Music emotion recognition using two level classification

Conference Paper · February 2014

DOI: 10.1109/IranianCIS.2014.6802519

2 authors:

Samira Pouyanfar
Florida International University
32 PUBLICATIONS **672** CITATIONS
SEE PROFILE

Hossein Sameti
Sharif University of Technology
133 PUBLICATIONS **963** CITATIONS
SEE PROFILE

Music Emotion Recognition Using Two Level Classification

Samira Pouyanfar, Hossein Sameti
Department of Computer Engineering
Sharif University of Technology,
Tehran, Iran
pouyanfar@ce.sharif.edu, sameti@sharif.edu

***Abstract*—Rapid growth of digital music data in the Internet during the recent years has led to increase of user demands for search based on different types of meta data. One kind of meta data that we focused in this paper is the emotion or mood of music. Music emotion recognition is a prevalent research topic today. We collected a database including 280 pieces of popular music with four basic emotions of Thayer's two Dimensional model. We used a two level classifier the process of which could be briefly summarized in three steps: 1) Extracting most suitable features from pieces of music in the database to describe each music song; 2) Applying feature selection approaches to decrease correlations between features; 3) Using SVM classifier in two level to train these features. Finally we increased accuracy rate from 72.14% with simple SVM to 87.27% with our hierarchical classifier.**

Keywords- Music emotion recognition; Feature extraction; Two level classiffication; Feature selection; Music information retrieval

I. INTRODUCTION

Music is a powerful tool and has many advantageous effects on mankind's body and soul, it can excite and calm. The area in human brain which perceives music is near to area where emotional expressions are learnt by man and it is why there is a direct relationship between music and emotional expressions. Two groups of researchers are studying the relation between music and emotion:

1. Music psychologists [1, 2] who study the relation between acoustic cues (such as beat, tempo, sound level, etc.) and various expressed emotions (such as angry, sad, happy, etc.). They mostly develop emotional models.
2. Computer researchers [3, 4, 5, 6, 7, 8, 9] who develop algorithms to detect music emotion automatically. They are trying to utilize emotion in addition to conventional meta data such as genre and title, for music retrieval.

Music emotion recognition methods can be divided into two general classes: categorical method [5, 6, 7, 8, 9] and dimensional method [3, 4]. In categorical method emotions are grouped into basic emotion classes but in the dimensional method, each emotion is modeled as a point on a continuous plane. In this study, categorical method which could be integrated more easily with music retrieval methods is used.

Thayer's two dimensional emotion model [1] is a successful psychological model of emotion that is used in most of emotion detection projects. This model includes two dimensions called arousal and valence. This emotion model and four basic emotion classes (happy, angry, sad, relax) are used in our study where each class is considered as a quadrant of the emotion plane.

In this paper, a two-level classification method is applied. The flow chart of the proposed algorithm is shown in Fig. 1. In the beginning, after preprocessing step, several features such as energy, rhythm and timbre are extracted from music pieces, then two algorithms are used for feature reduction; finally, a two-level framework is used to train and predict the emotion of each piece of music and SVM [10] applied as classifier in this framework. Then we compared performance of our two level frameworks with a simple SVM classifier. Also Different classifiers such as K nearest Neighbor (KNN) [11], Gaussian Mixture Model (GMM) [12], neural network (RBF) and Bagging are employed to evaluate the efficiency of proposed algorithm. The remaining structure of this paper is as follow. 2D Thayer emotion model is introduced in the next section, recognition steps are explained in Section 3, experimental results are presented in Section 4 and finally in Section 5, conclusion and future works are given.

II. EMOTION MODEL

Till now, a few emotion models have been suggested in psychology and physiological sciences [1, 2]. One of the models prevalently used in music emotion is the Thayer model because it is quit relevant to music features. Thayer'2D model is based on two basic and effective parameters, music energy and music pleasance, these parameters are also known as arousal and valence, respectively. People's heart rate beats faster and their blood pressure rises when they are listening to angry or happy songs (music with high energy). These parameters are related to the arousal dimension. Instead, increased Cortisol levels in the blood related to positive valence.

This model consists of a two dimensional plane divided according to the arousal and valence parameters in to four clusters each placed in one quadrant of the plane. The resultant four clusters are angry, happy, relax and sad as shown in Fig. 2; we used this emotion model in our proposed method.

978-1-4799-3351-8/14/$31.00 ©2014 IEEE

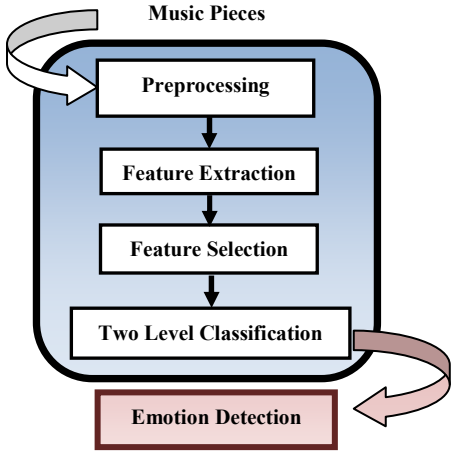


Figure 1. Flowchart of proposed scheme

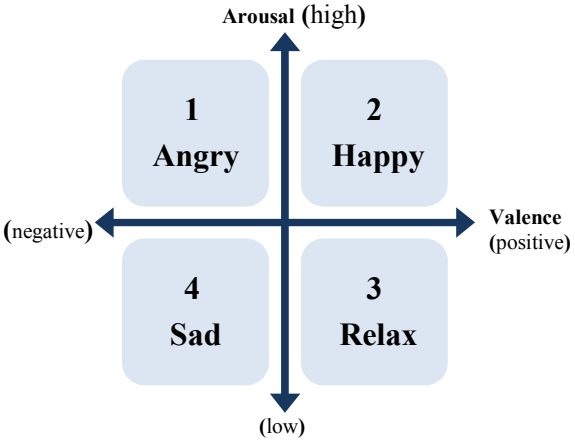


Figure 2. Thayer's two dimensional emotion model

III. MUSIC EMOTION RECOGNITION PROCESS

Steps of the proposed algorithm have been shown in Fig. 1 and explained in the following:

A. Data Collection

There is no universal standard database for music motion recognition, so different databases have been used in reported works [3, 5, 6].That is why there are differences in reported accuracies of each algorithm. All Music Guide (AMG)[a] is a music company which uses moods for music retrieval, this company uses 288 mood labels for emotional labeling, and these labels are assigned by music specialists. To build our database, we used 280 songs from AMG including four basic emotion labels. Each track of music contains a range of different emotions, so we used a period 30 seconds of each track to normalize emotional variation; a 30s music clip, could be informative enough to retrieve the mood.

We performed clip normalization to deal with the amplitude differences of the music tracks in the database (i.e., some tracks have a higher volume than other). For this purpose, Cool edit pro[b] is utilized as an audio editing tool.

[a] http://www.allmusic.com/.
[b] http://www.adobe.com/products/audition/.

TABLE I. EXTRACTED FEATURES FOR THE PROPOSED ALGORITHM

Group	Toolboxes	Features	Count
Energy	SDT	Audio power, total loudness, and Specific loudness sensation coefficients, Sharpness	88
Rhythm	MIRtoolbox	Rhythm strength, rhythm regularity, rhythm clarity , average onset frequency, and average tempo	5
Other features (Timbre, Spectrum and Harmony)	PsySound, MIRtoolbox, MATLAB	Roughness, irregularity, inharmonicity, Salient pitch, chromagram centroid, key clarity, musical mode, harmonic change Spectral centroid, spectral roll off, spectral flux, spectral flatness measures, and Mel-frequency cepstral coefficients Zero-crossings, temporal centroid, and log attack time , …	83
Total			**176**

B. Music Features

After data preprocessing, appropriate features are extracted from music data. Regarding Thayer model, two feature sets are needed: energy or intensity feature set is related to the arousal dimension and the rhythm feature set is more related to the valence dimension. For feature extraction, we also employed toolboxes such as MIRtoolbox[c] [13], Sound Description Toolbox (SDT)[d] [14], and PsySound[e] [15]. Table I shows the features in more detail.

1) Energy Features

One of the most significant features in music emotion detection is energy, because it's quite related to the arousal dimension. High energy music locates on top half of the plane and vice versa. The Energy of a signal x in a window of N sample is calculated by (1).

$$E(n) = \sum_1^N x(n).x*(n) \quad (1)$$

Other energy related features are loudness and sharpness. We measured perceived loudness by Chaluper and Fastl model [16]. This model uses the Bark critical band [17] for modeling auditory filters. Auditory temporal integration is included in this model. Also sharpness, a subjective rate of sound on a scale from dull to sharp, is measured by Zwicker and Fastl's model [18]. To extract this feature set we used Sound

[c] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/.
[d] http://www.ifs.tuwien.ac.at/mir/muscle/del/audio_tools.html#SoundDescrToolbox/.
[e] http://psysound.wikidot.com/.

Description Toolbox and PsySound. SDT is a MATLAB open source code that has the capability of extracting 187 features from audio signals where 40 of them are energy related. Psysound is a MATLAB open source software with a graphical user interface that analyzes the sounds using psychoacoustical algorithms [15]. Table I shows details of energy feature set.

Fig. 3 shows loudness of four different songs. It can be observed that happy and angry songs (high arousal songs located at the top of Thayer emotion plane), have higher total loudness than relax and sad songs (low arousal songs at the bottom of the plane). In addition to the total loudness, as it can be figured out from Fig. 3, amplitude variation range of loudness diagram in the first group is less than second group. Sharpness diagrams of these songs have this characteristic, either. So, we measured standard deviation of loudness and sharpness as new features.

2) *Rhythm Features*

Another significant feature in music application is rhythm which is related to both arousal and valence dimensions [6]. The systematic pattern of musical sounds, mainly according to duration and periodic stress is called rhythm.

Onset sequence or event density [6, 19], detection of the successive notes of each music clip, is extracted by MIRtoolbox. Fig. 4 shows onset curves of two example songs with different emotions. It can be observed that happy song has much firmer and stronger rhythm than sad song. Happy song rhythm is more regular, too. Therefore, two types of features from onset curve can be extracted: average onset strength and rhythm regularity [6]. Firstly, onset sequence is obtained by MIRtoolbox, and then onset strength mean indicating the rhythm strength is computed. Autocorrelation is applied on the onset curve to extract rhythm regularity; the more regular the rhythm pattern is the stronger peaks appear on the autocorrelation curve. Using this curve, other derivative features could be calculated, like mean of autocorrelation curve peaks and peaks strength to valley's strength ratio.

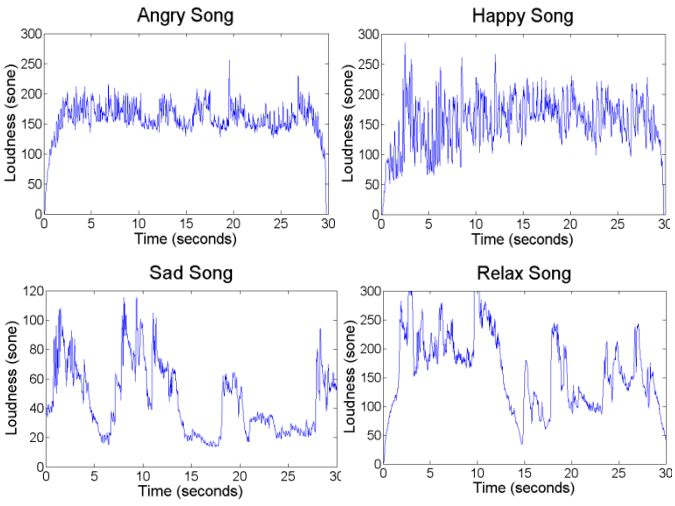
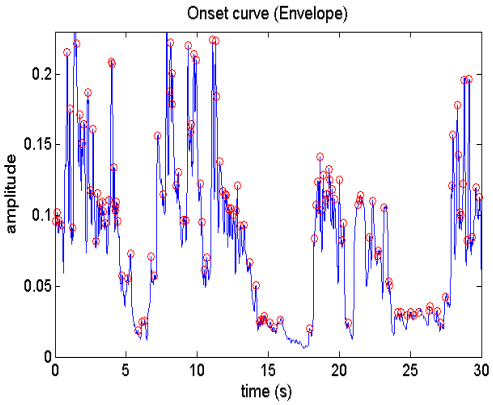
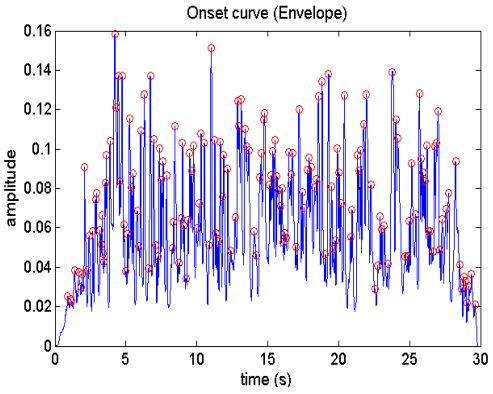


Figure 3. The loudness diagrams of the four example songs



a) *Sad song*



b) *Happy Song*

Figure 4. Onset curves of two songs with different emotions

Tempo (detection of periodicity from onsets per second), and rhythm clarity (autocorrelation of onset detection curve) [19] are also extracted by MIRtoolbox.

3) *Timbre Features*

Timbre features are usually used in speech recognition systems [20, 21]. Preprocessing of sound signals is needed to extract this feature set. First, the signals are divided into frames by applying a window function to remove the edge effects. Hamming windows is a good choice for this aim.

- Mel-Frequency Cepstral Coefficients (MFCC) representing the formant peaks of the spectrum are very popular features in speech processing. First, log power spectrum based on short-term Fourier transform is computed for each frame. To divide the frequencies, Mel- frequency scale is used. Then, discrete cosine transform (DCT) is applied to decorrelate the Mel vectors. Typically, the mean and standard deviation of MFCCs are taken.
- Spectral Centroid, the centroid of the amplitude spectrum of short-term Fourier transform (STFT), measures the spectral brightness.
- Spectral Rolloff, the frequency below where a certain fraction of the amplitude distribution is contained and

this ratio is 0.85 by default. Spectral rolloff measures the spectral shape of signal.

- Zero Crossing rate is the number of signal values crossing the zero axis in the time domain. It is a measure of signal noisiness.
- MIRtoolbox and SDT were utilized to extract other features such as temporal, spectrum and harmony; they are listed in Table I.

C. Feature Selection

For identifying the best subset of features in the data set and removing redundant and irrelevant features, some feature selection methods are applied. For this purpose, FilterSubsetEval algorithm of selection attribute panel of Weka[f] and the sequential floating forward selection algorithm (SFFS) of Feature Selection DEMO [22] are used. FilterSubsetEval algorithm passes data through an arbitrary filter before running an arbitrary subset evaluator. By default, we employ SpreadSubsample as the filter and CfsSubsetEval as the subset evaluator; also, GreedyStepwise search is our preferred method. Detailed explanation of each method described as following:

- **CfsSubsetEval**: Evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them.
- **SpreadSubsample**: Produces a random subsample of a dataset.
- **GreedyStepwise**: performs a greedy forward through the space of attribute subset.

SFFS algorithm finds an optimum subset of features by inserting a new feature to the previous subset of features and deleting a feature from new subset of features. By applying these feature reduction methods, final result is more reliable and accuracy is improved.

D. Two Level Classification

Next step of recognition process is training a machine learning model to learn the relation between emotion and music. We proposed a two level classification algorithm based on Thayer emotion model as illustrated in Fig. 5. Based on experimental results, high energy classes such as angry and happy are easy to classify but low energy classes (relax and sad) are very likely to be wrongly classified. So we tried to classify our dataset in two levels. Before training process, we divided our dataset into two categories. One includes all data with three classes (angry, happy and Relax-Sad) and other one includes half of data which are labeled with relax or sad.

In the first level of this procedure, we trained a SVM to classify first category of datasets. In this level, we used a feature subset extracted from feature reduction methods described in the previous section.

[f] WEKA Data Mining software in java. http://www.cs.waikato.ac.nz/ml/weka/.

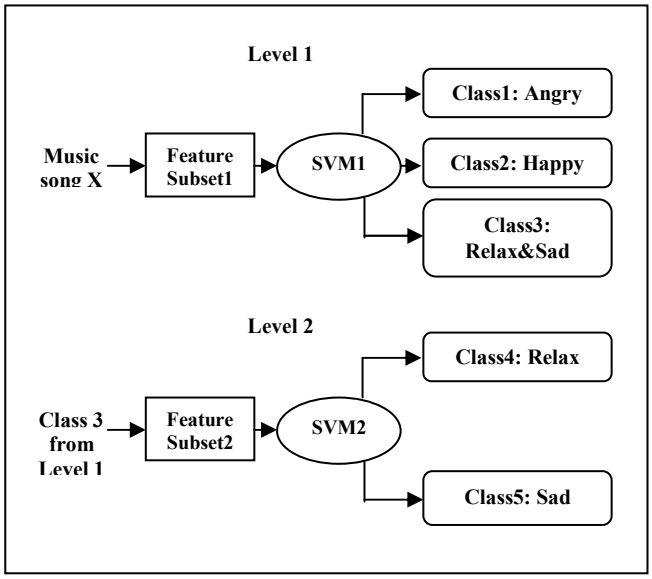


Figure 5. Two level classification framework

The first level feature subset includes the best features such as Energy, one of the most significant features in this subset, to separate three classes. In the second level of our methods, we trained another SVM to classify the second dataset. In this level we again applied feature selection methods to our second dataset to find the best features for separating relax and sad classes.

After training phase, each unlabeled music signal with its first feature subset, is predicted by SVM 1. If its predicted class is 3, it will enter to second level to be classified by SVM 2 with the second feature subset and finally its exact label gets specified. One of our difficulties in this project was the absence of enough music data and sparsity of training data. Hence, appropriate classifiers and methods to overcome this deficiency are needed. This algorithm can emphasize on different feature subsets in different classification tasks and it is appropriate for sparse dataset.

IV. EXPERIMENTAL RESULTS

We used accuracy to evaluate the validity of our algorithm. As mentioned in Section 3, in this project we used 280 music songs with 4 classes of emotion from AMG site and we chose 30 seconds of each one. Then each data was converted to a standard format (22.050 sampling frequency rates, 16 bit per sample and mono channel). Feature extraction led to 176 features for each song; then, linear normalization was applied to each component. This normalization, also known as standardization or z-scores is depicted as the (2):

$$x^*(i) = (x(i) - \mu(i))/\sigma(i) \qquad (2)$$

Where x(i) is ith feature component with mean μ(i) and variance σ(i) within the training data. Due to lack of enough music data, we applied 10 fold cross validation method. FilterSubsetEval and SFFS methods were employed to select the best features for each level. Details of features included in

each feature set are shown in Table II. Finally, we trained each SVM of our two level algorithm separately with their datasets. RBF was the best kernel for this experiment. An optimized SVM algorithm was applied that chooses the best SVM parameter such as C or mu.

To evaluate our proposed method, we also employed other classifiers such as simple SVM, GMM, RBF, KNN, Bagging on data as illustrated in detail in Table III. Highest accuracy was 87.27% that was achieved by proposed method.

Confusion matrices of simple SVM (with feature selection) and proposed approach are shown in Table IV and Table V respectively. Based on these tables, one can conclude that the accuracy of happy and angry is very high (about 90%-98%) for both methods but decreasing of the final accuracy to 82.5% in simple SVM is mostly because of the similarity between sad and relax classes. Therefore, we tried to increase the accuracy of these two classes by adding another level to classify them separately, and applied different feature set for this level. The final result indicates the efficiency of the proposed algorithm increasing accuracy to 87.27%.

V. CONCLUSION AND FUTURE WORKS

In this paper, a new algorithm is proposed and evaluated for automatic emotion recognition based on Thayer's two dimensional model. Various basic classifier methods such as simple SVM, GMM, Neural Network, Bagging and KNN and our two level classifier based on SVM are applied to music data and accuracy improved from 72.14% by SVM to 87.27% by proposed method.

TABLE II. COMPARISON BETWEEN FEATURE SUBSETS OF TWO LEVEL CLASSIFICATION ALGORITHM

Feature sets	**Feature groups**			
	Energy	***Rhythm***	***Spectral***	***total***
Feature subset1	Audio power, total loudness, Specific loudness sensation coefficients , Sharpness (27)	Rhythm strength (1)	chromagram centroid, harmonic change Spectral centroid, spectral roll off, spectral flux, spectral flatness measures, and Mel-frequency cepstral coefficients Zero-crossings, temporal centroid, and log attack time , …(30)	58
Feature subset2	-	Rhythm strength, rhythm regularity, rhythm clarity , average onset frequency, and average tempo (5)	Roughness, irregularity, inharmonicity, Salient pitch, key clarity, musical mode, spectral roll off, spectral flux, spectral flatness, Mel-frequency cepstral coefficients Zero-crossings, …(31)	36

TABLE III. COMPARISON BETWEEN DIFFERENT CLASSIFIER ACCURACY

Classifier	***Accuracy (%)***	***Accuracy +FS[a] (%)***	***Toolbox***	***Details***
SVM	72.14	82.5	C#,LIBSVM	Kernel:RBF, C=0.6
GMM	72	82.14	MATLAB	EM algorithm # mixtures=4
RBF	75.35	77.14	WEKA	#cluster of kmeans=3
KNN	75	79.64	WEKA	K=5
Bagging	77.85	80.35	WEKA	Classifier: NaiveBayes
Proposed method (Two level SVM)		**87.27**	C#,LIBSVM	-

a. FEATURE SELECTION

TABLE IV. CONFUSION MATRIX OF SIMPLE SVM (100%)

	Angry	***Happy***	***Relax***	***Sad***
Angry	**95.71**	4.28	0	0
Happy	5.71	**90**	4.28	0
Relax	0	5.71	**70**	24.28
Sad	1.42	0	24.28	**74.28**

TABLE V. CONFUSION MATRICES OF TWO LEVEL CLASSIFICATION (100%)

	Angry	***Happy***	***Relax-Sad***
Angry	**98.57**	1.43	0
Happy	7.14	**91.42**	1.43
Relax-Sad	0	1.43	**98.57**

	Relax	***Sad***
Relax	**80**	20
Sad	18.58	**81.42**

There are deficiencies in current works on music emotion recognition such as lack of standard database including music with emotion tags, limited number of emotion classes in comparison to the human rich perceived emotions of music, lack of effective attribute selection algorithms and not enough existing powerful machine learning methods for music emotion classification. In the following, some solutions for future works are recommended:

- Provide a standard and acceptable database including all types of music (jazz, pop, classic, etc.) with various cultures and labeling by music specialists.
- Extract more efficient features for music, for example vocal timbre recommended in [19].
- Combine categorical and dimensional methods to achieve higher performance.

REFERENCES

[1] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford Univ. Press, New York, 1989.

[2] P.N. Juslin, and J.A. Sloboda, *Music and Emotion: Theory and research*, Oxford Univ. Press,2001.

[3] B. Han, S. Rho, R. Dannenberg, and E. Hwang, “SMERS: Music emotion recognition using support vector regression,” in *Proc. of 10th International Conference on Music Information Retrieval ISMIR, 2009*.

[4] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, “A regression approach to music emotion recognition,” IEEE Transactions on Audio, Speech Language Processing, 16(2):448-457, 2008.

[5] Y. Song, S. Dixon, and M. Pearce, “Evaluation of Musical Features for Emotion Classification,” in *Proc. of 13th International Conference on Music Information Retrieval ISMIR*, *2012*.

[6] D. Lu, L. Liu, and H. Zhang, “Automatic mood detection and tracking of music audio signals,” IEEE Transactions on Audio, Speech Language Processing, 14(1):5-18, 2006.

[7] C.-H. Yeh, H.-H. Lin, and H.-T. Chang, “An effcient emotion detection scheme for popular music,” in *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS),* pages 1799-1802, 2009.

[8] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahava, “Multilabel classification of music into emotions,” in *Proc. of 9th International Conference on Music Information Retrieval ISMIR, 2008*.

[9] T. Li and M. Ogihara, “Content-based music similarity search and emotion detection,” in *Proc. of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* volume 5, pages 705-708, 2004.

[10] C. Cortes and V. Vapnik, “Support-vector network,” Machine Learning, 20:273-297, 1995.

[11] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, 13(1):21–27, 1967.

[12] A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” journal of the Royal Statistical Society 39(1) (1977) 1–38.

[13] O. Lartillot and P. Toiviainen, “A matlab toolbox for musical feature extraction from audio,” in *Proc. of the 10th International Conference on Digital Audio Effects, Bordeaux*, pp. 237-244, 2007.

[14] E. Benetos, M. Kotti, and C. Kotropoulos, “Large scale musical instrument identification,” in *Proc. of International Conference on Music Information Retrieval, 2007.*

[15] D. Cabrera, “Psysound: A computer program for the psychoacoustical analysis of music,” in *Proc. of the Australian Acoustical Society Conference, pages 47-54, 1999*.

[16] J. Chalupper and H. Fastl, “Dynamic loudness model (dlm) for normal and hearing-impaired listeners,” Acta Acustica United with Acustica, 88:378-386, 2002.

[17] E. Zwicker, “Subdivision of the audible frequency range into critical bands (frequenzgrup-pen) ,” journal of Acoustical Society of America, 33, 1961.

[18] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, New York, 1999.

[19] Y.-H. Yang and H.-H. Homer. *Music Emotion Recognition*, CRC Press (Taylor and Francis), Feb 2011.

[20] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*, Springer-Verlag, 1998.

[21] T. Li and M. Ogihara, “Toward intelligent music information retrieval,” IEEE Transactions on Multimedia, 8(3):564-574, 2006.

[22] C. Kotropoulos, and D. Ververidis, “Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition,” Elsevier Signal Processing, 88(12):2956-2970, 2008.