, cinema, advertising or the clinical area (in the motivation to compliance to sport activities prescribed by physicians, as well as stress management).

Music Emotion Recognition (MER) is still a recent research topic. Even though the area has received increased attention in recent years, many limitations and open problem can be found, particularly on emotion detection in audio music signals. In fact, the present accuracy of current audio MER systems shows there is plenty of room for improvement.

In the most recent comparisons between current state of the art approaches, the best audio MER algorithm achieved a classification accuracy of 69%[4], highlighting once more the complexity of music emotion recognition and that there is a lot of work still to be done.

Several aspects make music emotion recognition a challenging task. On one hand, the perception of the emotions evoked by a song is inherently subjective: different people often perceive different, sometimes opposite, emotions. Besides, even when listeners agree in the kind of emotion, there’s still much ambiguity regarding its description (e.g., the adjectives employed). Additionally, it is not yet well-understood how and why music elements create specific emotional responses in listeners [3].

One of the main problems in the field is the lack of a standard, good quality dataset with audio clips and emotional information. Due to this fact, each author has to gather his own dataset, presenting results based on it, making it impossible to compare results obtained between different studies. Some efforts have been developed to address this problem, the most notable is probably the Music Information Retrieval Evaluation eXchange (MIREX), an annual evaluation campaign for Music Information Retrieval (MIR) algorithms, coupled to the International Society (and Conference) for Music Information Retrieval (ISMIR). Audio music mood classification is one of the tasks included in MIREX, where researchers can submit their classification systems to be tested and ranked. The process is conducted using a collection of 600 30 second audio clips in 22.05kHz mono WAV format selected from the APM collection, and labeled by human judges using the Evalutron6000 system. [22], divided in five

[1] http://www.allmusic.com/explore/moods

[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME

[3] Even though mood and emotion can be defined differently, the two terms are used interchangeably in the literature and in this paper. For further details, see [4].

[4] http://www.music-ir.org/nema_out/mirex2011/results/act/mood_report/index.html

clusters and several subcategories[5]: Cluster 1: passionate, rousing, confident, boisterous, rowdy; Cluster 2: rollicking, cheerful, fun, sweet, amiable/good natured; Cluster 3: literate, poignant, wistful, bittersweet, autumnal, brooding; Cluster 4: humorous, silly, campy, quirky, whimsical, witty, wry; Cluster 5: aggressive, fiery, tense/anxious, intense, volatile, visceral. However, the dataset is secret and exclusive to the MIREX contests, thus unavailable to anyone.

In this paper our first goal was to create a dataset with the same organization used in MIREX, automatically gathering a set of 903 30-second clips from Allmusic, labeled with the exact same 5 categories. The resultant set was later used in three different audio frameworks to extract features, employed in the train and test phases of our classification system. The system presented is based on supervised learning techniques, using support vector machines (SVM) as the classifier. Results measurement and validation is done recurring to 20 repetitions of 10-fold cross validation. This process consists in dividing the dataset in 10 balanced subsets (folds), using nine folds to train the model and the remaining to test, repeating until all folds were used in the test process. Several other solutions were also tested, namely feature selection algorithms such as ReliefF [5] and principal component analysis (PCA), in order to reduce the feature space and extract information about the most relevant features and frameworks to emotion recognition.

This paper is organized as follows. In section 2, the relevant work that has been done in the area is described. Next, section 3 introduces the audio frameworks and feature extraction process that were employed, followed by details on the classification strategy. In section 4, the quality of our dataset is analyzed and experimental results are presented and discussed. Finally, conclusions from this study are drawn in section 5

2. RELATED WORK

Mood and emotions has been a major subject of psychologists for a long time and so several theoretical models have been proposed over the years. Such models can be grouped into two major approaches: categorical models or dimensional models. Categorical models consist of several categories or states of emotion, such as anger, fear, happiness and joy. Dimensional models, on the other hand, use several axes to map emotions into a plan. The most frequent approach uses two axes (e.g. arousal-valence (AV) or energy-stress), with some cases of a third dimension (dominance).

The advantage of dimensional models is the reduced ambiguity when compared with the categorical approach. However, some ambiguity remains, since each of the four quadrants represents more than one distinct emotion (happiness and excitation are both represented by high arousal and valence for example). Given this, dimensional models can be further divided into discrete (representing the ones described above) and continuous. Continuous models, unlike discrete ones, view the emotion plan as a continuous space where each point denotes a different emotional state. As a result, all ambiguity related with emotion states is removed [3].

In this paper a categorical model of emotion is used, based in the five clusters defined in MIREX. The aim of the study was to understand the importance of the current features and audio frameworks in music mood classification problems and to compare against other approaches, indentifying important musical features. The choice of a categorical model is inevitable, since we replicated MIREX dataset to compare results. To the date there is no contest being held using dimensional models of emotion. However, it is something gaining momentum and some studies have used it [3], [6] and [7], due to its elimination of ambiguity, existent in categorical models as well as the differentiation between songs that is possible by using distinct discrete values representing different emotions.

Research on the relations between music and emotion has a long history, with initial empirical studies starting in the 19th century [8]. This problem was studied more actively in the 20th century, when several researchers investigated the relationship between emotions and particular musical attributes such as mode, harmony, tempo, rhythm and dynamics [2]. However, only a few attempts have been made to build computational models. From these, most are devoted to emotion synthesis [9], elaborating on the relationships between emotion and music composition and music expressivity.

Only a few works addressing emotion detection in audio signals can be found. To the best of our knowledge, the first paper on mood detection in audio was published in 2003, by Feng et al. [10]. There, musical pieces are classified into 4 mood categories (happiness, sadness, anger and fear) using two musical attributes: tempo and articulation. Tempo is obtained via a beat-tracking system, from which a feature termed relative tempo is derived. Articulation is computed through the average silence ratio feature. These features are used to train a neural network classifier using 200 musical pieces. The classifier is then validated on a test corpus of 23 pieces, with average precision, the fraction of songs correctly identified, and recall, measuring the fraction of relevant songs identified, of 67 and 66%, respectively. This first attempt towards audio music mood detection suffered, naturally, from several limitations. Namely, only 2 music attributes were captured, only 4 mood labels were employed and, regarding system validation, a reduced test corpus was utilized, making it hard to provide evidence of generality.

Most of the described limitations were still present in following research works (e.g., [3], [11] and [12]). Overall, in each approach a different (and limited) set of features, mood taxonomies, number of classes and test sets are proposed. Also, some studies constrain the analysis to a particular musical style, e.g., [3], [12].

One of the most comprehensive approaches so far is the one proposed by Lu et al. [3]. They based their system on Thayer's model of mood, employing its four quadrants. Features of intensity, timbre and rhythm are used. Mood is then detected with recourse to a hierarchical framework based on Gaussian Mixture Models and feature de-correlation via the Karhunen-Loeve Transform. The algorithm was evaluated on a test set of 800 songs, reaching 86.3% average precision. This value should be regarded with caution, since the system was only evaluated on a corpus of classical music using only 4 classes. Its main limitations are the absence of important mood-related features, such as mode and articulation, and its short number of mood categories. One of its interesting points, besides the use of a hierarchical framework, is the assignment of different weights to features, so as to represent their importance in each class.

Several other studies approaching various MER problems have been proposed in the last years as the research in field keeps evolving. Among these, Wang et al. [13] proposed an audio classification system using a semantic transformation of the feature vectors based on music tags and classifier ensemble, obtaining interesting results in the MIREX 2010 competition. In another study, Schmidt et al [14] studied the emotion variations over time

[5] http://www.music-ir.org/mirex/wiki/2011:Audio_Classification_%28Train/Test%29_Tasks#Audio_Mood_Classification

in a musical piece, using dimensional models. In the study, it was demonstrated that conditional random fields can be an effectively tool for modeling time-varying musical emotion. Also, a few recent studies have proposed multi-model approaches, combining different strategies for mood detection. One example of this was proposed by McVicar et al [15] is the detection of mood by identifying common characteristics between lyrics and audio features.

3. METHODOLOGY

3.1. Audio Feature Extraction

Several authors have studied the most relevant musical attributes for mood analysis. Namely, it was found that major modes are frequently related to emotional states such as happiness or solemnity, whereas minor modes are associated with sadness or anger [4]. Simple, consonant, harmonies are usually happy, pleasant or relaxed. On the contrary, complex, dissonant, harmonies relate to emotions such as excitement, tension or sadness, as they create instability in a musical piece [4]. In a recent overview, Friberg [2] lists and describes the following features: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality and rhythm. Other common features not included in that list are, for example, mode, loudness or musical form [4]. Several of these features have already been studied in the MIDI domain, e.g., [16]. The following list contains many of the relevant features for music mood analysis:

- Timing: Tempo, tempo variation, duration contrast
- Dynamics: overall level, crescendo/decrescendo, accents
- Articulation: overall (staccato/legato), variability
- Timbre: Spectral richness, onset velocity, harmonic richness
- Pitch (high/low)
- Interval (small/large)
- Melody: range (small/large), direction (up/down)
- Harmony (consonant/complex-dissonant)
- Tonality (chromatic-atonal/key-oriented)
- Rhythm (regular-smooth/firm/flowing-fluent/irregular-rough)
- Mode (major/minor)
- Loudness (high/low)
- Musical form (complexity, repetition, new ideas, disruption).

However, many of the previous features are often difficult to extract from audio signals. Also, several of them require further study from a psychological perspective. Therefore, it is common to directly apply low-level audio descriptors (LLDs), studied in other contexts (e.g., genre classification, speech recognition), directly to mood detection. Such descriptors aim to represent attributes of audio such as pitch, harmony, loudness, timbre, rhythm, tempo and so forth. LLDs are generally computed from the short-time spectra of the audio waveform, e.g., spectral shape features such as centroid, spread, bandwidth, skewness, kurtosis, slope, decrease, rolloff, flux, contrast or MFCCs [4]. Other methods have been studied to detect tempo and tonality.

There are various audio frameworks available that can be used to process songs and extract audio features. These frameworks have several differences, namely the number and type of features available, stability, ease of use, performance and the system resources they require. In this work, features from PsySound 3, MIR Toolbox and Marsyas were used and their results compared, in order to study their importance and how their feature sets are valuable in MER.

PsySound 3 is a MATLAB toolbox for the analysis of sound recordings using physical and psychoacoustical algorithms. It does precise analysis using standard acoustical measurements, as well as implementations of psychoacoustical and musical models such as loudness, sharpness, roughness, fluctuation strength, pitch, rhythm and running IACC. Although PsySound is cited in some literature [3] as having several emotionally-relevant features, there are few works using this framework, possibly due to its slow speed and problems – some of the most interesting features such as tonality do not work properly, outputting the same value for all songs, or simply crash the framework.

The MIR (Music Information Retrieval) toolbox is an integrated set of functions written in MATLAB, that are specific to the extraction of musical features such as pitch, timbre, tonality and others [17]. A high number of both low and high-level audio features are available.

Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) is a software framework developed for audio processing with specific emphasis on MIR applications. Marsyas has been used for a variety of projects in both academia and industry, and it is known to be computationally efficient, due in part to the fact of being written in highly optimized C++ code. On the less bright side, it lacks some features considered relevant to MER.

The Marsyas framework provides a set of example applications that can be used for a variety of different audio processing tasks. One of these applications is the feature extraction tool used in previous editions of MIREX. Since the results of those editions are known, using this tool in our experiment and comparing how it performs against our dataset may serve as a possible point of comparison between our dataset and the MIREX dataset. To this end we extracted various sets of features with Marsyas, one of them with the tool, extracting only 65 features.

A brief summary of the features extracted and their respective framework is given in Table 1. Regarding Marsyas the analysis window for frame-level features is 512 samples, MIR toolbox was used with the default window size of 0.05 seconds. These frame-level features are integrated to song-level features by the MeanVar model [18], which represents the feature by mean and variance. All extracted features were normalized to the [0, 1] interval.

Table 1: *Used frameworks and respective features.*

Framework	Features
Marsyas (65)	Centroid, rolloff, flux, Mel frequency cepstral coefficients (MFCCs), and tempo.
MIR toolbox (177)	Among others: Root mean square (RMS) energy, rhythmic fluctuation, tempo, attack time and slope, zero crossing rate, rolloff, flux, high frequency energy, Mel frequency cepstral coefficients (MFCCs), roughness, spectral peaks variability (irregularity), inharmonicity, pitch, mode, harmonic change and key.
PsySound 3 (11)	Loudness, sharpness, timbral width, spectral and tonal dissonances, pure tonalness, multiplicity.

3.2. Classification Approach and Evaluation

There are numerous classification methods available and many are commonly in the MER area. The idea behind classification is to predict the class of a selected sample, based on a previous set of training examples used to create a model, or in some cases directly using all the available examples without training (for instance, K-nearest neighbours).

Based on this, a classifier algorithm was used to train a model based on the feature vectors extracted from the dataset as well as the cluster labels, gathered from Allmusic database. These trained models can then be fed with new feature vectors, returning the predicted classes for them.

Support Vector Machines (SVM) was the preferred classification algorithm, according to the quality report in previous comparative studies [3], [13]. The libSVM library [19] was the selected implementation, providing a fast and reliable implementation of SVMs. A grid parameter search was also carried out to retrieve the best values for parameters γ and C (cost), used by the radial basis function (RBF) kernel of the SVM model. Some additional tests using other algorithms such as k-nearest neighbours (KNN) were conducted with lower results.

In order to reduce the feature set and achieve a subset of features that are most suitable to our problem, ReliefF [5] was used. ReliefF is a feature selection algorithm that does not assume feature independence. In addition, it also provides a weight to each feature in the problem under analysis. The algorithm uses KNN, and thus one of the important parameters to select is a proper value of K. Using a small value may give unreliable results. On the other hand, if K is high it may fail to highlight important features. Taking this into consideration, several values of K were tested, later using a feature rank based on the average weights obtained with all the values tested. This features' ranking and weight was then used in order to calculate the number $\boldsymbol{m}$ of features needed to obtain the best results (top $\boldsymbol{m}$ features) with 20 repetitions of 10-fold cross validation. Forward feature selection (FFS) algorithm was also initially tried but discarded due to its slow speed with such a large dataset and considerable number of features. One of its main limitations is the fact that it does not take into consideration the relation that might exist between groups of features.

The dimensionality of the feature space can also be reduced with recourse to Principal Component Analysis (PCA) [20]. This is a widely used technique whose basic idea is to project the computed feature matrix into an orthogonal basis that best expresses the original data set. Moreover, the resulting projected data is decorrelated. As for the selection of the principal components, we kept the ones that retained 90% of the variance. Regarding implementation, we made use of the PCA MATLAB code provided in the Netlab toolbox [21].

In the previous years, MIREX audio train/test classification results were presented in terms of classification accuracy, the matching ratio or precision between correct and total predictions. Since precision only measures quality or exactness of the prediction, our results are presented in terms of F-measure or F-score, the harmonic mean between precision and recall – a measure of completeness. In simple terms, a high precision means the system classified more songs correctly than incorrectly. High recall means the system classified most of the correct songs. F-measure of class $\boldsymbol{i}$, $\boldsymbol{F_i}$, is defined as follows (1):

$$F_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{1}$$

where $\boldsymbol{P}$ in equation (1) represents precision and is defined as (2):

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

and $\boldsymbol{R}$ in equation (1) represents recall and is defined as (3):

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

In the equations 2 and 3, $\boldsymbol{TP_i}$, $\boldsymbol{FP_i}$ and $\boldsymbol{FN_i}$ represent respectively the true positives – number of songs correctly classified as belonging to class $\boldsymbol{i}$, false positives – songs incorrectly classified as belonging to class $\boldsymbol{i}$, and false negatives – songs which were not classified as belonging to class $\boldsymbol{i}$ but that should have been.

4. EXPERIMENTAL RESULTS

4.1. Dataset Collection and Analysis

The dataset used in the study was built based on the known characteristics of the MIREX dataset. To this end, a Python robot was created to automatically gather 30 second audio clips and music mood tags from the Allmusic website. Initially, a total of 1335 clips belonging to the 29 categories described in MIREX were obtained and organized in the same five clusters. Next, the dataset was reduced to 903 clips by ignoring the clips which appeared in several categories. This ambiguity happens due to the subjectivity in MER and also due to the way information in Allmusic is collected – user contribution based. Some of the clips available there are associated with more than one category between different clusters. In the case of MIREX mood dataset, songs were labeled based on the agreement between three experts [22].

Although the two datasets have some similarities in organization, they still differ in important aspects such as the annotation process and results must be analysed or compared with this in mind. One possible solution to get an acceptable idea of the similarity between both datasets would be to have results of the same system for both, something that we will consider in the next edition on MIREX.

In terms of distribution, the dataset is relatively balanced between clusters, with a slight advantage for clusters 3 and 4, as shown in figure 1, due to the removal of the ambiguous songs.

Another relevant aspect of the dataset is that, as pointed out in a few studies, there is a semantic overlap (ambiguity) between clusters 2 and 4, and an acoustic overlap between clusters 1 and 5 [23]. For illustration, the word fun (cluster 2) and humorous (cluster 4) share the synonym: amusing. As for songs from clusters 1 – 5, there are some acoustic similarities. Both are energetic, loud and many of both use electric guitar. [23]

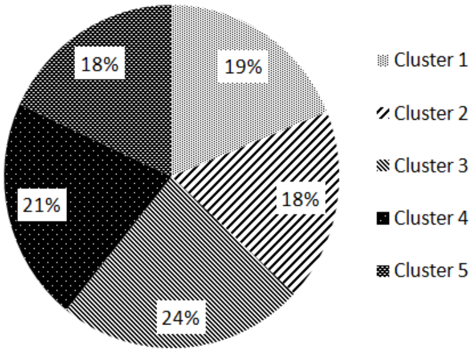

Figure 1: *Audio clips distribution between the five clusters.*

This dataset is available at our website[6] to any researchers willing to use it in future research.

4.2. Mood Classification

The classification tests were accomplished with 20 repetitions of 10-fold cross validation, guaranteeing that all songs are used in different groups for training and testing. Some tests were also repeated using 3 and 5-fold cross validation due to the fact that MIREX uses a 3 folds setup. However most of the analysis was done with 10-fold since, according to the literature, "there are more performance estimates, and the training set size is closer to the full data size, thus increasing the possibility that any conclusion made about the learning algorithm(s) under test will generalize to the case where all the data is used to train the learning model" and using fewer folds did not influenced the results noticeably. [24]

The various tests run gave some valuable insights about the importance of each framework and its features on mood detection. Based on them, the best results were obtained with a subset of the combination of all the feature sets from the three frameworks, selected by the feature selection algorithm ReliefF. Although the best results were obtained with a high number of features, similar results of 47.2% are observed using 39 features, selected by ReliefF, while only 19 features are sufficient to obtain 95% of that. The best results in terms of F-measure were 47.21%, with a precision of 46.86% and recall of 47.60%.

These tests also showed MIR Toolbox features as achieving better results, with Marsyas close behind. While PsySound was the third, it is important to note that it used a really small number of features when compared to the other two frameworks. Another interesting result was that the two Marsyas feature sets tested achieved very similar results, although one had much less features. Furthermore, using feature selection on the bigger set did not improve the results and the subset of features obtained from the algorithm actually performed worse than any of the initial two feature sets.

A brief summary of these results is presented in Table 2. All the tests where PCA was used, in order to reduce correlation between variables, did not achieve any noticeable improvement in results but actually resulted in lower values. The feature set referred as "Combined" represents the combination of features from the three frameworks used. The "CombinedRFF" represents a subset of the "Combined" dataset, selected with feature selection algorithm ReliefF.

Table 2: *Results obtained for each feature set.*

Feature Set	F-measure	Precision	Recall
Marsyas (65)	41.52%	40.71%	42.37%
MIR toolbox (177)	44.43%	44.05%	44.81%
PsySound 3 (11)	36.37%	35.64%	37.13%
Combined	47.21%	46.86%	47.60%
CombinedRFF	47.20%	46.82%	47.57%

A list of the top twelve features used to accomplish the best results is presented on Table 3.

Table 3: *The top 12 features obtained with ReliefF algorithm.*

Feature	Framework	Weight
Key strength – major (max)	MIR Toolbox	0.0236
mode	MIR Toolbox	0.0203
Key clarity	MIR Toolbox	0.0183
Key strength – major (std)	MIR Toolbox	0.0175
Tempo (std)	MIR Toolbox	0.0175
Key strength – minor (std)	MIR Toolbox	0.0166
High-freq energy (avg)	MIR Toolbox	0.0149
Tonal Dissonance (S)	PsySound3	0.0148
Zero crossing rate (skw)	MIR Toolbox	0.0146
Tonal centroid 3rd (std)	MIR Toolbox	0.0136
MFCC (3rd)	Marsyas	0.0125
Rolloff (avg)	Marsyas	0.0117

The list was obtained by running the feature selection algorithm on the combined feature set of all frameworks. It is important to notice that the best results were obtained with a big number of features, using only the listed features according to ReliefF results in lower results. Some features perform generally bad but will be important when combined with others. To obtain a better list of features and exploit the relation between features, some other feature selection algorithm could also be tested, combining the results. It could be also relevant to reduce the number of available features, dropping some of the less interesting features that may be introducing noise or duplicating information, resulting in worse ReliefF results.

Finally, the confusion matrix resulting from the best model, using the Combined feature set, is listed in the table 4.

Table 4: *Confusion matrix (results are in %).*

		Predicted				
		C1	**C2**	**C3**	**C4**	**C5**
Annotated	**C1**	42.36	18.27	8.01	16.66	21.84
	C2	19.79	44.03	14.63	17.33	6.80
	C3	4.08	12.43	54.62	13.98	7.39
	C4	14.13	21.01	17.20	38.99	10.97
	C5	19.64	4.26	5.54	13.04	53.00

In table 4, a considerable percentage of songs were wrongly classified between clusters 1 – 5 and 2 – 4. This seems to go in the direction of the previously identified semantic and acoustic overlap between these same clusters. In table 5 a confusion matrix grouping the overlapped clusters is presented.

Table 5: *Confusion matrix merging the clusters with semantic and acoustic overlap (results are in %).*

		Predicted		
		C1+5	**C2+4**	**C3**
Ann.	**C1+5**	68.82	27.46	13.55
	C2+4	25.34	59.05	31.83
	C3	5.84	13.50	54.62

Finally, table 6 contains the results – f-measure, precision and recall, for each cluster using the best feature set.

[6] http://mir.dei.uc.pt/

Table 6: *Results obtained for each cluster using the best performing feature set.*

Cluster	F-measure	Precision	Recall
C1	37.5%	40.0%	35.3%
C2	37.0%	50.0%	29.4%
C3	61.2%	53.6%	71.4%
C4	40.0%	38.1%	42.1%
C5	52.9%	52.9%	52.9%

5. CONCLUSIONS

In this paper, we have tested an approach for emotion classification in audio music. To this end, a dataset similar to the dataset used in the MIREX Mood Classification task was collected using the Allmusic database. However, the similarity between both sets is still to be properly accessed by using the same classification system in both. It is also known that the dataset used in MIREX suffers from some previous identified problems still to be fixed such as clusters overlap and lack of balance between clusters in each of the three folds [22]. The results available from previous editions also show some disparity of results between folds, possibly due to ambiguity existent in some clips, since songs were separated in folds based on the degree of agreement between the panel of experts [22].

Regarding our system accuracy, results of 47.2% in terms of F-measure were obtained, with a precision of 46.9% and recall of 47.6%. These results are lower than the best obtained in the last edition of MIREX, with 69% precision. Still, it is hard, to draw definitive conclusions only from these results, since two different datasets were used. As a possible point of comparison, we have conducted tests in our dataset extracting the same feature set that was extracted and used in MIREX 2008 and 2010 using Marsyas. This set of features achieved only 40.71% precision in our dataset, while obtaining 48.58% to 57.5% precision in 2010. This fact might suggest that our approach would have better precision in MIREX dataset, possibly due to the higher quality of its annotations, which were created by a panel of three experts, while Allmusic annotations were created by users' contribution.

It was also verified that the tested feature selection algorithms do not return the optimal feature combination but many times only follow a trend, resulting in a subset of features that may not be the optimal combination. Due to this fact, a better feature selection process is needed to point the best ones to the problem, possibly using also a pre-selection by manually eliminating some in order to reduce the initial number of features into something smaller and easier to work with.

Finally, the most important factor that may improve MER results overall is probably related with the creation of novel audio features that better represent emotions. Nowadays, most of the features being used were developed long ago for other problems such as speech recognition. The development of new, high level features specifically created to emotion recognition problems is a problem yet to be explored in the years to come.

6. ACKNOWLEDGMENTS

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e a Tecnologia (FCT) and Programa Operacional Temático Factores de Competitividade (COMPETE) – Portugal.

7. REFERENCES

[1] D. Huron, "Perceptual and Cognitive Applications in Music Information Retrieval", International Symposium on Music Information Retrieval, 2000.

[2] A. Friberg, "Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music", in Proc. 11th Int. Conf. on Digital Audio Effects, 2008, pp. 1-6.

[3] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition", IEEE Trans. on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 448-457, Feb. 2008.

[4] O. C. Meyers, "A mood-based music classification and exploration system", MSc thesis, Massachusetts Institute of Technology, 2007.

[5] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", Machine Learning, vol. 53, no. 1-2, pp. 23-69, 2003.

[6] R. Panda and R. P. Paiva, "Automatic Creation of Mood Playlists in the Thayer Plane: A Methodology and a Comparative Study," in 8th Sound and Music Computing Conference, 2011.

[7] M. D. Korhonen, D. a Clausi, and M. E. Jernigan, "Modeling Emotional Content of Music Using System Identification," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 36, no. 3, pp. 588-599, Jun. 2006.

[8] A. Gabrielsson and E. Lindström, "The Influence of Musical Structure on Emotional Expression", in Music and Emotion, vol. 8, Oxford University Press, 2001, pp. 223-248.

[9] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals", IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 5-18, Jan. 2006.

[10] Y. Feng, Y. Zhuang, and Y. Pan, "Popular Music Retrieval by Detecting Mood", Proc. 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, vol. 2, no. 2, p. 375, 2003.

[11] D. Yang and W. Lee, "Disambiguating Music Emotion Using Software Agents", in Proc. 5th Int. Conf. on Music Information Retrieval, 2004, p. 52–58.

[12] D. Liu and L. Lu, "Automatic Mood Detection from Acoustic Music Data", Int. J. on the Biology of Stress, vol. 8, no. 6, pp. 359-377, 2003.

[13] J. Wang, H. Lo, and S. Jeng, "Mirex 2010: Audio Classification Using Semantic Transformation And Classifier Ensemble," in Proc. of The 6th International WOCMAT & New Media Conference (WOCMAT 2010), 2010, pp. 2-5.

[14] E. Schmidt and Y. Kim, "Modeling Musical Emotion Dynamics with Conditional Random Fields," in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011, pp. 777-782.

[15] M. McVicar and T. Freeman, "Mining the Correlation Between Lyrical and Audio Features and the Emergence of Mood," in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011, pp. 783-788.

[16] Z. Cataltepe, Y. Tsuchihashi, and H. Katayose, "Music Genre Classification Using MIDI and Audio Features", EURASIP Journal on Advances in Signal Processing, vol. 2007, no. 1, pp. 275-279, 2007.

[17] O. Lartillot and P. Toiviainen, “A Matlab Toolbox for Musical Feature Extraction from Audio”, in Proc. 10th Int. Conf. on Digital Audio Effects, 2007, p. 237–244.

[18] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, “Temporal Feature Integration for Music Genre Classification”, IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 5, pp. 275-9, 2007.

[19] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines”, Computer. pp. 1-30, 2001.

[20] C. M. Bishop, “Neural Networks for Pattern Recognition”, Journal of the American Statistical Association, vol. 92, no. 440, p. 1642, 1995.

[21] I. Nabney and C. Bishop, “Netlab Neural Network Software”, Pattern Recognition. 1997.

[22] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, “The 2007 MIREX Audio Mood Classification Task: Lessons Learned,” in Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2011), 2008, pp. 462-467.

[23] C. Laurier, “Audio music mood classification using support vector machine,” in MIREX Audio Mood Classification Task, 2007, pp. 2-4.

[24] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” in Encyclopedia of Database Systems, 2nd ed., vol. 25, M. Tamer and L. Liu, Eds. Springer, 2009, pp. 1-12.