J Intell Inf Syst (2018) 50:363–384
DOI 10.1007/s10844-017-0464-5

Automatic music genre classification based on musical instrument track separation

Aldona Rosner[1] · **Bozena Kostek**[2]

Received: 12 January 2017 / Revised: 17 April 2017 / Accepted: 20 April 2017 /
Published online: 12 May 2017

Abstract The aim of this article is to investigate whether separating music tracks at the pre-processing phase and extending feature vector by parameters related to the specific musical instruments that are characteristic for the given musical genre allow for efficient automatic musical genre classification in case of database containing thousands of music excerpts and a dozen of genres. Results of extensive experiments show that the approach proposed for music genre classification is promising. Overall, conglomerating parameters derived from both an original audio and a mixture of separated tracks improve classification effectiveness measures, demonstrating that the proposed feature vector and the Support Vector Machine (SVM) with Co-training mechanism are applicable to a large dataset.

Keywords Music information retrieval (MIR) · Automatic music genre classification · Automatic separation of music tracks · Support vector machine (SVM)

1 Introduction

A special issue on automatic processing of music information research, edited by Herrera-Boyer et al. (2013), accounts on how the past of MIR (Music Information Retrieval), combined with realistic perspectives on the future of specific topics, influence the future

✉ Aldona Rosner
aldona.rosner@polsl.pl

Bozena Kostek
bokostek@audioacoustics.org

[1] Institute of Computer Science, Silesian University of Technology, Gliwice, Poland

[2] Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdańsk, Poland

of this area. One may argue that due to the amount of available music-related information, expert-based knowledge may start to be obsolete as it is already preserved in multiple data records. Thus, the future research in MIR domain may eventually lead to deep machine learning (Humprey et al. 2013). On the other hand, Lee and Cunningham (2013) show that understanding users' needs, behavior and requirements may have a great impact on developing a system that addresses critical concepts of MIR. Accordingly, they recommend to increase the visibility and impact of user-based studies in the field. A very comprehensive review of topics related to MIR was recently prepared by Schedl et al. (2014). The study contains over 300 references, however this is only a small fraction of the literature sources devoted to Music Information Retrieval. Even a more recent survey was prepared by Burgoyne et al. (2016), in which research performed within MIR was presented as an important part of a rapidly evolving area called digital humanities.

Automatic music genre classification (AMGC) has been exploited quite thoroughly in recent years by the research community (ISMIR conferences, ISMIR (2016)) and is one of the most popular search query choices within the MIR domain (Bergstra et al. 2006; Burred 2014; Kostek 2005; Ntalampiras 2013; Schedl et al. 2014; Silla et al. 2007; Sturm 2013; Tzanetakis et al. 2002)). On a smaller scale, a survey, focusing on AMGC was presented by Silla et al. (2007). They observed that the typical approach adopted to the AMGC is based on feature space decomposition and machine learning to assign music genre labels. Also some non-conventional machine learning strategies to AMGC exist, based on both space and time decomposition schemes, an example of which may be again the work of Silla et al. (2007). Features employed in their work were selected from several parts of a music excerpt, as well as from the entire music signal. They used a combination of binary classifiers, the results of which were merged to produce the final music genre labeling (Silla et al. 2007). Another, non-conventional approach was shown in the work by Sturm (2014), as well as by Bergstra et al. (2006). The AdaBoost algorithm, performing the classification iteratively by combining the weighted votes of several weak learners, was utilized. Yet, a novel data selection strategy based on Gaussian mixture model clustering for the creation of the Universal Modeling (UM) was introduced by Ntalampiras (2013). The scheme considered the dataset characteristics, adapted itself to them and achieved increased recognition rates in comparison to the conventional approach. Very recently, a Special Issue on Intelligent Audio Processing, Semantics, and Interaction was prepared, in which it was pointed out that semantic audio incorporates the processes of intelligent audio processing and augmented (semantic) interaction, thus broadening the area of music information retrieval (Kalliris et al. 2016).

AMGC is still an ongoing process, especially in the context of scalability, as most of the studies were carried out on the databases delivered either by ISMIR, MIREX, ISMIS conferences or those available in the Internet, e.g. GTZAN, RWC-MDB (Real World Computing Music Database) (Goto et al. 2002), Magnatune, etc., typically including approx. 1,000–2,000 music pieces assigned to a few popular music genres (see e.g. Bergstra et al. (2006)). There are some larger collections of music excerpts, e.g. Latin Music Database containing 3,160 music pieces categorized in 10 musical genres (Silla et al. 2007). In many cases, such databases are labeled manually, which means that audio files are correctly assigned to the corresponding music genre, however the assignment is carried out on subjective basis. This aspect may have a very positive impact on the effectiveness of classification experiment. Still, as reported in the literature, for low level-feature-based approach and multi-class recognition, the effectiveness of music genre classification is in the range of 60–80% (Bergstra et al. 2006; Tzanetakis et al. 2002; Holzapfel and Stylianou 2008; Kostek et al. 2011) with some exceptions (see e.g. Ntalampiras (2013)). It is worth mentioning that

the above-mentioned collections are not consistent among one another, as they differ in the number of musical pieces, file format, bit resolution, number of genres, etc., hence, a full comparison within these databases is justified only to some extent.

As pointed out by Silla et al. (2007) music genres are categorical labels created by human experts in order to identify the style of the music and organize music collections. "Music genre" notion may not precisely be defined, however the research comprising music categorization, as stated by Tekman and Hortacsu, still plays an essential role in music appreciation and cognition (Tekman and Hortacsu 2002). One may argue that the expanding consumer market for social music network services brought new ways for searching and analyzing musical information and examining their effectiveness and quality, i.e. based on collaborative filtering and similarity measures retrieved from large music archives (Schedl et al. 2014; Ness et al. 2009). However, the subject of a deeper content exploring, i.e. considering the sound source separation in the context of music recognition, starts to be useful in improving genre classification (McKay and Fujinaga 2004; Pérez-García et al. 2010; Zhu et al. 2004). This also is visible in some new applications (e.g. AudioScore Ultimate 7. (2016)).

Finally, music genre is related to the thematic identity of radio broadcasting shows, therefore to the underlying (semantic) relations between radio producers, content and consumers (Fu et al. 2011; Kotsakis et al. 2012; Romain et al. 2012) with many practical uses in media analytics and broadcasting programming. Similar audio-driven semantic analysis approaches (including Music Genre Recognition) can also be considered for the case of video content, thus leading to various semantic conceptualization outcomes (i.e. related to activities: dancing, signing, jogging, skiing etc., occasions: birthday, graduation etc., and others) (Lee and Ellis Daniel 2010). These are examples of intelligent information systems that will dominate in the upcoming (fully deployed) Semantic Web in the near future.

The presented work is a part of a larger framework carried out over the past several years. The authors and their collaborators performed several studies devoted to AMGC (Kostek 2013; Kostek et al. 2011; Kostek et al. 2014; Plewa and Kostek 2015; Rosner et al. 2014; Rosner and Kostek 2015), in which decision algorithms, such as: kNNs (k Nearest Neighbors), SVM by Sequential Minimal Optimization (SMO) algorithm, Rough Sets and Bayesian Networks were used. Recently, a paper was published by one of the authors and her Ph.D. student (Hoffmann and Kostek 2015), which presents a novel approach to the Virtual Bass Synthesis (VBS) applied to mobile devices, called Smart VBS (SVBS). Improving the low frequency sound of mobile devices is a problem that appears in many studies (Hill and Hawksford 2010; Mu and Gan 2012, 2015; Oo et al. 2000). The proposed algorithm uses a rule-based settings of bass synthesis parameters adjusted according to the recognized music genre. To perform harmonic generation based on a nonlinear device (NLD) method an intelligent controlling system, automatically adapting to the recognized music genre, was proposed (Hoffmann and Kostek 2015). Lately, a patent application was prepared in which the above described approach has been extended to separating music tracks before the NLD settings are adjusted. Thus, the motivation behind the presented study is to provide answers with regard to the content of the feature vector derived from separated tracks, also to what extent separating tracks helps to distinguish between genres, and which genres make the most use of track separation. The last question has alaredy been asked by Wieczorkowska et al. (2011) with regard to recognition the dominating musical instrument in sound mixes.

The aim of this research study is two-fold, namely: to propose a feature vector created based on separated audio tracks but retaining parameters derived from the original excerpt. This may be important in the context of the nature of musical genres. For example, it is well known that some genres (e.g. rock, hard rock, techno, etc.) are characterized by rich rhythmic patterns that possibly translate, among others, into the values of energy and temporal

descriptors. The authors' approach differs from other studies shown in the literature. When the track separation is performed on audio source, a music excerpt is separated into harmonic and drum tracks. We have expanded this by extracting features that are related to individual music instruments that may be characteristic to the specific genre. Then, having several individual tracks, we checked whether it is sufficient to build a feature vector based on descriptors derived only from individual tracks or whether to include those from the whole music excerpt as well, assuming the separation is not perfect because of the estimation inaccuracies. The second goal is to check whether the feature vector derived from this study enables to effectively classify musical genres in the case of a database containing thousands of records and a dozen of musical genres, i.e. with a similar correctness comparing to earlier experiments carried out on much smaller music databases.

The paper is organized as follows, Section 2 presents the experimental setup starting with a concise description of the database employed and parameters utilized. Research studies devoted to the music separation process are then recalled and a methodology involving music track separation that was utilized by the authors is explained. Finally, the preprocessing stage with regard to building a feature vector for the genre classification process is shown. Section 3 contains a short description of the classification algorithm used in this study, focusing on the so-called co-training mechanism. Section 4 discusses results of experiments that were carried out for optimizing feature vectors. Overall comments are included in Summary.

2 Experimental setup

2.1 Music database

For the purpose of experiments a subset of audio excerpts extracted from the Synat database belonging to 13 popular music genres was used (SYNAT 2016). In addition, a dataset of musical instrument samples was collected from the Sampleswap music service (Sampleswap 2016). As it contains samples of various instrument sounds, as well as examples of instruments playing in the loop, the dataset also provided longer sections of particular instruments. The samples of three musical instruments were collected for the experiment: piano, trumpet and saxophone.

The Synat database (Kostek et al. 2014; SYNAT 2016) stores over 50.000 music tracks of 30-second long song excerpts in mp3 format, representing the following 22 genres: Alternative Rock, Blues, Broadway & Vocalists, Children's Music, Christian&Gospel, Classic Rock, Classical, Country, Dance & DJ, Folk, Hard Rock & Metal, International, Jazz, Latin Music, Miscellaneous, New Age, Opera & Vocal, Pop, Rap & Hip-Hop, Rock, R&B, and Soundtracks. The whole database is parameterized employing a feature vector shown in the subsequent Section. For the experiments carried out within this study over 8,000 music excerpts representing 13 music genres were selected. They are as follows: Alternative Rock, Blues, Classical, Country, Dance & DJ, Hard Rock & Metal, Jazz, Latin Music, New age, Pop, R&B, Rap & Hip-Hop, Rock. Music genres chosen for the analysis represent sufficiently diverse, yet similar music material. Also, they were utilized in other research works. This way we could indirectly compare the obtained results with findings from the literature sources.

It should be pointed out that the constructed music robot assigned songs to the genres (i.e. classes in the Synat database) according to their ID3 tags. These tags were saved in a fully automatic way without human control. It should be reminded that the ability of humans

to distinguish between complex music genres is based strongly on context-dependent inferences and is far from being perfect (Tzanetakis et al. 2002). Hence, the decision systems trying to mimic human's way of analyzing music may not be capable to do it with a very high effectiveness.

2.2 Parametrization

Feature extraction plays a crucial part in the genre recognition process, thus this stage should be carefully controlled and optimized. Feature vectors (FVs) for music genre classification are usually based on low-level descriptors from the MPEG-7 standard (Lindsay and Herre 2001; Hyoung-Gook et al. 2005), Mel-Frequency Cepstral Coefficients (MFCCs) (Tzanetakis et al. 2002) or finally, dedicated parameters suggested by researchers (Kostek 1999; Kostek et al. 2011; Liu et al. 2007; Nayak and Bhutani 2011; Salamon et al. 2012; Silla et al. 2007). Table 1 presents a list of parameters contained in the Synat database (Kostek et al. 2014). Most parameters are based on the MPEG-7 standard, and the remaining ones are the MFCC descriptors and time-related dedicated parameters (Kostek et al. 2011). Since definitions of these parameters are well-known or easily found in the literature sources, they are not to be recalled here. It is interesting, however, that the same set of parameters was used in a study on music mood classification and brought sufficient effectiveness (Plewa and Kostek 2015).

2.3 Music track separation

In recent years, an extensive research has also been conducted on the subject of audio sound separation, and resulted in interesting ideas and solutions. Among the most promising, one finds sinusoidal modeling (SM) (Serra and Smith 1990) that was extensively exploited over the last two decades. There are also many examples of algorithms that were implemented within many research studies (Bregman 1990; Casey and Westner 2000; de Cheveigne 1993; Dziubiński et al. 2005; Eweret et al. 2014; Gerber et al. 2012; Gillet and Richard 2008; Herrera et al. 2000).

Uhle et al. (2003) designed a system for drum beat separation based on Independent Component Analysis. In contrast, Smaragdis and Brown (2003) applied Non-Negative Matrix Factorization (NMF) to create a system for transcription of polyphonic music with a special focus on piano music. In the study of Helen and Virtanen (2005) NMF is used, combined with a feature extraction and classification process. They have got good results in drum beat separation from pop music. The same methodology was used by Paulus and Virtanen (2005) for drum transcription.

In this study, a semi-supervised instrument separation based on NMF is adopted to the authors' needs. The main principles of the NMF-based methodology are first recalled with a focus on cost function minimization.

The main principle of the drum separation algorithm is employing a semi-supervised approach based on non-negative matrix factorization (NMF). The aim of unsupervised learning algorithms such as vector quantization is to factorize a data matrix according to different constraints (Lee and Seung 1999). This results in clustering the data into mutually exclusive prototypes. The general idea of NMF is to separate input audio track into several isolated audio tracks, representing specified components such as rhythmic or melodic part.

NMF is an efficient method used in the blind separation of drums and melodic parts of music recordings. NMF performs a decomposition of the magnitude spectrogram

Table 1 Audio features: identifier (ID) and description per type

#	ID	Audio Feature Description	*comment*
1	TC	Temporal Centroid	
2	SC, SC_V	Spectral Centroid – average and its variance	
34	ASE 1-34	Audio Spectrum Envelope (ASE) – average values in 34 frequency bands	*29 subbands as audio files are in .mp3 format*
1	ASE_M	Mean ASE (for all frequency bands)	
34	ASEV 1-34	ASE variance in 34 frequency bands	*as above*
1	ASE_MV	Mean ASE variance (for all frequency bands)	
2	ASC, ASC_V	Audio Spectrum Centroid (ASC) – average and its variance	
2	ASS, ASS_V	Audio Spectrum Spread (ASS) – average and its variance	
24	SFM 1-24	Spectral Flatness Measure (SFM) – average values for 24 frequency bands	*20 subbands*
1	SFM_M	Mean SFM (for all frequency bands)	
24	SFMV 1-24	SFM variance (for 24 frequency bands)	*20 subbands*
1	SFM_MV	Mean SFM variance (for all frequency bands)	
20	MFCC 1-20	Mel Frequency Cepstral Coefficients (MFCC) – first 20 (mean values)	
20	MFCCV 1-20	MFCC Variance – first 20	
3	THR_[1,2,3] RMS_TOT	No of samples higher than a single/ double/triple RMS value	*Dedicated parameters (24) in time domain based on the analysis of the distribution of the signal envelope in relation to the RMS value*
6	THR_[1,2,3]RMS_10 FR_[MEAN,VAR]	Mean/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_RMS10 FR_[MEAN,VAR]	A mean/variance of PEAK_RMS_TOT for 10 time frames	
1	ZCD	Number of transition by the level Zero	
2	ZCD_10 FR_[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	
6	[1,2,3]RMS_TCD_10 FR_[MEAN,VAR]	Mean/Variance value of [1,2,3]RMS_TCD for 10 time frames	
	TOTAL number of parameters	173	

$\mathbf{V}\ \mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$) obtained by Short-Time Fourier Transform (STFT), with spectral observations in columns, into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ where $\mathbf{W} \in R_{\geq 0}^{m \times r}$, $\mathbf{H} \in R_{\geq 0}^{r \times n}$ and a constant $r \in N$. Columns of matrix $\mathbf{W}$ resembles characteristic spectra of the audio

events occurring in the signal (such as notes played by an instrument), and rows in matrix **H** measures their time-varying gains. Columns **W** are not required to be orthogonal as in Principal Component Analysis (PCA).

For $r \ll n, m$, there exists generally only an approximate solution. Factorization is achieved by iterative algorithms minimizing cost-functions, as presented in (1) (Schuller et al. 2009):

$$\begin{array}{ll} (\mathbf{V}-\mathbf{WH})^2 & \text{Squared error} \\ \|(\mathbf{V}-\mathbf{WH})\|_F & \text{Frobenius norm} \\ \sum_{ij}\left(V_{ij}\log\frac{V_{ij}}{(\mathbf{WH})_{ij}}-V_{ij}+(\mathbf{WH})_{ij}\right) & \text{Modified KL Divergance} \end{array} \tag{1}$$

The first two cost-functions are closely related to each other, both minimizing some form of quadratic error, the Modified KL Divergence interprets the matrices **V** and (**WH**) as probability distributions and minimizes their divergence. The modification of the Kullback-Leibler (KL) divergence lies in an additional term $(\mathbf{WH})_{ij}-V_{ij}$, added not only to introduce a measurement of the absolute error, but also to ensure non-negativity.

In the experiments carried out, an approach employing an iterative algorithm for computing two factors based on the Modified Kullback-Leibler divergence of **V** given **W** and **H** was used. A pre-trained SVM (Support Vector Machine) classifier was applied to each NMF component (column of **W** and the corresponding row of **H**) to distinguish between percussive and non-percussive components based on such features as harmonicity of the spectrum and periodicity of the gains. 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 Schuller et al. (2009).

2.3.1 OpenBliSSART

The openBliSSART application is a C++ toolbox that provides Blind Source Separation for Audio Recognition Tasks (Weninger et al. 2011). Besides the basic blind (unsupervised) source separation, classification by Support Vector Machines (SVM) using common acoustic features from speech and music processing is implemented. A GUI is available based on cross-platform application framework Qt (Qt 2016) for the source component playback and data set creation. It includes various source separation algorithms, with a strong focus on variants of Non-Negative Matrix Factorization (NMF). Furthermore, supervised NMF can be performed for source separation as well as audio feature extraction (Weninger et al. 2017). It should be noted that openBliSSART has built-in components to separate the HARMONIC and DRUM instruments. However, the toolkit also enables to import audio files (in order to define new instrument components), create label (to define new instrument's name), and create response (to define which instruments should be considered in the separation process). In our study, we have introduced samples of new instruments (piano, trumpet, saxophone) to teach the built-in SVM classifier. Musical instrument samples were collected from the Sampleswap music service (Sampleswap 2016).

2.3.2 Feature vectors built on separated music tracks

This part of the experiment includes separating the input signal in order to obtain the signal of the specific instrument, such as: harmonic part of the input audio track, drum signal (percussion), piano, trumpet, saxophone. Therefore the same parameters as presented in Table 1 were calculated additionally for the separated music tracks. In that way, vectors of parameters (VoPs) were obtained, i.e. the FV containing the original track was extended by new parameters derived from the separated signal. Therefore, feature vectors derived from original and harmonic signals (denoted as OH), original and drum (OD), original and piano (OP), etc., as well as from mixtures of more than two signals (e.g. original + drum + harmonic resulted in OHD FVs) were created. This strategy assumes that the separation process may not be perfect.

2.3.3 Normalization methods

Data normalization is a scaling of original data to the specified range, e.g. $[-1,\ 1]$ or $[0,\ 1]$, which is useful in data exploration and specifically for neural networks. In the study performed, the most popular methods of normalizations, such as: Min-Max and Zero-Mean (ZScore) were applied and tested in the pre-study. Min-Max normalization is a linear transformation on the original data usually to the range $[0,\ 1]$. Zero-Mean normalization takes into account the fact that the mean value should equal zero after the normalization process. The normalization of training and test datasets designated for the decision algorithms are performed in the same way as for Min-Max normalization – the mean and standard deviation values are calculated only for training dataset, and only the current value is retained from training and test datasets (used respectively for normalization of training and test datasets).

2.4 Experimental setup

As described above, the experimental setup consisted of several steps (see Fig. 1). As observed in Fig. 1, the feature extraction is performed on original (O) audio and separated (harmonic (H), drum (D), trumpet (T), piano (P), saxophone (S)) signals. All feature vectors (FVs) are then normalized and optimized. When performing the Non-Negative Matrix Factorization-based separation, the following configuration was used: cost function (Modified KL-divergence), window sizes (20 ms, 30 ms or 40 ms), window function (square root of Hann function), window overlap (0.5), number of components (5, 10, 20 or 30). After including parameters derived from separated signals to the original FV, they form an expanded feature vector, which is also optimized (based on the reduction of the number of attributes – i.e. Best First, Greedy Stepwise, Ranker). This feature vector is called the vector of parameters (VoP). Finally, the derived VoPs are employed in the classification process by means of the co-training mechanism applied for SVM (which is described in the next Section). The last step involves selection of optimum classification algorithm parameters and settings.

3 Classification process

As already mentioned, there were two stages of experiments, namely one focused on FVs optimization and the other one was devoted to evaluating music genre classification effectiveness. Several algorithms were employed in the pre-study phase, namely *k*Nearest

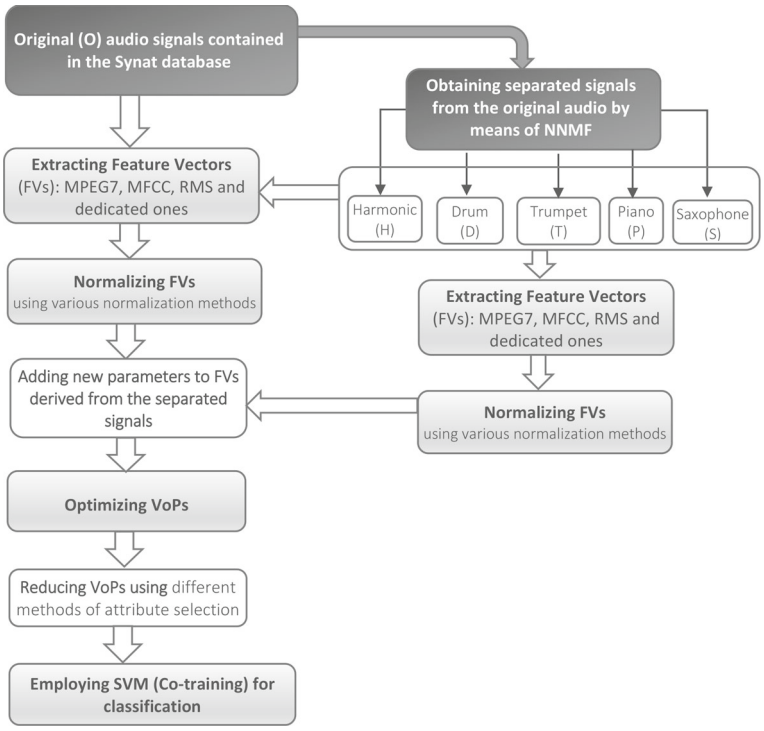

Fig. 1 Experimental setup

Neighbors (*k*NNs) algorithm, Support Vector Machine (SVM), both algorithms with- and without the co-training mechanism, as well as Random Forests. The results achieved for music genre classification using these algorithms are approximately within the same range of accuracy. As the best effectiveness was obtained while using the Support Vector Machine algorithm with a co-training mechanism, consequently, the results for SVM (co-training), will only be presented. First, some basic information concerning SVM is recalled below.

SVM uses a nonlinear mapping to transform the original training data into new space. Within this new space, it searches for the optimal separating hyperplane (i.e., a "decision boundary" separating the instances of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane, which is found by using support vectors ("essential" training data elements) and margins (defined by the support vectors). The SVM method is accurate thanks to its ability to model complex, nonlinear decision boundaries. It is much less prone to overfitting (especially as the cross-validation procedure is utilized) (Hsu et al. 2003) than other methods and can also provide a compact description of the learned model. Weka implementation of the SVM algorithm is the SMO function (Weka library 2016) that allows for using normalization or standardization of the input data as the preprocessing step, additionally enabling to determine different kernel functions, such as linear, polynomial or RBF (Gaussian radial basis function). Details of the decision-making stage involving machine learning with the cross-validation approach are presented in Fig. 2.

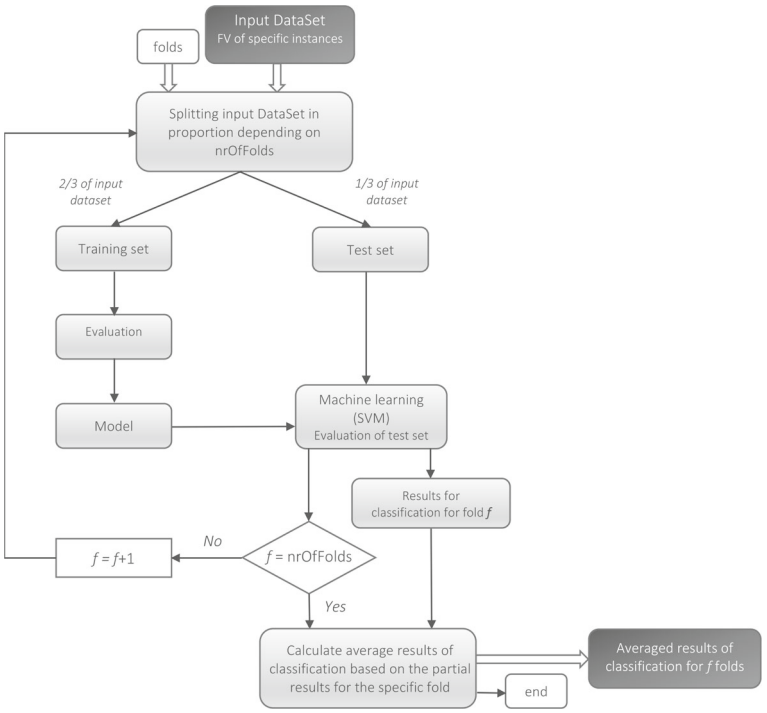

Fig. 2 Classification process using the cross-validation approach

3.1 Co-training method

Co-training (Blum and Mitchell 1998) is an example of semi-supervised machine learning technique, which uses labeled and unlabeled data to build a classifier. It initially learns on small training set, then during the classification of unlabeled data, the elements of the most confident predictions are used to iteratively extend the original training set (Xiaojin and Goldberg 2009). This is done by adding a threshold criterion in the process of classifying the data from the test set. If the prediction of classification of unlabeled data is sufficiently high (i.e. higher than the threshold criterion), such data are marked as classified and they are added to the training set. This is repeated up to the stage when all the elements from the test set are classified.

The main advantage of such an approach is that in each iteration, the training set is extended by new information based on the classification of new elements from the test set, which can improve the learning process. Contrarily, the disadvantage of the method is that if the elements are not classified correctly, it introduces misleading information to the training set. Regardless of that, co-training is a common approach in the machine learning-based problem solutions and usually gives much better results than the standard methods. Since the co-training method enhances the performance of classification, it was decided to be applied in the experiments (Rosner et al. 2013, 2014; Rosner and Kostek 2015).

3.2 Effectiveness measures

Sturm (2013) indicated that presenting accuracy alone is not sufficient in accurate interpretation of results obtained in the evaluation of music recognition. Therefore, in this study, the following measures: *True Positive* (TP) *Rate*, *Precision, Recall, Accuracy* and $F1$ were used:

$$TPR(Id) = \frac{CCP(Id)}{TNE(Id)} \cdot 100\% \tag{2}$$

where: *Id* – class identifier, *TPR* – percentage of true positives of class *Id*, *CCP* – Correctly Classified Positives of class *Id*, *TNE* – total number of elements in class *Id*, meaning that a class *Id* will be classified correctly with *TP* probability.

$$Precision(Id) = \frac{CCP(Id)}{TCP(Id)} \cdot 100\% \tag{3}$$

where: *Id, CCP* – as above, *Precision* – proportion of the examples which truly have class *Id* among all those which were classified as class *Id*, in [%], *TCP* – total number of objects classified as class *Id* (including *FP*(*Id*) – false positives), meaning that if an instance X is classified as an object in class *Id*, then with probability equal to *Precision* value it is truly class *Id*.

$$Recall(Id) = \frac{TPR(Id)}{TCN(Id)} \cdot 100\% \tag{4}$$

where: *Id*, *TPR*(*Id*) – as above, *Recall*(*Id*) – equivalent to true positive rate (or sensitivity), *TCN* – total number of objects classified as class *Id* (including *FN*(*Id*) – false negatives).

$$Accuracy(Id) = \frac{TP(Id) + TN(Id)}{TP(Id) + TN(Id) + FP(Id) + FN(Id)} \cdot 100\% \tag{5}$$

where: *Id*, *TP, FP, FN* – as above, *TN*(*Id*) – true negatives of class *Id*.

$$F1 = 2 \cdot \frac{Precision(Id) \cdot Recall(Id)}{Precision(Id) + Recall(Id)} \cdot 100\% \tag{6}$$

$F1$ – a combined measure for precision and recall (harmonic mean).

3.3 Feature vector optimization

Optimization of FVs is focused on selecting optimum parameters in the process of music genre classification. The first step is to reduce the original vector of parameters to eliminate strongly correlated parameters and replace them with one parameter, so the derived feature vector consists only of uncorrelated features. The second step is to add new parameters representing a specific instrument, typical for a given music genre.

Several optimization methods were performed using Weka implementation (Weka library), which resulted in a new vector of parameters (VoP). As mentioned before, VoP is understood here as the optimized FV.

4 Experiments

4.1 Reducing feature vector

4.1.1 Attribute subset selection

Among the methods of feature vector reducing one may discern those based on the attribute subset selection such as Best First, Greedy or Ranker. They were all tested on music excerpts extracted from the Synat database in the pre-study phase. For each experiment different settings were considered (direction of search, direction of search, number of non-improving nodes to consider before search termination). It occurred that the Best First (Weka library) method (based on 5-node test results for the Best First method in direction Forward) returned the best results for reducing FV of 173 parameters. The algorithm was implemented in the Weka environment, and the analysis searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controlling the level of the backtracking. Best First search uses the node depth as its cost.

As a result of such an optimization, a VoP_59 was obtained containing the following 59 descriptors: (VoP_59={TC,ASE1,ASE4,ASE5,ASE21,ASE23,ASE25–E29,ASEV1,ASEV29, ASE_MV,ASC,ASC_V,ASS_V,SFM1,SFMV1–3,SFMV5,SFMV6,SFMV8–12,SFMV14, MFCC2–7,MFCC9–13,MFCC17,MFCC20,MFCCV1–6,MFCCV8,MFCCV19,THR_2RMS_ TOT,THR_3RMS_TOT,THR_1RMS_10FR_MEAN,THR_2RMS_10FR_MEAN,THR_3RMS_ 10FR_MEAN,1RMS_TCD,2RMS_TCD,ZCD_10FR_VAR,1RMS_TCD_10FR_MEAN}).

The Principal Component Analysis (PCA) was also used to reduce data dimensionality. Resulted from PCA there were 74 new components that retain information contained previously in FV of 173 attributes. For the set of 74 components, a similar correctness to the Best First algorithm was obtained. Since these two data reduction approaches returned similar results, thus the smaller VoP_59 was employed for further analysis.

4.1.2 Adding parameters extracted from separated tracks

The importance of the parameters selected for the specific instrument may play a quite significant part in the classification process. Thus, the next series of experiments that involved adding parameters to the feature vector related to separated music tracks were performed. It was also checked that the mixture of more than three signals, in most cases, returned less promising results. Further analysis involved optimization of VoPs using the Best First method.

The optimization process was performed according to the scheme presented below:

a) The reduced VoP p_59 (as shown above in Section 4.1.1) was applied to each separated signal (the same attributes for original and separated signals). In that way new VoPs of 118 attributes were created for each mixture of two signals, and then subjected to the Best First method.

b) The Best First method was applied for each single FV (173 attributes) of the separated signal and added to VoP p_59 (original signal). As a result, new VoPs of different length (OH p_125, OD p_90, OP p_105, OT p_102, OS p_74) were created.
c) The Best First method was applied for the VoP of mixture of two FVs (173 (original) + 173 (separated)). This way the following VoPs were extracted: OH p_79, OD p_65, OP p_61, OT p_60, OS p_60.

Even though not every of the optimized VoPs as shown below gave the best overall correctness of classification, they were chosen for further experiments:

a) In the case of the **OH** signal the correctness of classification of the following genres: Classical, Pop, Latin Music and New Age, was taken into consideration with regard to the importance of Harmonic part for those genres. Based on those criteria, **VoP p_59** (as shown above) was chosen for the OH signal.
b) In the case of Drum mixture Rock, Hard Rock & Metal and Alternative Rock genres were taken into consideration. Based on those criteria, **VoP p_90** was chosen for the OD signal: (**VoP p_90**={**VoP_59**+drum_SC_v,drum_ASE1,7,16–19,29,drum_ASEv25,29, drum_ASE_MV,drum_SFM12–15,17,19,drum_SFMv18,drum_MFCCV1–4,drum_THR_3RMS_TOT,drum_THR_1RMS_10FR_MEAN,drum_THR_2RMS_10FR_MEAN,drum_THR_3RMS_10FR_MEAN,drum_THR_3RMS_10FR_VAR,drum_PEAK_RMS_TOT, drum_1RMS_TCD,drum_ZCD_10FR_VAR,drum_1RMS_TCD_10FR_VAR}).
c) In the case of **Piano** Classical, Blues, Jazz and New Age genres were taken into consideration. Based on those criteria, **VoP p_61** was chosen for the OP signal (OP VoP_61={TC,ASE1,4,5,21,23,25–29,ASEV1,29,ASE_MV, ASC,ASC_V,ASS_V,SFM1, SFMV1–5,8–12,14,MFCC2–7,9–13,17,20,MFCCV1–6,8,19, THR_2RMS_TOT,THR_3RMS_TOT,THR_1RMS_10FR_MEAN,THR_2RMS_10FR_MEAN,THR_3RMS_10FR_MEAN,1RMS_TCD, 2RMS_TCD,ZCD_10FR_VAR,1RMS_TCD_10FR_MEAN,piano_ASE_MV, piano_1RMS_TCD_10FR_MEAN}).
d) In the case of **Trumpet** and **Saxophone** Blues, Jazz and New Age genres were taken into consideration. Therefore, OT **VoP p_60** for OT and OS **VoP p_60** for OS were chosen, correspondingly. Their structures are as follows: OT VoP p_60={OP VoP_61+Trumpet_MFCC1} and OS VoP p_60={OP VoP_61+sax_MFCC2}.

4.1.3 Results and discussion

All results presented further on refer to the Co-SVM-based classification method. Cross-validation with 3-folds was used in this stage of experiments. Three iterations of cross-validation were performed. Then the individual performance values are aggregated, by calculating the mean over the three rounds. Tables 2 and 3 show confusion matrices along with the effectiveness measures, i.e. *Precision*, *Recall* (*True Positive Rate*), $F1$ and *Accuracy* obtained for VoP_59 for the original audio and for the OH (original + harmonic) mixture, correspondingly. As seen from Tables 2 and 3, there is a high degree of misclassifications between alternative rock and rock. It should be emphasized that a song in the Synat database was assigned to the particular music genre automatically by its label, however these two genres may be very similar both in perception and an automatic evaluation by a decision algorithm, hence difficult to be distinguished between each other.

Table 2 Confusion matrix for the Original signal based on VoP_p59 using Co-SVM classification method along with the effectiveness measures

O	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock	Sum	TPR [%]
Alternative Rock	**23.67**	1.33	0.00	2.33	1.67	2.33	0.33	0.67	1.67	11.33	1.67	0.33	21.67	69.00	34.30
Blues	3.67	**35.33**	0.00	14.00	0.00	0.00	7.33	3.67	1.00	6.67	0.67	7.33	8.33	88.00	40.15
Classical	1.33	0.33	**289.00**	0.67	0.00	0.00	4.00	0.33	18.00	2.67	0.00	0.00	1.33	317.67	90.98
Country	3.33	8.00	0.33	**269.33**	0.00	0.67	7.67	6.67	1.67	20.67	1.00	6.67	20.33	346.33	77.76
DanceDJ	4.33	0.33	0.00	0.00	**63.33**	1.00	0.00	0.67	1.00	3.67	9.33	2.00	0.67	86.33	73.36
Hard Rock & Metal	3.67	0.33	0.00	1.33	1.33	**176.67**	0.00	0.00	0.67	1.33	0.00	0.00	15.33	200.67	88.04
Jazz	1.00	2.67	5.00	8.00	0.33	0.00	**133.00**	2.00	15.00	14.33	0.00	6.33	1.67	189.33	70.25
Latin Music	0.33	3.67	0.00	13.67	0.33	0.00	4.33	**103.00**	0.00	13.33	5.33	3.33	0.67	148.00	69.60
New Age	1.33	0.67	24.00	0.33	2.67	2.67	10.67	0.00	**140.00**	2.00	0.33	2.67	1.67	189.00	74.07
Pop	6.67	10.33	8.00	34.00	4.33	5.33	16.00	11.33	13.00	**101.00**	9.33	16.00	30.00	265.33	38.07
Rap & Hip Hop	1.67	0.33	0.00	2.00	10.33	0.33	0.00	5.33	1.00	5.00	**303.67**	7.33	0.33	337.33	90.02
R&B	0.33	6.00	0.33	6.00	2.67	0.00	12.33	5.33	2.67	24.00	11.00	**128.00**	4.67	203.33	62.95
Rock	14.33	6.33	1.00	18.33	1.67	22.00	2.33	1.00	9.33	28.67	0.67	3.33	**198.67**	307.67	64.57
Precision [%]	36.05	46.70	88.20	72.79	71.43	83.73	67.18	73.57	68.29	43.04	88.53	69.82	65.07		
*F*1 [%]	35.16	43.18	89.57	75.20	72.38	85.83	68.68	71.53	71.06	40.40	89.27	66.21	64.82		
Accuracy [%]	96.92	96.73	97.61	93.93	98.27	97.92	95.77	97.10	96.02	90.22	97.41	95.46	92.72		

Bold data in diagonal indicates the number of correct answers

Table 3 Confusion matrix for the OH (VoP_p59) for the Co-SVM classification method along with the effectiveness measures

OH	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	Jazz	Latin Music	NewAge	Pop	Rap & Hip Hop	R&B	Rock	Sum	TPR [%]
Alternative Rock	**28**	1.33	0	2	2	2	0.33	0.33	0.67	10	2	0.33	20	68.99	40.59
Blues	2.33	**39.33**	0	11.67	0	0.33	6.67	2.67	0.67	8.67	1	6.67	8	88.01	44.69
Classical	0.67	0.67	**291.33**	0.67	0	0	4.67	0	13.33	5	0	0.33	1	317.67	91.71
Country	3	11.67	0.33	**271.67**	0.33	1	5	5	0	23.67	1.67	5.67	17.33	346.34	78.44
DanceDJ	4	0.67	0	0.33	**66**	1	0	0	0.33	3.67	8.33	1.33	0.67	86.33	76.45
Hard Rock & Metal	5	0.33	0	1.67	0.67	**170.67**	0	0	0.33	2	1.33	0	18.67	200.67	85.05
Jazz	0.33	6.33	7	4.67	0.67	0	**139.33**	3.33	7	11.67	0	7	2	189.33	73.59
Latin Music	0.33	4	0	11.67	0.67	0	3	**107**	0	13	3.67	3.67	1	148.01	72.29
New Age	1	0.67	18	1	4	2.67	4.67	0	**150.33**	2.67	0	1	3	189.01	79.54
Pop	9.67	9	8.33	32.67	3	5.33	17.33	10.33	7	**109.33**	7.33	15	31	265.32	41.21
Rap & Hip Hop	1.67	0.67	0	1.33	11	0.67	0.33	7	0	9	**293.33**	11.67	0.67	337.34	86.95
R&B	1	10	0	7	1.67	0	10.33	6	1.67	22.67	12.33	**126**	4.67	203.34	61.97
Rock	20	9	1	17.67	1.67	20	3.33	1	4.67	33	0.67	3.33	**192.33**	307.67	62.51
Precision [%]	36.36	41.99	89.37	74.63	71.99	83.80	71.45	75.00	80.82	42.98	88.44	69.23	64.04		
$F1$ [%]	38.36	43.30	90.52	76.49	74.15	84.42	72.51	73.62	80.17	42.08	87.69	65.40	63.27		
Accuracy [%]	96.83	96.39	97.83	94.27	98.35	97.76	96.30	97.29	97.37	90.13	97.09	95.37	92.48		

Bold data in diagonal indicates the number of correct answers

Overall, the improvement in classification results occurred for almost all music genres when expanding the original audio-based feature vector by parameters derived from harmonic, drum or a sum of harmonic and drum signals. For example, for the OH signal, Alternative Rock was less confused than in the case of the Original signal, Pop was less confused with New Age than for the original signal, Blues with Country, and Country with Rock, Latin with Country and with Jazz, Rock with New Age, etc. However, these results were statistically significant only in the case of: Alternative Rock, Blues, Classical, DanceDJ, Hard Rock & Metal, Latin Music, Rap & Hip Hop, R&B genres (see Table 8). Even though the difference in classification accuracy for Country, Jazz, New Age, Pop, Rock genres is still visible when using the expanded VoPs, the results obtained are not statistically significant.

The next step of experiments was to use different VoPs in the classification process depending on the type of music genre and taking into account pre-study classification results. Therefore, experiments were designed according to Section 4.1.2, i.e. selecting the most effective VoPs and the Co-SVM settings for the particular mixture of separated signals. In Tables 4, 5, 6 and 7 effectiveness measures, i.e. *Precision, Recall*, $F1$ and *Accuracy* obtained for OH, OD, OP, OT and OS signals are shown. It should be emphasized that both *Precision* and *Recall* measures have high values for most music classes except for Alt. Rock, Blues and Pop. Even though parameters selected for OH, OD, OP, OT and OS seem to return quite similar results, the list of attributes related to the original signal for these mixtures differs, this especially concerns OH and OD VoPs. It was also observed that for the VoPs of OP, OT and OS mixtures, the difference between those VoPs is only between the attributes related to the instrument part, i.e. two parameters for piano (piano_ASE_MV and piano_1rms_TCD_10FR_MEAN), one selected for trumpet (trum_mfcc1) and one for saxophone (sax_MFCC2).

To confirm the statistical significance of the results, T-Student test was carried out. The value of the T-student parameter above the 2.201 value indicates that the null hypothesis can be rejected, thus all the results with values above the threshold are statistically significant. Statistical significance threshold was set at 0.05. In Table 8 T-Student's test values are contained for all the combinations of VoPs. As observed from Table 8, the statistical analysis returned values above the threshold for: Alternative Rock, Blues, Classical, DanceDJ, Hard

Table 4 Effectiveness measures for the OD signal based on VoP_p90 using Co-SVM classification method [%]

OD p_90	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	
Precision	35.65	42.37	88.75	73.13	74.35	84.86	
Recall	39.61	42.05	91.08	77.29	77.21	86.55	
$F1$	37.53	42.21	89.90	75.15	75.75	85.70	
Accuracy	96.79	96.44	97.69	93.95	98.47	97.93	
OD p_90	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock
Precision	69.95	73.30	75.31	42.31	90.80	68.54	64.44
Recall	70.07	70.50	75.85	40.07	88.74	63.94	63.81
$F1$	70.01	71.87	75.58	41.16	89.76	66.16	64.12
Accuracy	96.03	97.11	96.74	90.04	97.57	95.38	92.60

Table 5 Effectiveness measures for the OP signal based on VoP_p61 using Co-SVM classification method [%]

OP p_61	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	
Precision	37.36	49.36	88.89	72.92	71.74	84.87	
Recall	32.85	43.56	91.50	77.48	76.45	88.54	
*F*1	34.96	46.28	90.18	75.13	74.02	86.67	
Accuracy	97.02	96.86	97.75	93.93	98.34	98.05	
OP p_61	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock
Precision	68.07	74.88	72.73	41.86	88.99	69.14	63.86
Recall	71.30	69.14	77.60	37.81	89.43	64.26	64.14
*F*1	69.65	71.90	75.09	39.74	89.21	66.61	64.00
Accuracy	95.89	97.17	96.58	90.03	97.41	95.45	92.53

Rock & Metal, Latin, Rap & Hip Hop and R&B genres. In the case of Pop and New Age using mixtures of OP and OT signals, correspondingly, makes the difference statistically significant.

The empirical study performed by the authors brought several findings:

– In most cases of the mixture of signals the improvement of the effectiveness measures was observed in comparison to the original signal.
– For each of the genres where Harmonic plays important part (Classical, Latin Music, New Age and Pop) the improvement of *TPR* values is observed for the OH signal. For the three of four selected genres (Classical vs. Latin Music and vs. New Age and vs. Pop), the improvement of *Precision* is also observed. In particular, an increase of 12.53 percent points in *Precision* for the New Age is achieved. Jazz genre deserves a special attention, in which case *Precision* was higher for over 4.27 percent points and *TPR* for over 3.34 percent points, as well as DanceDJ where *TPR* got over 3.1 percent points

Table 6 Effectiveness measures for the OT signal based on VoP_p60 using Co-SVM classification method [%]

OT p_60	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	
Precision	36.02	48.18	88.35	72.47	71.48	84.36	
Recall	32.37	40.15	90.66	78.06	76.44	86.88	
*F*1	34.10	43.80	89.49	75.16	73.88	85.60	
Accuracy	96.95	96.81	97.60	93.90	98.33	97.91	
OT p_60	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock
Precision	67.45	74.39	69.40	43.00	88.75	69.54	64.54
Recall	70.42	69.37	74.43	37.81	89.62	64.75	65.65
*F*1	68.91	71.79	71.83	40.24	89.18	67.06	65.09
Accuracy	95.80	97.15	96.14	90.22	97.40	95.50	92.69

Table 7 Effectiveness measures for the OS signal based on VoP_p60 using Co-SVM classification method [%]

OS p_60	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	
Precision	36.90	45.42	87.86	72.20	70.50	85.28	
Recall	33.33	41.29	91.08	77.77	75.67	88.54	
*F*1	35.03	43.26	89.44	74.88	72.99	86.88	
Accuracy	96.99	96.65	97.57	93.83	98.27	98.08	
OS p_60	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock
Precision	67.92	73.85	71.66	42.67	89.11	70.02	64.49
Recall	70.07	69.37	75.84	37.31	89.72	63.93	65.11
*F*1	68.98	71.54	73.69	39.81	89.41	66.84	64.80
Accuracy	95.84	97.11	96.41	90.18	97.46	95.52	92.66

higher. The improvement of Jazz should be stressed out especially in the context of the lower rate of misclassification between Pop and Jazz. It was also shown that for genres such as Rock and Hard Rock & Metal, the decrease of correctness for OH was observed, what confirms that the harmonic part does not play an important part for those genres. Surprisingly, Alternative Rock got over 6 percent points of improvement of the *TPR*. The behavior of Blues is also interesting, where the *TPR* was also improved for over 4.5 percent points, while *Precision* decreased by almost 5 percent points.

– The improvement of *Recall* (*TPR*) value for Alternative Rock was gained in the case of the OD signal. Surprisingly, higher *Precision* values of New Age (over 7 percent points)

Table 8 Student's T values under the null hypothesis for independent samples (statistical significance threshold set at 0.05)

2.201	Alternative Rock	Blues	Classical	Country	DanceDJ	Hard Rock & Metal	
O–OH p_59	8.791	3.392	3.415	1.709	4.101	3.808	
O–OD p_90	7.575	3.785	3.501	1.732	4.071	3.72	
O–OP p_61	7.477	3.782	3.456	1.748	4.053	3.764	
O–OT p_60	6.088	4.082	3.564	1.769	4.096	3.825	
O–OS p_60	6.735	4.01	3.538	1.774	4.187	3.784	
2.201	Jazz	Latin Music	New Age	Pop	Rap & Hip Hop	R&B	Rock
O–OH p_59	1.988	3.531	1.96	1.848	3.264	3.631	2.023
O–OD p_90	2.081	3.689	2.111	1.901	3.197	3.705	2.046
O–OP p_61	2.062	3.731	2.122	2.305	3.235	3.612	2.062
O–OT p_60	2.065	3.667	2.217	1.964	3.193	3.798	2.015
O–OS p_60	2.078	3.606	2.166	2.055	3.204	3.754	2.014

Values greater than $t > 2.201$ are statistically significant

and *Recall* (*TPR*) of Dance & DJ (almost 4 percent points) were also gained. In the case of classes such as Latin Music, New Age, Pop and Classical, a slight improvement of classification was also observed. That proves that the lack of Drum element (the percussion signal was present only for 89.6% of elements from the input audio dataset) is a piece of information/feature for the classifier with the significance in the training process.

– The improvement of classification for the genres where the piano plays an important part was not so visible for the OP signal. *Precision* was improved in the case of several genres (e.g. Classical, Dance & DJ, Jazz, Latin, but also Blues, Alternative Rock, etc.), along with *Recall* (*TPR*) values. Improvement of over 3 percent points for the DanceDJ genre was obtained.
– A slight improvement of *Precision* is to be observed, i.e. for the OT – Hard Rock & Metal, Latin, New Age, and in the case of OS – New Age, Rap & Hip Hop and R&B. This is also visible for *Recall* (*TPR*) values (e.g. DanceDJ, R&B).

5 Summary

The article focuses on automatic music genre classification while using the original and separated tracks. The instrument separation approach was selected to improve the results of music genre classification, and in particular to decrease the misclassification between selected genres in the context of the influence of the specific instrument on selected genres. For that case, a Non-Negative Matrix Factorization (NMF) method was adapted from the literature, and a new way of using the separated and original signals for parameterization was proposed. Since other researchers (Lampropoulos et al. 2005; Rump et al. 2010) applied just one specific single signal in the process of music classification, which did not result in high accuracy, the authors' approach was based on creating a new VoP, i.e. extending the original FV by new attributes representing a specific instrument. In that way five different separated audio signals were obtained: harmonic, drum, piano, trumpet and saxophone. It should also be noted that this is a multi-instrument separation process, as the "drum" signal consists of a few instruments: snare drum, bass drum, tom-tom, timpani, crash cymbals, etc. With regard to the piano, we have to keep in mind that classical piano has quite different kind of sound than e.g. a jazzy piano. VoP consisted of an original audio and separated piano (OP) did not improve the results for classical music but decreased the misclassification of Jazz.

In the analysis performed, the overall correctness of classification was higher in almost each case of the mixed VoP in comparison to the Original signal. Also, it was observed that the specific mix of signals improved the correctness of classification of genres where this signal played an important part. This means that for genres where harmonic instruments play an important part, e.g. New Age, Pop, Latin Music, the correctness of classification increased. The same tendency was observed for other mixed VoPs: OD signal for Alternative Rock, Hard Rock & Metal, as well as DanceDj, and New Age. In the case of the OP signal, the improvement in classification of Blues, Classical and New Age was also visible. Overall, a decrease in misclassification between the similar, as well as opposite genres was obtained.

In the process of the analysis over 8.000 music tracks, representing 13 music genres, were extracted from the Synat database. Although many research works were published in the area of music genre classification, most of them, with some exceptions, analyze only a few genres represented by $\sim$ 1.000 songs in total. The results shows that the overall classification obtained by the authors reaches $\sim$ 72%, what is $\sim$ 10 percent points better

in comparison to the results shown in the literature, i.e. $\sim$ 60% for 10 musical genres (Tzanetakis et al. 2002) and $\sim$ 57.8% for 13 music genres (Burred 2014).

References

AudioScore Ultimate 7. (2016). http://www.sibelius.com/products/audioscore/ultimate.html.

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kegl, B. (2006). Aggregate features and AdaBoost for music classification. *Machine Learning*, *65*(2/3), 473–484. doi:10.1007/s10994-006-9019-7.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory,* (pp. 92–100). Morgan Kaufmann.

Bregman, A. (1990). Auditory scene analysis: the perceptual organization of sound. MIT Press.

Burgoyne, J., Fujinaga, I., & Downie, J.S. (2016). Music information retrieval. In Schreibman, S., Siemens, R., & Unsworth, J. (Eds.) *A new companion to digital humanities,* Chapter 15. 1st Ed. New York: Wiley.

Burred, J.J. (2014). Hierarchical approach to automatic musical genre classification. *Journal of the Audio Engineering Society*, *52*(7/8), 724–739.

Casey, M., & Westner, A. (2000). Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference (ICMA),* (pp. 154–161). Berlin.

de Cheveigne, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 3271–3290.

Dziubiński, M., Dalka, P., & Kostek, B. (2005). Estimation of musical sound separation algorithm effectiveness employing neural networks. *Journal of Intelligent Information Systems*, *24*(2), 133–157.

Eweret, S., Prado, B., Muller, M., & Plumbley, M. (2014). Score-Informed Source separation for musical audio recordings. *Signal Processing Magazine*, *31*(3), 116–124.

Fu, Z., Lu, G., Ting, K.M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, *13.2*, 303–319.

Gerber, T., Dutasta, M., Girin, L., & Fevotte, C. (2012). Professionally-Produced Music Separation Guided by Covers. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR),* (pp. 85–90). Portox.

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.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music database: popular, classical, and jazz music databases. In *Proceedings of ISMIR*, (Vol. 2002 pp. 287–288).

Helen, M., & Virtanen, T. (2005). Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO),* Antalaya.

Herrera-Boyer, P., Gouyon, F., & et al. (2013). MIRrors: music information research reflects on its future, *J. Intel. Infor. Systems*, 41 (3).

Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2000). Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR),* Plymouth.

Hill, A.J., & Hawksford, M.O.J. (2010). A hybrid virtual bass system for optimized steady state and transient performance. CEEC Conf. 8–9.09.

Hoffmann, P., & Kostek, B. (2015). Bass enhancement settings in portable devices based on music genre recognition. *Journal of the Audio Engineering Society*, *12*(63), 980–989. doi:10.17743/jaes.2015.0087.

Holzapfel, A., & Stylianou, Y. (2008). Musical genre classification using nonnegative matrix factorization-based features. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 424–434. doi:10.1109/TASL.2007.909434.

Hsu, C.-W., Chang C.-C., & Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. Technical report. Department of Computer Science, National Taiwan University, Taipei 106, 2003 Taiwan (retrieved from http://www.csie.ntu.edu.tw/~cjlin2016).

Humprey, E.J., Bello, J., & LeCun, Y. (2013). Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information System*, *41*(3), 461–481.

Hyoung-Gook, K., Moreau, N., & Sikora, T. (2005). MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. Wiley Sons.

ISMIR (2016). International conference on music information retrieval, Malaga, Spain. http://ismir2015.ismir.net.

Kalliris, G.M., Dimoulas, C.A., & Uhle, C. (2016). Guest Editors' Note, Special Issue on Intelligent Audio Processing, Semantics, and Interaction. *Journal of the Audio Engineering Society*, *64*(7/8), 464–465.

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*. New York: Physica Verlag.

Kostek, B. (2005). Perception-Based Data processing in acoustics. Applications to music information retrieval and psychophysiology of hearing. In *Series on cognitive technologies,* Berlin: Springer.

Kostek, B. (2013). A kaczmarek music recommendation based on multidimensional description and similarity measures. *Fundamenta Informaticae*, *127*(1–4), 325–340. doi:10.3233/FI-2013-912.

Kostek, B., Kupryjanow, A., Zwan, P., Jiang, W., Ras, Z., Wojnarski, M., & Swietlicka, J. (2011). In Kryszkiewicz, M., et al. (Eds.) *Report of the ISMIS 2011 contest: Music information retrieval, foundations of intelligent systems. ISMIS 2011, LNAI 6804*, (pp. 715–724). Berlin: Springer.

Kostek, B., Hoffmann, P., Kaczmarek, A., & Spaleniak, P. (2014). Creating a Reliable Music Discovery and Recommendation System. In Bembenik, R., et al. (Eds.) *Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation, Studies in Computational Intelligence 541,* (pp. 107–130). Switzerland: Springer Intern. Publishing. doi:10.1007/978-3-319-04714-07.

Kotsakis, R., Kalliris, G., & Dimoulas, C. (2012). Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. *Speech Communication*, *54.6*, 743–762.

Lampropoulos, A., Lampropoulou, P., & Tsihrintzis, G. (2005). Musical Genre Classification Enhanced by Improved Source Separation Techniques. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR),* (pp. 576–581). London.

Lee, D.D., & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature – International Weekly Journal of Science*, 788–791.

Lee, J.H., & Cunningham, S.J. (2013). Toward an understanding of the history and impact of user studies in music information retrieval. *Journal of Intelligent Information System*, *41*(3), 499–521.

Lee, K., & Ellis Daniel, P.W. (2010). Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, *18.6*, 1406–1416.

Lindsay, A., & Herre, J. (2001). MPEG-7 and MPEG-7 Audio – An Overview. *Journal of the Audio Engineering Society*, *49*(7/8), 589–594.

Liu, Y., Xu, J., Wei, L., & Tian, Y. (2007). The study of the classification of Chinese folk songs by regional style. In *Proceedings of the International Conference on Semantic Computing (ICSC),* (pp. 657–662). IEEE.

McKay, C., & Fujinaga, I. (2004). Automatic genre classification using large High-Level musical feature sets. In *Proceedings of the International Conference on Music Information Retrieval,* (pp. 525–30).

Mu, H., & Gan, W.-S. (2012). A psychoacoustic bass enhancement system with improved transient and steady-state performance. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP,* (pp. 141–144), Kyoto. doi:10.1109/ICASSP.2012.6287837.

Mu, H., & Gan, W.-S. (2015). *Journal of the Audio Engineering Society*, *63*(11), 900–913. doi:10.17743/jaes.2015.0079.

Nayak, S., & Bhutani, A. (2011). Music genre classification using GA-induced minimal feature-set. In *3rd National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, (pp. 33–36). Hubli, Karnataka.

Ness, S., Theocharis, A., Tzanetakis, G., & Martins L.G. (2009). Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In *17 ACM International Conference on Multimedia*, New York.

Ntalampiras, S. (2013). A novel holistic modeling approach for generalized sound recognition. *IEEE Signal Processing Letters*, *20*(2), 185–188. doi:10.1109/LSP.2013.2237902.

Oo, N., Gan, W.-S., & Hawksford, M.O.J. (2000). Perceptually-motivated objective grading of nonlinear processing in virtual-bass systems. *Journal of the Audio Engineering Society*, *59*(11), 804–824.

Paulus, J., & Virtanen, T. (2005). Drum transcription with non-negative spectrogram factorisation. In *Proceedings of 13th European Signal Processing Conference (EUSIPCO), 1*, (pp. 1059–1062), Antalya: Curran Associates, Inc.

Pérez-García, T., Pérez-Sancho, C., & Iñesta, J.M. (2010). Harmonic and instrumental information fusion for musical genre classification. In *MML'10 Proceedings of 3rd International Workshop on Machine Learning and Music*. New York: ACM. doi:10.1145/1878003.1878020.

Plewa, M., & Kostek, B. (2015). Music mood visualization using self-organizing maps. *Archives of Acoustics*, *40*(4), 513–525. doi:10.1515/aoa-2015-0051.

Qt (2016). cross-platform application framework; https://www.qt.io/.

Romain, O., Tietche, B.H., Denby, B., Dieuleveult, F., Granado, B., Kemiri, H., Chollet, G., & Blouet, R. (2012). Prototype of a Radio-On-Demand Broadcast Receiver with real time musical genre classification. In *Conference on Design and Architectures for Signal and Image Processing (DASIP 2012)*, (pp. 23–25). Germany: Karlsruhe.

Rosner, A., & Kostek, B. (2015). Musical instrument separation applied to music genre classification. In *International Symposium on Methodologies for Intelligent Systems, (ISMIS)*, Springer.

Rosner, A., Weninger, F., Schuller, B., Michalak, M., & Kostek, B. (2013). Influence of Low-Level Features Extracted from Rhythmic and Harmonic Sections on Music Genre Classification. In Gruca, A., Czachórski, T., & Kozielski, S. (Eds.) *Man-Machine Interactions 3, Proceedings of International Conference on Man-Machine Interactions (ICMII). 242,* (pp. 467–473). Beskidy: Springer.

Rosner, A., Schuller, B., & Kostek, B. (2014). Classification of music genres based on music separation into harmonic and drum components. *Archives of Acoustics*, *39*(4), 629–638. doi:10.2478/aoa-2014-0068.

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

Sampleswap (2016). http://sampleswap.org/.

Salamon, J., Rocha, B., & Gomez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *IEEE international conference on acoustics speech and signal processing, ICASSP*, Kyoto.

Schuller, B., Lehmann, A., Weninger, F., Eyben, F., & Rigoll, G. (2009). Blind Enhancement of the Rhythmic and Harmonic Sections by NMF: Does it help? In *Proceedings International Conference on Acoustics including the 35th German Annual Conference on Acoustics, (NAG/DAGA)*, Rotterdam. The Netherlands.

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

Schedl, M., Gómez, E., & Urba, J. (2014). Music Information Retrieval: Recent developments and applications. *Foundations and Trends R in Information Retrieval*, *8*(2-3), 127–261. http://dx.doi.org/978-1-60198-807-2.

Silla, C.N., Kaestner, C.A., & Koerich, A.L. (2007). Automatic Music Genre Classification Using Ensemble of Classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, (pp. 1687–1692). Montreal. doi:10.1007/BF03192561.

Smaragdis, P., & Brown, J.C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of WASPAA*, (pp. 177–180).

Sturm, B.L. (2013). Classification accuracy is not enough. on the evaluation of music genre recognition systems. *Journal of Intelligent Information Systems*, *41*(3), 371–406. doi:10.1007/s10844-013-0250-y.

Sturm, B.L. (2014). A survey of evaluation in music genre recognition. In Nurnberger, A., Stober, S., Larsen, B., & Detyniecki, M. (Eds.) *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation* (pp. 29–66). LNCS 8382.

SYNAT (2016). https://synat.eti.pg.gda.pl/.

Tekman, H.G., & Hortacsu, N. (2002). Aspects of stylistic knowledge: what are different styles like and why do we listen to them? *Psychology of Music*, *30*(1), 28–47.

Tzanetakis, G., Essl, G., & Cook, P. (2002). Automatic musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302.

Uhle, C., Dittmar, C., & Sporer, T. (2003). Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, (pp. 843–848). Nara.

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), 85–104.

Weka library (2016). http://sourceforge.net/projects/weka/files/weka-3-7/3.7.5/.

Weninger, F., Lehmann, A., & Schuller, B. (2011). openbliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Prague: IEEE.

Weninger, F., Lehmann, A., & Schuller, B. (2017). OpenBliSSART, http://openblissart.github.io/openBliSSART/.

Xiaojin, Z., & Goldberg, A.B. (2009). Introduction to Semi-supervised Learning. In Brachman, R.J., & Dietterich, T.G. (Eds.) *Synthesis Lectures on artificial Intelligence ad Machine Learning*: Morgan & Claypool Publishers.

Zhu, J., Xue, X., & Lu, H. (2004). Musical genre classification by instrumental features. In *Proceedings of the ICMC*.