www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **39**, No. 4, pp. 629–638 (2014)

DOI: 10.2478/aoa-2014-0068

Classification of Music Genres Based on Music Separation into Harmonic and Drum Components

Aldona ROSNER[1], Björn SCHULLER[2,3], Bozena KOSTEK[4]

[1] *Institute of Informatics, Silesian University of Technology*
Akademicka 16, 44-100 Gliwice, Poland; e-mail: aldona.rosner@polsl.pl

[2] *Technische Universität München*
Machine Intelligence and Signal Processing Group
80333 München, Germany; e-mail: schuller@tum.de

[3] *Imperial College London, Department of Computing*
SW7 2AZ, London, United Kingdom

[4] *Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics*
Gdańsk University of Technology
Narutowicza 11/12, 80-233 Gdańsk, Poland; e-mail: bokostek@audioacoustics.org

(*received September 9, 2014; accepted December 2, 2014*)

This article presents a study on music genre classification based on music separation into harmonic and drum components. For this purpose, audio signal separation is executed to extend the overall vector of parameters by new descriptors extracted from harmonic and/or drum music content. The study is performed using the ISMIS database of music files represented by vectors of parameters containing music features. The Support Vector Machine (SVM) classifier and co-training method adapted for the standard SVM are involved in genre classification. Also, some additional experiments are performed using reduced feature vectors, which improved the overall result. Finally, results and conclusions drawn from the study are presented, and suggestions for further work are outlined.

Keywords: Music Information Retrieval, musical sound separation, drum separation, music genre classification, Support Vector Machine, co-training, Non-Negative Matrix Factorization.

1. Introduction

The expanding consumer market for social network services and the large number of music databases demand the implementation of new functionalities for searching and analyzing musical information and examining their effectiveness and quality. Automatic genre classification has been exploited quite thoroughly in recent years, not only by the research community but also by music services and applications (ISMIR, 2014; Kostek, Hoffmann, 2014; Kostek, 2013; Kostek, Kaczmarek, 2013; Marxer, Janer, 2013; Ras, Wieczorkowska, 2010; Wieczorkowska *et al.*, 2011). However, the subject of a more deep content exploring, i.e., taking into consideration sound source separation in the context of music recognition is to some extent less visible in the literature, even though separation of individual auditory sources, apart from instrument recognition and automatic transcription systems, may be very useful in genre classification.

In recent years, extensive research has been conducted on this subject, and resulted in interesting ideas and solutions. Among the most promising one finds sinusoidal modeling (SM) (Serra, Smith, 1990) which was extensively exploited over the last two decades. Also, there are many examples of algorithms that were implemented within many research studies by e.g. Bregman (1990), Casey and Westner (2000), de Cheveigné (1993), Dziubinski *et al.* (2005), Gerber *et al.* (2012), Gillet and Richard (2008), Herrera *et al.* (2000), Klapuri (2001), Kostek and Dziubinski (2010), Tolonen (1999), Eweret *et al.* (2014).

The main objective of this article is to improve the classification results of musical genres such as Metal and Rock with high occurrence of percussion-type instruments. We separate the input signal into drum

and harmonic components, in addition to the non-separated signals in order to expand the set of features for each song. For that purpose some soft computing techniques may be used, as classification is considered to be supervised learning based on labeled training examples (Kostek, 2004; 2005). It should be remembered that data may be wrongly labeled, some data instances may be missing, etc., thus this creates many problems.

Most popular methods for music genre classification are: Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Decision Trees, Rough Sets and Minimum-distance methods, to which a very popular k-Nearest Neighbor (k-NN) method belongs (Kostek, 1999; 2005). Generally, SVMs are very widely used in Music Information Retrieval. They are considered efficient, robust and they perform relatively well in supervised classification. SVMs also 'protect' against overfitting because of structural risk minimization at the core of the algorithm (Wack *et al.*, 2009).

In our previous study (Rosner *et al.*, 2013a; 2013b) we also utilized k-NN based classification, and the results obtained showed that the SVM algorithm is a better choice for music genre classification when the original signal is used together with the signal separated into percussion and harmonic component tracks. Thus, this is the motivation for using the algorithm employed in our previous study.

In the remainder of this article we present the experimental setup which includes details on audio data and feature vectors, as well as algorithms that are employed in classification (Sec. 2). Then, the description of experiments follows. The results can be found in Sec. 3, while conclusions are presented in Sec. 4.

2. Methods

2.1. Drum separation algorithm

The effective and reliable separation of music sounds is the key to searching musical phrases in multi-pitch material (Gunawan, Sen, 2012; Nikunen *et al.*, 2012). It also forms the basis of some automatic transcription systems. Separation algorithms usually operate on a spectral analysis basis in order to determine the fundamental tones of individual voices and their harmonics. However, there are a number of technical difficulties to overcome which result from the compromise between time and frequency resolution of the analyzed signal.

Mel-frequency cepstral coefficients (MFCC) are often chosen as a metric for spectral envelope perception because of their linearity, orthogonality, and multidimensionality (Terasawa *et al.*, 2012). They were also applied in a study by Rump *et al.* (2007), which aimed at the improvement of accuracy of MFCC-based genre classification by applying the Harmonic-Percussion Signal Separation (HPSS) algorithm to the music signal, and then calculating the MFCCs on the separated signals. The authors' conclusion was that, by analyzing the MAR (Multivariate Autoregressive) features calculated on the separated signals, it was possible to achieve a good performance when all three signals (original, harmonic and percussion) were used. However, that study concentrated on an improvement of overall performance and relative error rates, while our aim is to present more specific results for each genre by using different algorithmic methodology. For this purpose, the open-source accessible drum separation algorithm implemented by Schuller *et al.* (2009) and a considerably expanded set of music audio features compared to previous studies are utilized.

Applying drum-beat separation for tempo and key detection shows that the separation into single signals parts (only drums or only harmonic parts) does not necessarily improve the results in comparison with the original signal (Schuller *et al.*, 2009). Due to that fact, we consider different mixtures of at least two signal representation types, namely: original, drum and/or harmonic. The results reported in the paper by Rosner *et al.* (2013b) confirm the assumptions that such a mixture of signals is a promising approach to music classification.

The main principle of the drum separation algorithm is employing a semi-supervised approach based on non-negative matrix factorization (NMF). Data and components are assumed to be non-negative in this approach. The aim of unsupervised learning algorithms such as vector quantization is to factorize a data matrix according to different constraints (Lee, Seung, 1999). This results in clustering the data into mutually exclusive prototypes.

NMF is an efficient method in the blind separation of drums and melodic parts of music recordings. NMF performs a decomposition of the magnitude spectrogram $\mathbf{V}$ ($V \approx W \cdot H$) obtained by Short-Time Fourier Transform (STFT), with spectral observations in columns, into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ (where $W \in R_{\geq 0}^{n \times m}$, $H \in R_{\geq 0}^{n \times m}$ and constant $r \in N$). Matrix $\mathbf{W}$ resembles characteristic spectra of the audio events occurring in the signal (such as notes played by an instrument), and matrix $\mathbf{H}$ measures their time-varying gains. Columns of $\mathbf{W}$ are not required to be orthogonal as is in principal component method.

In our experiments, we use an approach based on an iterative algorithm for computing two factors based on the Kullback-Leibler divergence of $\mathbf{V}$ given $\mathbf{W}$ and $\mathbf{H}$. This means that the factorization process is achieved by iterative algorithms minimizing cost-functions, which interprets the matrices $\mathbf{V}$ and ($\mathbf{W}$, $\mathbf{H}$) as probability distributions.

Then, to each NMF component (column of $\mathbf{W}$ and corresponding row of $\mathbf{H}$) we apply a pre-trained SVM

classifier to distinguish between percussive and non-percussive components. The task of this pre-trained SVM classification which bases on features such as harmonicity of the spectrum and periodicity of the gains is to distinguish between percussive and non-percussive signals bases. By selecting the columns of **W** that are classified as percussive and multiplying them with their estimated gains in **H**, we obtain an estimate of the contribution of percussive instruments to each time-frequency bin in **V**. Thus, we can construct a soft mask that is applied to **V** to obtain an estimated spectrogram of the drum part, which is transferred back to the time domain through the inverse STFT using the OLA (overlap-add) operation between the short-time sections in the inverting process. It should be reminded that the redundancy within overlapping segments and the averaging of the redundant samples averages out the effect of the window analysis (windowing). More details on the drum separation procedure can be found in the introductory paper by Weninger *et al.* (2011).

For straightforward reproducibility of our experiments, we used the default parameters of the publicly available drum beat separation application of the source separation toolkit openBliSSART as implemented by part of the authors (Weninger, Schuller, 2012). These parameters are as follows: frame rate 30 ms, window size 60 ms, 100 iterations, and separation into 20 NMF components.

2.2. Analyzed data and extracted audio features

In this article, we utilize samples of 30 seconds long music tracks of the ISMIS[1] music database, which are 44.1 kHz, 16 bit, stereo music excerpts. The ISMIS database consists of 470 music audio files. For 465 files it was possible to separate the drum path – these tracks were considered in our analysis. The tracks represent four music genres: Blues (11 files), Metal (77 files), Pop (129 files) and Rock (148 files). The chosen selection of genres is challenging as there is high resemblance between these music styles.

The audio feature vector consists of 191 acoustic features per representation of the music track (cf. Table 1). We prepared the audio feature vector for five different 'mixtures' of the input signals as shown in Table 2. The description of features is not given here, as most of the descriptors have roots in the MPEG 7 standard or earlier research (e.g. Kostek, Czyzewski, 2001) and are explained very thoroughly in the paper on the ISMIS competition (Kostek *et al.*, 2011).

[1]The ISMIS music database was prepared for a data mining contest associated with the 19th International Symposium on Methodologies for Intelligent Systems (ISMIS 2011, Warsaw), http://tunedit.org/challenge/music-retrieval, March 2013 (Kostek *et al.*, 2011).

Table 1. Audio features (191 in total): overview by the total number, identifier (ID), and description per type (Kostek *et al.*, 2011; Rosner *et al.*, 2013).

#	ID	Audio Feature Description
1	TC	Temporal Centroid
2	SC, SC_V	Spectral Centroid and its variance
34	ASE 1-34	Audio Spectrum Envelope (ASE) in 34 subbands
1	ASE_M	ASE mean
34	ASEV 1-34	ASE variance in 34 subbands
1	ASE_MV	Mean ASE variance
2	ASC, ASC_V	Audio Spectrum Centroid (ASC) and its variance
2	ASS, ASS_V	Audio Spectrum Spread (ASS) and its variance
24	SFM 1-24	Spectral Flatness Measure (SFM) in 24 subbands
1	SFM_M	SFM mean
24	SFMV 1-24	SFM variance
1	SFM_MV	SFM variance of all subbands
20	MFCC 1-20	Mel Function Cepstral Coefficients (MFCC) –first 20
20	MFCCV 1-20	MFCC Variance –first 20
3	THR_[1,2,3] RMS_TOT	No of samples higher than single/double/triple RMS value
3	THR_[1,2,3] RMS_10FR_MEAN	Mean of THR_[1,2,3]RMS_TOT for 10 time frames
3	THR_[1,2,3] RMS_10FR_VAR	Variance of THR_[1,2,3]RMS_TOT for 10 time frames
1	PEAK_RMS_TOT	A ratio of peak to RMS (Root Mean Square)
2	PEAK_RMS10FR_[MEAN,VAR]	A mean/variance of PEAK_RMS_TOT for 10 time frames
1	ZCD	Number of transition by the level Zero
2	ZCD_10FR_[MEAN,VAR]	Mean/Variance value of ZCD for 10 time frames
3	[1,2,3]RMS_TCD	Number of transitions by single/double/triple level RMS
3	[1,2,3]RMS_TCD_10FR_MEAN	Mean value of [1,2,3]RMS_TCD for 10 time frames
3	[1,2,3]RMS_TCD_10FR_VAR	Variance value of [1,2,3]RMS_TCD for 10 time frames

Table 2. Audio feature sets with regard to identifier (ID), their description, and the number of contained audio features.

ID	Description	#features
O	original signal	191
HD	harmonic and drum signals	382
OD	original and drum signals	382
OH	original and harmonic signals	382
OHD	original, harmonic and drum signals	573

2.3. Classification and co-training

In this study we use the co-training method which is incorporated into the SVM-based classification. Since SVMs are widely used in many classification problems, including those related to the acoustics domain, only the co-training methodology will be explained here. Co-training is a semi-supervised machine learning technique (Blum, 1998), which first learns on small training set, and then during classification of unlabelled data, the elements of the most confident predictions are used to iteratively extend the original training set. This is done by adding threshold criteria in the process of classifying data from the test set. If the prediction of classification of unlabelled data is sufficiently high (i.e., higher than the threshold criteria), then those data are marked as classified and they are added to the training set. Those steps are repeated iteratively until all elements from the test set are classified. The main advantage of this approach is that in each iteration the training set is extended by new information based on classification of new elements from the test set, so that the learning process can be improved. Unfortunately, not all elements are classified correctly and each incorrect classification introduces misleading information to the training set, which is the main disadvantage of the co-training method. Despite this, co-training is a common approach in many problems which are solved when applying machine learning, such as speech recognition, information extraction, classification and filtering, and usually gives much better results than standard methods. This is the motivation for choosing this method for our experiments, since it enhances the performance of classification.

3. Experiments and results

The experiments were conducted using the commonly used free WEKA machine learning library[2], which supports the evaluation model, implementation of classification algorithms and cross-validation methods.

SVM and co-training adapted for the standard SVM were employed as classification and learning algorithms. The SVM algorithm was selected for testing with the normalization and standardization, both using the polynomial kernel. For each test setting, a cross-validation method was used and average results were calculated. Cross-validation splits the original dataset (465 tracks) into n subsets in equal proportions, and for each run $(n-1)$ subsets form the training set, the remaining data instances form the test set. Here, the original dataset was split into three 155 data instance subsets: two for training and one test subset.

[2]WEKA 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

As described above, the co-training technique was applied with "standard" SVM. The algorithm selects the unknown and unlabeled test set instances which meet a threshold criteria of minimum confidence in the class membership after evaluation with a classification model trained on the training data. At the beginning, this threshold was set at 0.5 minimum confidence level, and the instances for which the membership to the given class was greater or equal were added to the training set and removed from the test set, so that in each iteration the training set was extended by new elements for which the probability of correctness was sufficiently high. If any elements from the test set met the threshold criteria, then the threshold was reduced by 0.1. Those steps were repeated until all elements from test sets were classified. For the sake of evaluation of the gain reached by such co-training, in each iteration the accuracy was monitored.

Figures 1 and 2 present results of the correctly classified tracks, for SVMs and co-training adapted for

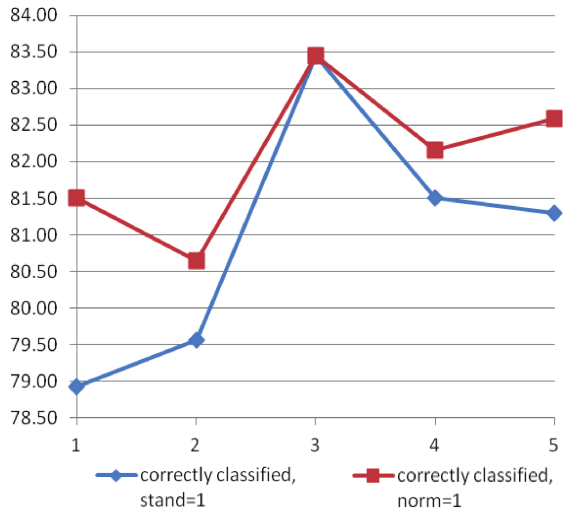

Fig. 1. Results of correctly classified tracks for SVM, where: 1 – HD, 2 – OD, 3 – OH, 4 – OHD, 5 – original signal, stand – standarization, norm – normalization.

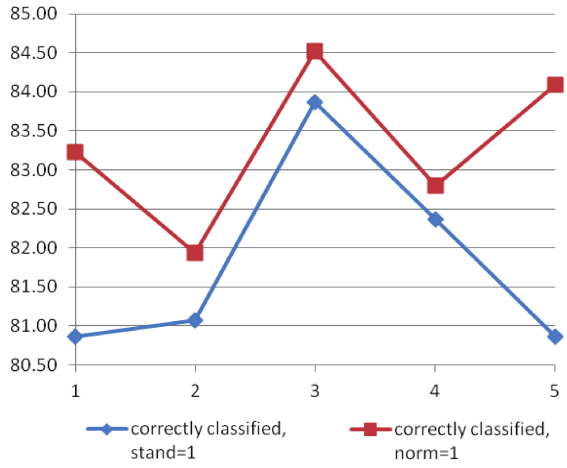

Fig. 2. Results of correctly classified tracks for Co-SVM, denotations as previously explained.

SVM (Co-SVM) methods, using normalization or standardization for different mixtures of signal, using polynomial kernel.

Figures 1 and 2 show that the best results were achieved for the OH signal, both for normalization (∼1% of improvement in comparison to the original signal in the standard SVM case and ∼0.9% in case of Co-SVM) and for standardization (∼2.1% of improvement in the standard SVM case and ∼3% in case of Co-SVM). The general correctness was the best for normalization settings, both for the SVM and the Co-SVM case.

To compare results of the experiments, the True Positives (TP) and precision measures were taken into account, where TP stands for the number of correctly classified positives of class to the total number of elements in that class and precision stands for the total number of objects classified as a specific class (including false positives) to the total number of elements in that class.

Figure 3 presents TP and precision values for the OH signal with normalization settings, both for the SVM and the Co-SVM methods.

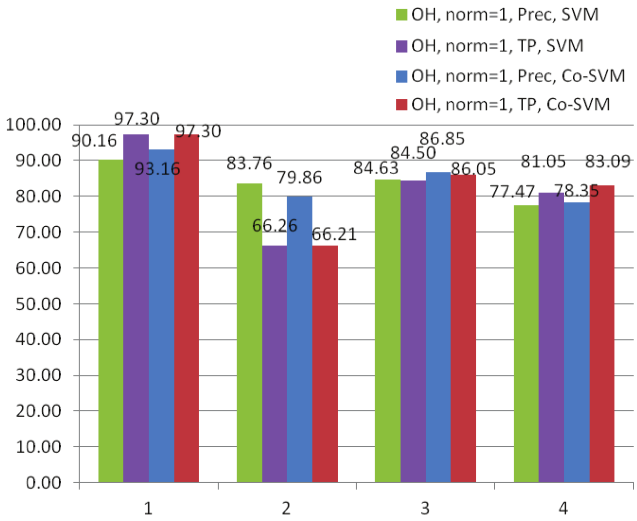

Fig. 3. True Positives (TP) and Precision (Prec) for the OH signal with normalization settings (norm=1) for SVM and Co-SVM, where: 1 – Blues, 2 – Metal, 3 – Pop, 4 – Rock.

Figure 3 shows that for three of the four genres the co-training version of SVM improves the results of classification, both while examining true positive (TP) and precision performance indicators. As seen from Fig. 3 the precision of Metal genre is worse for Co-SVM, what means that in the process of extending the training set, test instances are wrongly classified as Metal genre, which apparently disturbed the process of rebuilding the training set.

Figures 4 and 5 show the results of precision (Fig. 4) and TP (Fig. 5) for different genres using different audio features, in each case when utilizing normalization.

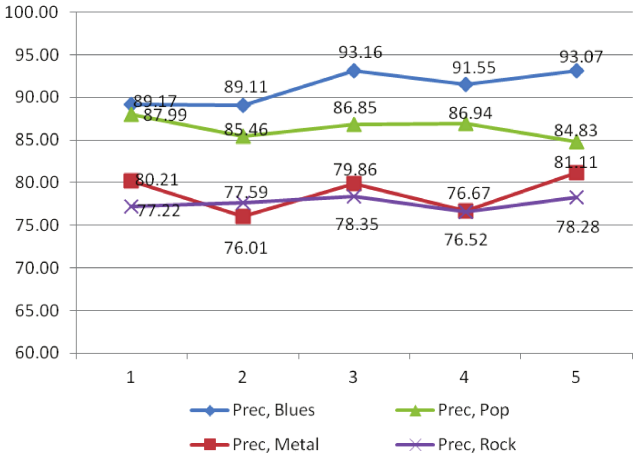

Fig. 4. Precision for different genres, where: 1 – HD, 2 – OD, 3 – OH, 4 – OHD, 5 – original signal.

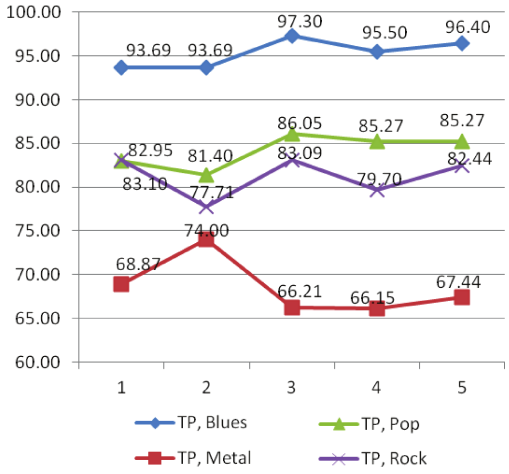

Fig. 5. TP for different genres, denotations as previously explained.

Generally the best results vere achieved for the OH signal, however, for the rock genre, the improvement in classification in comparison to the original signal is rather small (∼0.1%) and for Metal genre the best result is still achieved for the original signal.

Contrarily, the TP value for Blues and Pop for the OH signal is ∼0.9 in comparison to the original signal, thus eliminating the drum information from the signal apparently improves the classification of music styles where drums do not occur too often in our case. It is also worth to mention that the TP value for Metal is approximatelly 6.5% better for the OD signal in comparison to the original signal, what confirms the importance of drum information in cases of classes where drums occur more often.

Tables 3 and 4 present the results of classification for various genres involving different mixtures of audio signals in case of co-training, employing SVM with polynomial kernel and normalization (Table 3) or standardization (Table 4).

Table 3. Results of classification for various genres involving different mixtures of audio signals in case of co-training of SVM, employing polynomial (abrev. poly) kernel and and standardization (abrev. stand=1). Numbers in cells marked with a boldface present the highest TP and precision score for specific test settings.

Blues		Metal		Pop		Rock		Correctly classified [%]	Audio signals	Settings
TP [%]	Prec [%]	TP [%]	Prec [%]	TP [%]	Prec [%]	TP [%]	Prec [%]			
94.59	91.55	70.05	70.96	83.72	83.60	73.62	75.82	80.86	HD	poly stand =1
93.69	89.83	70.05	74.14	83.72	82.70	74.97	76.97	81.08	OD	poly stand =1
98.20	92.45	66.10	**83.13**	86.05	**85.60**	**80.37**	76.83	**83.87**	OH	poly stand =1
95.50	91.70	**70.10**	73.12	**86.82**	85.48	74.94	**77.43**	82.37	OHD	poly stand =1
95.50	**95.75**	67.44	68.56	82.95	82.57	74.97	75.05	80.86	Original signal	poly stand =1
98.20	**95.75**	**70.10**	**83.13**	**86.82**	**85.60**	**80.37**	**77.43**	**83.87**	**BEST**	

Table 4. Results of classification for various genres involving different mixtures of audio signals in case of co-training of SVM, employing polynomial (abrev. poly) kernel and normalization (abrev. norm=1). Denotations are the same as in Table 3.

Blues		Metal		Pop		Rock		Correctly classified [%]	Audio signals	Settings
TP [%]	Prec [%]	TP [%]	Prec [%]	TP [%]	Prec [%]	TP [%]	Prec [%]			
93.69	89.17	68.87	80.21	82.95	**87.99**	**83.10**	77.22	83.23	HD	poly norm=1
93.69	89.11	**74.00**	76.01	81.40	85.46	77.71	77.59	81.94	OD	poly norm=1
97.30	**93.16**	66.21	79.86	**86.05**	86.85	83.09	**78.35**	**84.52**	OH	poly norm=1
95.50	91.55	66.15	76.67	85.27	86.94	79.70	76.52	82.80	OHD	poly norm=1
96.40	93.07	67.44	**81.11**	85.27	84.83	82.44	78.28	84.09	Original signal	poly norm=1
97.30	**93.16**	**74.00**	**81.11**	**86.05**	**87.99**	**83.10**	**78.35**	**84.52**	**BEST**	

Table 5. Correctness of classification per each iteration.

# iteration	Threshold	Number of instances – in the current iteration	Number of classified instances – in total	Number of correctly classified instances – in the current iteration	Number of correctly classified instances – in total
1	0.5	75	75	65	65
2	0.5	39	114	33	98
3	0.5	19	133	16	114
4	0.5	10	143	9	123
5	0.5	5	148	5	128
6	0.5	3	151	2	130
7	0.5	2	153	2	132
8	0.4	1	154	0	132
9	0.4	1	155	0	132

The results presented in Tables 3 and 4 show that co-training technique improves the results of genre classification, even if in consecutive iterations the training set is corrupted by new, incorrectly classified elements.

As shown in Table 4, in the first iterations approximately 50% instances of the test set are classified and ∼86.67% of them are classified correctly. In the next iterations the number of classified elements is still about 50% of the remaining test set, and the overall accuracy is also above 84%.

As seen from that analysis, if the feature vector is properly chosen, this a may lead to the improvement of the automatic genre classification accuracy. For that purpose a preliminary experiment was prepared using one of Weka methods for selecting best attributes which helped achieving best accuracy, i.e. CfsSubsetEval with BestFirst method in the backward direction of search. From 382 parameters that were obtained for the OD signal, we select top 52 attributes. They are presented in Table 6.

Parameters selected for this experiment had at least 80% of effectiveness in 10-fold Cross-Validation test for CfsSubsetEval with BestFirst conducted on the OD mix of signals. As seen from Table 6 many parameters occur for both original signals as well as for drum

Table 6. Parameters used for the reduced feature vector of parameters for the OD signal.

No.	Effectiveness for 10-fold Cross-Validation [%]	# of attribute in the original feature vector	Attribute name
1	80	17	ASE14
2	100	24	ASE21
3	80	25	ASE22
4	100	26	ASE23
5	100	28	ASE25
6	80	32	ASE29
7	90	34	ASE31
8	100	35	ASE32
9	100	36	ASE33
10	100	37	ASE34
11	100	39	ASEv1
12	100	65	ASEv27
13	100	67	ASEv29
14	90	72	ASEv34
15	80	77	**ASS_v**
16	80	92	SFM15
17	100	99	SFM22
18	100	101	SFM24
19	90	102	SFM_m
20	90	109	SFMv7
21	100	110	SFMv8
22	100	111	SFMv9
23	100	115	SFMv13
24	100	117	SFMv15
25	80	118	SFMv16
26	100	124	SFMv22
27	100	125	SFMv23
28	100	126	SFMv24
29	100	127	SFM_mv
30	90	132	mfcc5
31	100	137	mfcc10
32	100	139	mfcc12
33	80	143	mfcc16
34	100	150	mfccv3
35	90	174	thr_2rms_10fr_var
36	100	185	ZCD_10fr_var
37	100	194	**d_SC_v**
38	100	225	**d_ASE31**
39	100	226	**d_ASE32**
40	100	227	**d_ASE33**
41	80	228	**d_ASE34**
42	100	229	**d_ASE_m**
43	90	255	**d_ASEv26**
44	100	258	**d_ASEv29**
45	80	276	**d_SFM8**
46	100	283	**d_SFM15**
47	80	289	**d_SFM21**
48	100	290	**d_SFM22**
49	90	292	**d_SFM24**
50	100	293	**d_SFM_m**
51	90	316	**d_SFMv23**
52	100	376	**d_ZCD_10fr_var**

signals (parameters which starts with "d_" letter). In green there are features that have the same effectiveness in the 10-fold Cross-Validation test for both types of signals (original and drum) and in blue – the ones which have different effectiveness. There is also one parameter (ASS_v) marked in red color which didn't occur for the same test conducted in case of original signal.

The total correctness of classification increased from 80.4% to 81.72% while using reduced vector of 52 parameters resulted from the Weka SVM classification using normalization and 10-fold Cross-Validation method. Also the TP for Rock genre increased from 77% to 82% and a significant increase of precision was observed for Metal genre (from 68% to 78%). As expected, the improvement for the class with high occurrence of drum instrument was gained. That confirms that the reduction parameters of the feature vector is a proper direction for further experiments.

4. Conclusion and further work

The experiments demonstrated that separating the input signal may have a positive impact on genre classification. The results confirm also that the co-training technique improves the results in case of the original signal by ∼1.5% in comparison to the standard SVM and by 1% for the OH (original and harmonic) signal, which actually produced the best general correctness of the classification. On the whole, it may be said that ranking SVM in a co-training algorithm produces better ranking results than the standard ranking SVM algorithm.

The study conducted revealed some promising areas of further research. We plan to investigate the presented approach using larger dataset with more genres that would provide a more representative instrument content context. Also, for further improvement of classification performance, the feature vector should be optimized, i.e. the number of parameters should be reduced. This was already confirmed in our preliminary experiments in which the feature vector was reduced. The True Positives predictions indicator (TP) increased for Rock genre from 77% to 82%, and for Metal genre the TP gain was even more promising as it rose from 68% to 78%.

It should, however, be remembered that certain phenomena occurring in music, as e.g. musical articulation like tremolo, glissando, transients with non-harmonic spectra make problematic towards seamless music source separation. This is stated in many related studies in the literature (Kleczkowski, 2012; Beauchamp, 2011; Lohri *et al.*, 2012; Mika, Kleczkowski, 2011; Sofianos *et al.*, 2012; Terasawa *et al.*, 2012). Another type of problem concerns the overlapping harmonics of individual sounds. This phenomenon impedes obtaining the original timbre of the separated sound sources. An obvious solution to this problem is having all sound sources recorded separately and then mixed. This, however, requires a lot of additional effort and resources, and in most cases is unattainable.

In future experiments the reduced vector could be expanded with new parameters, which would represent the features of a specific instrument. Following this, the full drum feature set might be reduced to a single feature vector containing only few most important parameters.

Acknowledgments

The work has partially been supported by the European Union from the European Social Fund (grant agreement number: UDA-POKL.04.01.01-00-106/09) and by the project no. PBS1/B3/16/2012 entitled "Multimodal system supporting acoustic communication with computers" financed by the Polish National Centre for R&D.

References

1. Beauchamp J. (2011), *Perceptually Correlated Parameters of Musical Instrument Tones*, Archives of Acoustics, **36**, 2, 225–238.
2. Blum A., Mitchell T. (1998), *Combining labeled and unlabeled data with co-training*, Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann, 92–100.
3. Bregman A. (1990), *Auditory scene analysis: the perceptual organization of sound*, MIT Press.
4. Casey M., Westner A. (2000), *Separation of mixed audio sources by independent subspace analysis*, Proceedings of International Computer Music Conference, 154–161, Berlin.
5. de Cheveigné A. (1993), *Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing*, J. Acoust. Soc. Am..
6. Dziubinski M., Dalka P., Kostek B. (2005), *Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks*, J. Intel. Inform. Systems, **24**, 2, 133–157.
7. Eweret S., Prado B., Muller M., Plumbley M. (2014), *Score-Informed Source Separation for Musical Audio Recordings,* IEEE Signal Proc. Magazine, 116–124.
8. Gerber T., Dutasta M., Girin L., Févotte C. (2012), *Professionally-produced music separation guided by covers*, 13th International Society for Music Information Retrieval Conference.
9. Gillet O., Richard G. (2008), *Transcription and separation of drum signals from polyphonic music*,

IEEE Transactions on Audio, Speech and Language Processing, **16**, 529–540.

10. Gunawan D., Sen S. (2012), *Separation of Harmonic Musical Instrument Notes Using Spectro-Temporal Modeling of Harmonic Magnitudes and Spectrogram Inversion with Phase Optimization*, JAES, **60**, 12, 1004–1014.

11. Herrera P., Amatriain X., Batlle E., Serra X. (2000), *Towards instrument segmentation for music content description: a critical review of instrument classification techniques*, Proceedings of International Symp. on Music Information Retrieval, Plymouth, Massachusetts.

12. ISMIR, Intern. Conference on Music Information Retrieval website (http://ismir2014.ismir.net).

13. Klapuri A. (2001), *Multipitch estimation and sound separation by the spectral smoothness principle*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3381–3384, Salt Lake City.

14. Kleczkowski P. (2012), *Perception of Mixture of Musical Instruments with Spectral Overlap Removed*, Archives of Acoustics, **37**, 3, 355–363.

15. Kostek B. (1999), *Soft Computing in Acoustics, Applications of Neural Networks, Fuzzy Logic and Rough Sets to Musical Acoustics, Studies in Fuzziness and Soft Computing,* Physica Verlag.

16. Kostek B. (2004), *Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques*, Proceedings of the IEEE, **92**, 4, 712–729.

17. Kostek B. (2005), *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.

18. Kostek B. (2013), *Music Information Retrieval in Music Repositories*, Chapter 17, in: Rough Sets and Intelligent Systems (Skowron A., Suraj Z., Eds.), Vol. 1, ISRL, 42, 463–489, Springer Verlag, Berlin Heilderberg.

19. Kostek B., Czyzewski A. (2001), *Representing Musical Instrument Sounds for Their Automatic Classification*, J. Audio Eng. Soc., **49**, 9, 768–785.

20. Kostek B., Dziubinski M. (2010), *Evaluation of the separation algorithm performance employing ANNs*, 34, Springer Verlag, in: Advances in Intelligent and Soft Computing, 80, 27–37, Berlin, Heidelberg.

21. Kostek B., Hoffmann P. (2014), *Music Data Processing and Mining in Large Databases for Active Media*, Active Media Technology, 2014, LNCS 8610, pp. 85–95, Springer International (Slezak et al., Eds.).

22. Kostek B., Kaczmarek A. (2013), *Based on Based on Multidimensional Description and Similarity Measures*, Fundamenta Informaticae, 1001–1017, 1001, DOI 10.3233/FI-2012-0000.

23. Kostek B., Kupryjanow A., Zwan P., Jiang W., Ras Z., Wojnarski M., Swietlicka J. (2011), *Report of the ISMIS 2011 Contest: Music Information Retrieval, Foundations of Intelligent Systems*, ISMIS 2011, Springer Verlag, 715–724, Berlin, Heidelberg.

24. Lee D.D., Seung H.S. (1999), *Learning the parts of objects by non-negative matrix factorization*, Nature, 401, 788–791.

25. Liutkus A., Pinel J., Badeau R., Girin L., Richard G. (2012), *Informed source separation through spectrogram coding and data embedding*, Signal Processing, **92**, 8, 1937–1949.

26. Lohri A., Carral S., Chatziioannou V. (2012), *Combination Tones in Violins*, Archives of Acoustics, **36**, 4, 727–740.

27. Marxer R., Janer J. (2013), *Study of regularizations and constraints in NMF-based drums monaural separation*, International Conference on Digital Audio Effects Conference (DAFx-13).

28. Mika D., Kleczkowski P. (2011), *ICA-based Single Channel Audio Separation: New Bases and Measures of Distance*, Archives of Acoustics, **36**, 2, 311–331.

29. Nikunen J., Virtanen T., Vilermo M. (2012), *Multichannel Audio Upmixing by Time-Frequency Filtering Using Non-Negative Tensor Factorization*, JAES, **60**, 10, 794–806.

30. Ras Z., Wieczorkowska A. (2010), *Advances in Music Information Retrieval*, Springer Publishing Company.

31. Rosner A., Michalak M., Kostek B. (2013a), *A Study on Influence of Normalization Methods on Music Genre Classification Results Employing kNN Algorithm*, Proceedings 9th National Conference on Databases: Applications and Systems, pp. 411–423, Ustron.

32. Rosner A., Weninger F., Schuller B., Michalak M., Kostek B. (2013b), *Influence of Low-Level Features Extracted from Rhythmic and Harmonic Sections on Music Genre Classification*, ICMMI'2013, Gruca A., Czachrski T., Kozielski S. [Eds.], Man-Machine Interactions 3, volume 242 of Advances in Intelligent Systems and Computing (AISC), pages 467–473, Springer.

33. Rump H., Miyabe S., Tsunoo E., Ono N., Sagama S. (2010), *Autoregressive MFCC Models For Genre Classification Improved By Harmonic-Percussion Separation*, Proceedings of the 11th International Society for Music Information Retrieval Conference, pp. 87–92, Utrecht.

34. Serra X., Smith J.O. (1990), *Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition*, Computer Music Journal, **14**, 4, 12–24.

35. Schuller B., Lehmann A., Weninger F., Eyben F., Rigoli G. (2009), *Blind Enhancement of the Rhythmic and Harmonic Sections by NMF: Does it help?*, NAG/DAGA 2009, Rotterdam, The Netherlands, Sofianos S., Ariyaeeinia A., Polfreman R., Sotudeh R. (2012) H-Semantics: *a Hybrid Approach to Singing Voice Separation*, JAES, **60**, 10, 831–841.

36. Terasawa H., Berger J., Makino S. (2012), *In Search of a Perceptual Metric for Timbre: Dissimilarity Judgments among Synthetic Sounds with MFCC-Derived Spectral Envelopes*, JAES, **60**, 9, 674–685.

37. Tolonen T. (1999), *Methods for Separation of Harmonic Sound Sources using Sinusoidal Modeling*, 106th Audio Engineering Society Conv., Munich.

38. Wack N., Guaus E., Laurier C., Meyers O., Marxer R., Bogdanov D., Serra J., Herrera P. (2009), *Music Type Groupers (Mtg): Generic Music Classification Algorithms*, International Society for Music Information Retrieval.

39. Weninger F., Durrieu J., Eyben F., Richard G., Schuller B. (2011), *Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition*, Proceedings of International Conference on Acoustics Speech and Signal Processing, pp. 2196–2199, IEEE, Prague, Czech Republic.

40. Weninger F., Schuller B. (2012), *Optimization and parallelization of monaural source separation algorithms in the openblissart toolkit*, J. Signal Processing Systems, **69**(3), 267–277.

41. Wieczorkowska A., Kubera E., Kubik-Komar A. (2011), *Analysis of Recognition of a Musical Instrument in Sound Mixes Using Support Vector Machines*, Fundamenta Informaticae, **107**, 1.