Report of the ISMIS 2011 Contest: Music Information Retrieval

Bozena Kostek[1], Adam Kupryjanow[1], Pawel Zwan[1], Wenxin Jiang[2],
Zbigniew W. Raś[3,4], Marcin Wojnarski[5] and Joanna Swietlicka[5]

[1] Multimedia Systems Department, Gdansk University of Technology,
Narutowicza 11/12, 80-233 Gdansk, PL
{bozenka, adamq, zwan}@sound.eti.pg.gda.pl

[2] Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
wjiang2@fhcrc.org

[3] Univ. of North Carolina, Dept. of Computer Science, Charlotte, NC 28223, USA

[4] Warsaw Univ. of Technology, Institute of Comp. Science, 00-665 Warsaw, Poland
ras@uncc.edu

[5] TunedIT Solutions, Zwirki i Wigury 93/3049, 02-089 Warszawa, Poland
{marcin.wojnarski,j.swietlicka}@tunedit.org

Abstract. This report presents an overview of the data mining contest organized in conjunction with the 19th International Symposium on Methodologies for Intelligent Systems (ISMIS 2011), in days between Jan 10 and Mar 21, 2011, on TunedIT competition platform. The contest consisted of two independent tasks, both related to music information retrieval: recognition of music genres and recognition of instruments, for a given music sample represented by a number of pre-extracted features. In this report, we describe aim of the contest, tasks formulation, procedures of data generation and parametrization, as well as final results of the competition.

Keywords: Automatic genre classification, instrument recognition, music parametrization, query systems, intelligent decision systems.

1 Introduction

Internet services expose nowadays vast amounts of multimedia data for exchange and browsing, the most notable example being YouTube. These digital databases cannot be easily searched through, because automatic understanding and indexing of multimedia content is still too difficult for computers – take, for example, the diversity of musical trends and genres, uncommon instruments or the variety of performers and their compositions present in typical multimedia databases. In ISMIS 2011 Contest, we invited all researchers and students interested in sound recognition and related areas (signal processing, data mining, machine learning) to design algorithms for two important and challenging problems of Music Information Retrieval: recognition of music genres (jazz, rock, pop, ...) and recognition of instruments playing together in a given music sample.

The contest comprised 2 tracks:

- Music Genres – Automatic recognition of music genre (jazz, rock, pop, ...) from a short sample. The dataset and feature vectors were prepared by the Multimedia Systems Department of the Gdansk University of Technology.
- Music Instruments – Automatic recognition of instruments playing together in a given sample. Prepared by Wenxin Jiang and Zbigniew Raś.

The tasks were independent. Contestants could have participated in both of them or in a selected one. The challenge was organized at TunedIT Challenges[6] platform as an on-line *interactive* competition: participants were submitting solutions many times, for the whole duration of the challenge; solutions were immediately automatically evaluated and results were published on the Leaderboard, to allow comparison with other participants and introduction of further improvements. TunedIT Challenges is an open platform that can be freely used by everyone for scientific and didactic purposes [23].

Contestants were given descriptions of both tasks along with vectors of parameters – features extracted from raw sound samples. For each track, the full set of vectors was divided into two subsets: (1) *training set*, disclosed publicly to participants for classifier building; (2) *test set*, disclosed without decisions, which had to be predicted by contestants – the predictions were subsequently compared with ground truth kept on the server and scores were calculated. Test set was further divided into preliminary (35%) and final (65%) parts: the former being used for calculating results shown on the Leaderboard during contest; the latter one used for final unbiased scoring and picking up the winner.

The competition attracted very large interest among Data Mining and Music Information Retrieval community: 292 teams with 357 members had registered, 150 of them actively participated, submitting over 12.000 solutions in total, largely outperforming baseline methods. The winners are:

- Music Genres: **Amanda C. Schierz and Marcin Budka**, Bournemouth University, UK. Achieved 59% lower error rate than baseline algorithm.
- Music Instruments: **Eleftherios Spyromitros Xioufis**, Aristotle University of Thessaloniki, Greece. Achieved 63% lower error than baseline.

The winning teams were awarded prizes of 1000 USD each. See also the contest web page: `http://tunedit.org/challenge/music-retrieval`.

2 Task 1: Music Genres Recognition

2.1 Database and Parametrization Methods

The automatic recognition of music genres is described by many researchers [1, 3, 4, 6, 7, 11, 15–18, 20, 21], who point out that the key issue of this process is a proper parametrization. Parametrization has so far experienced extensive development [5, 8–12, 24], however, there are still some important areas of Music Information Retrieval, such as for example music genre classification, that is

[6] http://tunedit.org

researching this aspect. The parameters that are most often used are MPEG-7 descriptors [6], mel cepstral coefficients [19] and bit histograms [18]. The classification is typically based on the analysis of a short (3–10 sec.) excerpts of a musical piece which are parameterized and later classified.

A database of 60 music musicians/performers was prepared for the competition. The material is divided into six categories: classical music (class No. 1), jazz (class No. 2), blues (class No. 3), rock (class No. 4), heavy metal (class No. 5) and pop (class No. 6). For each of the performers 15–20 music pieces were collected. The total number of music pieces available for each of music genres is presented in Tab. 1. An effort was put to select representative music pieces for each of music genres.

All music pieces are partitioned into 20 segments and parameterized. The descriptors used in parametrization also those formulated within the MPEG-7 standard, are only listed here since they have already been thoroughly reviewed and explained in many research studies.

Table 1. The number of music pieces extracted for a given genre

Genre	**Number of music pieces**
Classical music	320
Jazz music	362
Blues	216
Rock music	124
Heavy metal music	104
Pop music	184

For each of music pieces 25-second-long excerpts were extracted and parameterized. The excerpts are evenly distributed in time. Each of these excerpts was parameterized and the resulting feature vector contains 171 descriptors. 127 of the features used are MPEG-7 descriptors, 20 are MFCC and additional 24 are time-related ‘dedicated’ parameters. Those ‘dedicated’ parameters are original authors’ contribution to the parametrization methods. Since MPEG-7 features and mel-frequency cepstral-coefficients are widely presented in a rich literature related to this subject there they will only be listed. Dedicated parameters will be described in the next Section.

The parameters utilized are listed below:

a) parameter 1: Temporal Centroid,
b) parameter 2: Spectral Centroid average value,
c) parameter 3: Spectral Centroid variance,
d) parameters 4–37: Audio Spectrum Envelope (ASE) average values in 34 frequency bands,
e) parameter 38: ASE average value (averaged for all frequency bands),
f) parameters 39–72: ASE variance values in 34 frequency bands,

g) parameter 73: averaged ASE variance parameters,
h) parameters 74, 75: Audio Spectrum Centroid – average and variance values,
i) parameters 76, 77:Audio Spectrum Spread – average and variance values,
j) parameters 78–101:Spectral Flatness Measure (SFM) average values for 24 frequency bands,
k) parameter 102: SFM average value (averaged for all frequency bands),
l) parameters 103–126: Spectral Flatness Measure (SFM) variance values for 24 frequency bands,
m) parameter 127: Averaged SFM variance parameters,
n) parameters 128–147: 20 first mel-frequency cepstral coefficients (mean values),
o) parameters 168–191: Dedicated parameters in time domain based on the analysis of the distribution of the envelope in relation to the rms value.

The dedicated parameters are related to the time domain. They are based on the analysis of the distribution of sound sample values in relation to the root mean square values of the signal (rms). For this purpose three reference levels were defined: r_1, r_2, r_3 – equal to namely 1, 2, 3 rms values of the samples in the analyzed signal frame.

The first three parameters are related to the number of samples that are exceeding the levels: r_1, r_2 and r_3.

$$p_n = \frac{count(samples\_exceeding\_r_n)}{length(x(k))} \tag{1}$$

where $n = 1, 2, 3$ and $x(k)$ is the signal frame analyzed.

The initial analysis of the values of parameters p_n showed a difficulty, because the rms level in the excerpts analyzed sometimes significantly varies within the analyzed frame. In order to cope with this problem another approach was introduced. Each 5-second frame was divided into 10 smaller segments. In each of these segments parameters p_n (Eq. 1) were calculated. As a result a sequence P_n was obtained:

$$P_n = \{p_n^1, p_n^2, p_n^3, \ldots.., p_n^{10}\} \tag{2}$$

where p_n^k, $k = 1 \ldots 10$ and $n = 1, 2, 3$ as defined in Eq. 1.

In this way, 6 new features were defined on the basis of sequences P_n. New features were defined as mean (q_n) and variance (v_n) values of P_n, $n = 1, 2, 3$. The index n is related to the different reference values of r_1, r_2 and r_3.

$$q_n = \frac{\sum_{k=1}^{10} p_n^k}{10} \tag{3}$$

$$v_n = var(P_n) \tag{4}$$

In order to supplement the feature description, additional three parameters were defined. They are calculated as the ‘peak to rms’ ratio, but in 3 different ways described below:

- parameter k_1 calculated for the 5-second frame,
- parameter k_2 calculated as the mean value of the ratio calculated in 10 sub-frames,
- parameter k_3 calculated as the variance value of the ratio calculated in 10 sub-frames.

The last group of dedicated parameters is related to the observation of the rate of the threshold value crossing. This solution may be compared to the more general idea based on classical zero crossing rate parametrization (ZCR). The ZCR parametrization is widely used in many fields related to automatic recognition of sound. The extension of this approach is a definition of a threshold crossing rate value (TCR) calculated analogically as ZCR, but by counting the number of signal crossings in relation not only to zero, but also to the r_1, r_2 and r_3 values. These values (similarly as in the case of the other previously presented parameters) are defined in 3 different ways: for the entire 5-second frame and as the mean and variance values of the TCR calculated for 10 sub-frames. This gives 12 additional parameters to the feature set.

The entire set of dedicated parameters consists of 24 parameters that supplement 147 parameters calculated based on MPEG-7 and mel-cepstral parametrization.

2.2 Database Verification

Before the publication of the database for the purpose of the contest the data collected needed to be tested. The set of feature vectors was divided into two parts: training and test set, with a proportion of 50:50 in such a way that excerpts extracted from one music piece could belong either to the training or to the test set (in order to test effectiveness of the system for recognizing music pieces that were not used during the training).

Next, three classifiers, namely: SVM (Support Vector Machine), J48 tree and Random Forest were trained on the basis of these data. Those classifiers were verified on a training data. In the case of SVM classifiers the C-SVC SVM with the Radial Basis kernel function was used. Settings of the algorithm as well as parameters of the kernel were chosen experimentally. The cost parameter was set to 62.5 and gamma for the kernel to 0.5. The tolerance of the termination criterion (epsilon) was equal to 10e-3. In addition, linear kernel was tested to check if the parameters could be separated linearly. Since no classification accuracy improvement has been noticed, thus the radial basis was used. For this classifier the libSVM library [2] linked to the WEKA environment [22] was employed.

Random Forest and Decision Tree classifiers were tested in the WEKA system. The Random Forest model was created with the unlimited depth of trees and the number of trees was equal to 20. For the J48 the confidence factor used for the pruning was set to 0.25 and the minimum number of instances per leaf was equal to 2.

In Tab. 2 the average efficiency obtained for these three classifiers is presented. The best results were obtained for the SVM classifier due to its ability to the non-linear separation of classes.

Table 2. Comparison of the average efficiency for 3 classifiers (first experiment)

Classifier	**Average efficiency** [%]
Support Vector Machine	**90.87**
J48	77.40
Random Forest	84.72

The second experiment was related to different division of data: excerpts extracted from one music artist/performer could belong either to the training or to the test set. In this case smaller accuracy was expected than in the first scenario, since testing was performed on excerpts that were not earlier used in the training phase. Results of this experiment are presented in Fig. 1 and Tab. 3. The same classifier settings as earlier described were used in this experiment.

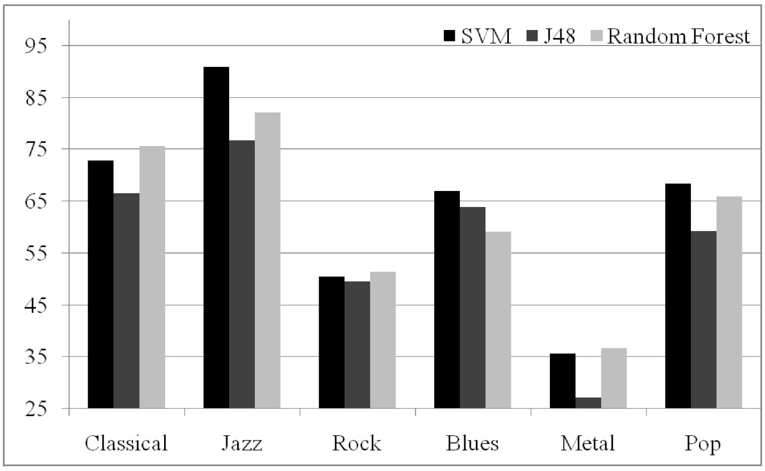

Fig. 1. Classification efficiency of music genre using three classifiers (second experiment)

Table 3. Comparison of the average efficiency for 3 classifiers (second experiment)

Classifier	**Average efficiency** [%]
Support Vector Machine	**70.0**
J48	62.1
Random Forest	67.3

Similarly to the previous experiment, the best results were obtained by the SMV classifier. As was expected the results had smaller efficiency. In addition, the confusion matrix for the SVM classifier is presented in Tab. 4.

Table 4. Confusion matrix obtained in the second experiment

[%]	Classical	Jazz	Rock	Blues	H. Metal	Pop
Classical	**72**	28	0	0	0	0
Jazz	8	**91**	0.5	0.5	0	0
Rock	0.5	21.5	**50.5**	7	14	7
Blues	3	3	1	**67**	10.5	15
H. Metal	0	2	42	6	**35**	14
Pop	0.5	2.5	11	13	5	**68.5**

The diagonal of the matrix shows the efficiency of automatic recognition for each of the classes. However the main misclassifications are additionally marked (underlined). The 'classical music' class was mainly misclassified as 'jazz' genre (28%). It is worth mentioning that any classical excerpts were classified as rock, blues, heavy metal, or pop genres. Similarly, nearly all misclassification for jazz excerpts were related to the classical music. Other genre misclassification groups are {Jazz, Rock, Heavy Metal} {Heavy Metal, Rock}, so it can be concluded that those misclassifications are related to the similarity between music genres.

The Organizers of the competition have decided to use the data divided as in the second experiment – excerpts extracted from one music performer could belong either to the training or to the test set, thus one could expect results close or better to the ones obtained in our preliminary experiment.

3 Task 2: Music Instruments Recognition

In recent years, rapid advances in digital music creation, collection and storage technology have enabled organizations to accumulate vast amounts of musical audio data, which are manually labeled with some description information, such as title, author, company, and so on. However, in most cases those labels do not have description on timbre or other perceptual properties and are insufficient for content-based searching.

Timbre is the quality of sound that distinguishes different musical instruments playing the same note with the identical pitch and loudness. Recognition and separation of sounds played by various instruments is very useful in labelling audio files with semantic information. However timbre recognition, one of the main subtasks of Music Information Retrieval, has proven to be extremely challenging especially in multi-timbre sounds, where multiple instruments are playing at the same time.

It is fairly easy to automatically recognize single instruments. However, when minimum two instruments are playing at the same time, the task becomes much more difficult. In this contest, the goal is to build a model based on data collected for both single instruments and examples of mixtures in order to recognize pairs of instruments.

3.1 Data

The data are taken from the musical sounds of MUMS (McGill Univ. Master Samples) and samples recorded in the KDD Lab at UNC Charlotte. The original audio sounds have a format of a large volume of unstructured sequential values, which are not suitable for traditional data mining algorithms. The process of feature extraction is usually performed to build a table representation from the temporal or spectral space of the signal. This will reduce the raw data into a smaller and simplified representation while preserving the important information for timbre estimation. Sets of acoustical features have been successfully developed for timbre estimation in monophonic sounds where single instruments are playing [14].

Training data consists of two datasets: large one containing data for single instruments and much smaller one containing mixtures of pairs of different instruments. 26 music instruments are involved : electric guitar, bassoon, oboe, b-flat clarinet, marimba, c trumpet, e-flat clarinet, tenor trombone, French horn, flute, viola, violin, English horn, vibraphone, accordion, electric bass, cello, tenor saxophone, b-flat trumpet, bass flute, double bass, alto flute, piano, Bach trumpet, tuba, and bass clarinet. Both sets have the same attributes (the single instruments set contains some additional information, though).

The test set contains only data for mixtures, which are the features extracted from 52 music recording pieces synthesized by Sound Forge sound editor, where each piece was played by two different music instruments. The pairs of instruments that are playing in each piece are to be predicted. Note that test and training sets contain different pairs of instruments (i.e. the pairs from the training set do not occur in the test set). Moreover, not all instruments from the training data must also occur in the test part. There also may be some instruments from the test set that only appear in the single instruments part of the training set.

The following attributes are used to represent the data:

- BandsCoef1-33, bandsCoefSum - Flatness coefficients
- MFCC1-13 - MFCC coefficients
- HamoPk1-28 - Harmonic peaks
- Prj1-33, prjmin, prjmax, prjsum, prjdis, prjstd - Spectrum projection coefficients
- SpecCentroid, specSpread, energy, log spectral centroid, log spectral spread, flux, rolloff, zerocrossing - The other
- acoustic spectral features
- LogAttackTime, temporalCentroid - Temporal features

These attributes are the acoustic features either in the frequency domain or in the time domain. Some of the features are extracted according to the MPEG-7 standard [14](such as spectral flatness coefficients and spectral centroid). Other non-MPEG7 features are also extracted such as MFCC (Mel frequency cepstral coefficients) which describes the spectrum according to the human perception system in the mel scale [13]. Additionally, the single instruments set contains:

- Frameid - Each frame is 40ms long signal
- Note - Pitch information
- Playmethod - One schema of musical instrument classification according to the way they are played
- Class1,class2 - Another schema of musical instrument classification according to Hornbostel-Sachs

Both sets contain also the actual labels, i.e. the instruments. For the mixture data, there are two instruments, thus - two labels.

3.2 Solutions and Evaluation

The goal of the contest is to select the best features provided in the training set to build the appropriate classifier that yields the high confidence of the instrument classification for the test set.

Solution should be a text file containing one pair of labels per line. The names of the instruments in each line can be given in any order and should be separated by a comma. Different number of labels in a line will result in an error. Labels are not case sensitive. The evaluation metric is modified accuracy:

- If no recognized instrument matches the actual ones, 0.0 score is assigned
- If only one instrument is correctly recognized, 0.5 is assigned
- If both instruments match the target ones, 1.0 is assigned The final score is equal to arithmetic mean of individual scores.

Acknowledgements

This work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”. It is supported by the grants N N516 368334 and N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland. Also, this work is supported by the National Science Foundation under Grant Number IIS-0968647.

References

1. Aucouturier, J.-J., Pachet, F.: Representing musical genre: A state of art, Journal of New Music Research, vol. 32 (1), pp. 83–93, 2003.

2. Chang, C., Lin, C.: LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
3. Fingerhut, M.: Music Information Retrieval, or how to search for (and maybe find) music and do away with incipits, IAML-IASA Congress, Oslo, Aug 8–13, 2004.
4. Foote, J.: Content-based retrieval of music and audio, Multimed. Storage Archiv. Syst. II, pp. 138–147, 1997.
5. Herrera, P., Amatriain, X., Batlle, E., Serra, X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques, International Symposium on Music Information Retrieval ISMIR (2000).
6. Hyoung-Gook, K., Moreau, N., Sikora, T.: MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. Wiley & Sons, 2005.
7. The International Society for Music Information Retrieval/Intern. Conference on Music Information Retrieval website http://www.ismir.net/
8. Kostek, B., Czyzewski, A.: Representing Musical Instrument Sounds for their Automatic Classification, J. Audio Eng. Soc., vol. 49, 768–785, 2001.
9. Kostek, B.: Soft Computing in Acoustics, Applications of Neural Networks, Fuzzy Logic and Rough Sets to Musical Acoustics, Studies in Fuzziness and Soft Computing, Physica Verlag, Heildelberg, New York, 1999.
10. Kostek, B.: Perception-Based Data Processing in Acoustics. Applications to Music Information Retrieval and Psychophysiology of Hearing, Springer Verlag, Series on Cognitive Technologies, Berlin, Heidelberg, New York 2005.
11. Kostek B., Kania L., Music information analysis and retrieval techniques, Archives of Acoustics, vol. 33, No. 4, pp. 483–496, 2008.
12. Lindsay A., Herre J.: MPEG-7 and MPEG-7 Audio – An Overview, vol. 49, No. 7/8, pp. 589–594, 2001.
13. Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling, in Proceedings of the First Int. Symp. on Music Information Retrieval (MUSIC IR 2000)
14. O/IEC JTC1/SC29/WG11 (2004). MPEG-7 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
15. Pachet, F., Cazaly, D.: A classification of musical genre, Proc. RIAO Content-Based Multimedia Information Access Conf., pp. 2000, 2003.
16. Panagakis, I., Benetos, E., Kotropoulos, C.: Music Genre Classification: A Multilinear Approach, Proc. Int. Symp. Music Information Retrieval, ISMIR 2008.
17. Pye, D.: Content-based methods for the management of digital music, Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2000.
18. Scheirer, E.D.: Tempo and beat analysis of acoustic musical signals, J. Acoust. Soc. Am., vol. 103(1), January 1998.
19. Tyagi, V., Wellekens, C.: On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition, Proc. ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 529–532, 2005.
20. Tzanetakis, G., Essl, G., Cook, P.: Automatic musical genre classification of audio signals, Proc. Int. Symp. Music Information Retrieval (ISMIR), 2001.
21. Tzanetakis, G., Cook, P.: Musical genre classification of audio signal, IEEE Transactions on Speech and Audio Processing, vol. 10, No. 3, pp. 293–302, July 2002.
22. WEKA: http://www.cs.waikato.ac.nz/ml/weka/
23. Wojnarski, M., Stawicki, S.,Wojnarowski, P.: TunedIT.org: System for Automated Evaluation of Algorithms in Repeatable Experiments. RSCTC 2010, LNAI vol. 6086, pp. 20–29. Web: http://tunedit.org.
24. Zwan, P., Kostek, B.: System for Automatic Singing Voice Recognition, J.Audio Eng. Soc., vol. 56, No. 9, 710–723, 2008.