Recognition of Instrument Timbres in Real Polytimbral Audio Recordings

Elżbieta Kubera[1,2], Alicja Wieczorkowska[2],
Zbigniew Raś[2,3], and Magdalena Skrzypiec[4]

[1] University of Life Sciences in Lublin, Akademicka 13, 20-950 Lublin, Poland
[2] Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland
[3] University of North Carolina, Dept. of Computer Science,
Charlotte, NC 28223, USA
[4] Maria Curie-Skłodowska University in Lublin,
Pl. Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland
elzbieta.kubera@up.lublin.pl,
alicja@poljap.edu.pl,
ras@uncc.edu,
mskrzypiec@hektor.umcs.lublin.pl

Abstract. Automatic recognition of multiple musical instruments in polyphonic and polytimbral music is a difficult task, but often attempted to perform by MIR researchers recently. In papers published so far, the proposed systems were validated mainly on audio data obtained through mixing of isolated sounds of musical instruments. This paper tests recognition of instruments in real recordings, using a recognition system which has multilabel and hierarchical structure. Random forest classifiers were applied to build the system. Evaluation of our model was performed on audio recordings of classical music. The obtained results are shown and discussed in the paper.

Keywords: Music Information Retrieval, Random Forest.

1 Introduction

Music Information Retrieval (MIR) gains increasing interest last years [24]. MIR is multi-disciplinary research on retrieving information from music, involving efforts of numerous researchers – scientists from traditional, music and digital libraries, information science, computer science, law, business, engineering, musicology, cognitive psychology and education [4], [33]. Topics covered in MIR research include [33]: auditory scene analysis, aiming at the recognition of e.g. outside and inside environments, like streets, restaurants, offices, homes, cars etc. [23]; music genre categorization – an automatic classification of music into various genres [7], [20]; rhythm and tempo extraction [5]; pitch tracking for query-by-humming systems that allows automatic searching of melodic databases using

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part II, LNAI 6322, pp. 97–110, 2010.

sung queries [1]; and many other topics. Research groups design various intelligent MIR systems and frameworks for research, allowing extensive works on audio data, see e.g. [20], [29].

Huge repositories of audio recordings available from the Internet and private sets offer plethora of options for potential listeners. The listeners might be interested in finding particular titles, but they can also wish to find pieces they are unable to name. For example, the user might be in mood to listen to something joyful, romantic, or nostalgic; he or she may want to find a tune sung to the computer's microphone; also, the user might be in mood to listen to jazz with solo trumpet, or classic music with sweet violin sound. More advanced person (a musician) might need scores for the piece of music found in the Internet, to play it by himself or herself.

All these issues are of interest for researchers working in MIR domain, since meta-information enclosed in audio files lacks such data – usually recordings are labeled by title and performer, maybe category and playing time. However, automatic categorization of music pieces is still one of more often performed tasks, since the user may need more information than it is already provided, i.e. more detailed or different categorization. Automatic extraction of melody or possibly the full score is another aim of MIR. Pitch-tracking techniques yield quite good results for monophonic data, but extraction of polyphonic data is much more complicated. When multiple instruments play, information about timbre may help to separate melodic lines for automatic transcription of music [15] (spatial information might also be used here). Automatic recognition of timbre, i.e. of instrument, playing in polyphonic and polytimbral (multi-instrumental) audio recordings, is our goal in the investigations presented in this paper.

One of the main problems when working with audio recordings is labeling of the data, since without properly labeled data, testing is impossible. It is difficult to recognize all notes played by all instruments in each recording, and if numerous instruments are playing, this task is becoming infeasible. Even if a score is available for a given piece of music, still, the real performance actually differs from the score because of human interpretation, imperfections of tempo, minor mistakes, and so on. Soft and short notes pose further difficulties, since they might not be heard, and grace notes leave some freedom to the performer - therefore, consecutive onsets may not correspond to consecutive notes in the score. As a result, some notes can be omitted. The problem of score following is addressed in [28].

1.1 Automatic Identification of Musical Instruments in Sound Recordings

The research on automatic identification of instruments in audio data is not a new topic; it started years ago, at first on isolated monophonic (monotimbral) sounds. Classification techniques applied quite successfully for this purpose by many researchers include k-nearest neighbors, artificial neural networks, rough-set based classifiers, support vector machines (SVM) – a survey of this research is presented in [9]. Next, automatic recognition of instruments in audio data

was performed on polyphonic polytimbral data, see e.g. [3], [12], [13], [14], [19], [30], [32], [35], also including investigations on separation of the sounds from the audio sources (see e.g. [8]).

The comparison of results of the research on automatic recognition of instruments in audio data is not so straightforward, because various scientists utilized different data sets: of different number of classes (instruments and/or articulation), different number of objects/sounds in each class, and basically different feature sets, so the results are quite difficult to compare. Obviously, the less classes (instruments) to recognize, the higher recognition rate was achieved, and identification in monophonic recordings, especially for isolated sounds, is easier than in polyphonic polytimbral environment. The recognition of instruments in monophonic recordings can reach 100% for a small number of classes, more than 90% if the instrument or articulation family is identified, or about 70% or less for recognition of an instrument when there are more classes to recognize. The identification of instruments in polytimbral environment is usually lower, especially for lower levels of the target sounds – even below 50% for same-pitch sounds and if more than one instrument is to be identified in a chord; more details can be found in the papers describing our previous work [16], [31]. However, this research was performed on sound mixes (created by automatic mixing of isolated sounds), mainly to make proper labeling of data easier.

2 Audio Data

In our previous research [17], we performed experiments using isolated sounds of musical instruments and mixes calculated from these sounds, with one of the sounds being of higher level than the others in the mix, so our goal was to recognize the dominating instrument in the mix. The obtained results for 14 instruments and one octave shown low classification error, depending on the level of sounds added to the main sound in the mix - the highest error was 10% for the level of accompanying sound equal to 50% of the level of the main sound. These results were obtained for random forest classifiers, thus proving usefulness of this methodology for the purpose of the recognition of the dominating instrument in polytimbral data, at least in case of mixes. Therefore, we applied the random forest technique for the recognition of plural (2–5) instruments in artificial mixes [16]. In this case we obtained lower accuracy, also depending of the level of the sounds used, and varying between 80% and 83% in total, and between 74% and 87% for individual instruments; some instruments were easier to recognize, and some were more difficult.

The ultimate goal of such work is to recognize instruments (as many as possible) in real audio recordings. This is why we decided to perform experiments on the recognition of instruments with tests on real polyphonic recordings as well.

2.1 Parameterization

Since audio data represent sequences of amplitude values of the recorded sound wave, such data are not really suitable for direct classification, and

parameterization is performed as a preprocessing. An interesting example of a framework for modular sound parameterization and classification is given in [20], where collaborative scheme is used for feature extraction from distributed data sets, and further for audio data classification in a peer-to-peer setting.

The method of parameterization influences final classification results, and many parameterization techniques have been applied so far in research on automatic timbre classification. Parameterization is usually based on outcomes of sound analysis, such us Fourier transform, wavelet transform, or time-domain based description of sound amplitude or spectrum. There is no standard set of parameters, but low-level audio descriptors from the MPEG-7 standard of multimedia content description [11] are quite often used as a basis of musical instrument recognition. Since we have already performed similar research, we decided to use MPEG-7 based sound parameters, as well as additional ones.

In the experiments described in this paper, we used 2 sets of parameters: average values of sound parameters calculated through the entire sound (being a single sound or a chord), and temporal parameters, describing evolution of the same parameters in time. The following parameters were used for this purpose [35]:

– MPEG-7 audio descriptors [11], [31]:
  - *AudioSpectrumCentroid* - power weighted average of the frequency bins in the power spectrum of all the frames in a sound segment;
  - *AudioSpectrumSpread* - a RMS value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame;
  - *AudioSpectrumFlatness*, $flat_1, \ldots, flat_{25}$ - multidimensional parameter describing the flatness property of the power spectrum within a frequency bin for selected bins; 25 out of 32 frequency bands were used for a given frame;
  - $HarmonicSpectralCentroid$ - the mean of the harmonic peaks of the spectrum, weighted by the amplitude in linear scale;
  - $HarmonicSpectralSpread$ - represents the standard deviation of the harmonic peaks of the spectrum with respect to the harmonic spectral centroid, weighted by the amplitude;
  - $HarmonicSpectralVariation$ - the normalized correlation between amplitudes of harmonic peaks of each 2 adjacent frames;
  - $HarmonicSpectralDeviation$ - represents the spectral deviation of the log amplitude components from a global spectral envelope;

– other audio descriptors:
  - *Energy* - energy of spectrum in the parameterized sound;
  - MFCC - vector of 13 Mel frequency cepstral coefficients, describe the spectrum according to the human perception system in the mel scale [21];
  - *ZeroCrossingDensity* - zero-crossing rate, where zero-crossing is a point where the sign of time-domain representation of sound wave changes;

- $FundamentalFrequency$ - maximum likelihood algorithm was applied for pitch estimation [36];
- $NonMPEG7 - AudioSpectrumCentroid$ - a differently calculated version - in linear scale;
- $NonMPEG7 - AudioSpectrumSpread$ - different version;
- $RollOff$ - the frequency below which an experimentally chosen percentage equal to 85% of the accumulated magnitudes of the spectrum is concentrated. It is a measure of spectral shape, used in speech recognition to distinguish between voiced and unvoiced speech;
- $Flux$ - the difference between the magnitude of the DFT points in a given frame and its successive frame. This value was multiplied by 10^7 to comply with the requirements of the classifier applied in our research;
- $FundamentalFrequency'sAmplitude$ - the amplitude value for the predominant (in a chord or mix) fundamental frequency in a harmonic spectrum, over whole sound sample. Most frequent fundamental frequency over all frames is taken into consideration;
- $Ratio\ r_1, \ldots, r_{11}$ - parameters describing various ratios of harmonic partials in the spectrum;
  - r_1: energy of the fundamental to the total energy of all harmonic partials,
  - r_2: amplitude difference [dB] between 1^{st} partial (i.e., the fundamental) and 2^{nd} partial,
  - r_3: ratio of the sum of energy of 3^{rd} and 4^{th} partial to the total energy of harmonic partials,
  - r_4: ratio of the sum of partials no. 5-7 to all harmonic partials,
  - r_5: ratio of the sum of partials no. 8-10 to all harmonic partials,
  - r_6: ratio of the remaining partials to all harmonic partials,
  - r_7: brightness - gravity center of spectrum,
  - r_8: contents of even partials in spectrum,

$$r_8 = \frac{\sqrt{\sum_{k=1}^{M} A_{2k}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}}$$

    where A_n - amplitude of n^{th} harmonic partial,
    N - number of harmonic partials in the spectrum,
    M - number of even harmonic partials in the spectrum,
  - r_9: contents of odd partials (without fundamental) in spectrum,

$$r_9 = \frac{\sqrt{\sum_{k=2}^{L} A_{2k-1}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}}$$

    where L – number of odd harmonic partials in the spectrum,
  - r_{10}: mean frequency deviation for partials 1-5 (when they exist),

$$r_{10} = \frac{\sum_{k=1}^{N} A_k \cdot |f_k - kf_1| / (kf_1)}{N}$$

where $N = 5$, or equals to the number of the last available harmonic partial in the spectrum, if it is less than 5,

* r_{11}: partial (i=1,...,5) of the highest frequency deviation.

Detailed description of popular features can be found in the literature; therefore, equations were given only for less commonly used features.

These parameters were calculated using fast Fourier transform, with 75 ms analyzing frame and Hamming window (hop size 15 ms). Such a frame is long enough to analyze the lowest pitch sounds of our instruments and yield quite good resolution of spectrum; since the frame should not be too long because the signal may then undergo changes, we believe that this length is good enough to capture spectral features and changes of these features in time, to be represented by temporal parameters. Our descriptors describe the entire sound, constituting one sound event, being a single note or a chord.

The sound timbre is believed to depend not only on the contents of sound spectrum (depending on the shape of the sound wave), but also on changes of spectrum (and the shape of the sound wave) over time. Therefore, the use of temporal sound descriptors was also investigated - we would like to check whether adding of such (even simple) descriptors will improve the accuracy of classification. The temporal parameters in our research were calculated in the following way. Temporal parameters describe temporal evolution of each original feature vector p, calculated as presented above. We were treating p as a function of time and searching for 3 maximal peaks. Maximum is described by k - the consecutive number of frame where the maximum appeared, and the value of this parameter in the frame k:

$M_i(p) = (k_i, p[k_i]), \quad i = 1, 2, 3 \quad k_1 < k_2 < k_3$

The temporal variation of each feature can be then presented by a vector T of new temporal parameters, built as follows:

$T_1 = k_2 - k_1$

$T_2 = k_3 - k_2$

$T_3 = k_3 - k_1$

$T_4 = p[k_2]/p[k_1]$

$T_5 = p[k_3]/p[k_2]$

$T_6 = p[k_3]/p[k_1]$

Altogether, we obtained a feature vector of 63 averaged descriptors, and another vector of $63 \cdot 6 = 378$ temporal descriptors for each sound object. We made a comparison of performance of classifiers built using only 63 averaged parameters and built using both averaged and temporal features.

2.2 Training and Testing Data

Our training and testing data were based on audio samples of the following 10 instruments: B-flat clarinet, cello, double bass, flute, French horn, oboe, piano, tenor trombone, viola, and violin. Full musical scale of these instruments was used for both training and testing purposes. Training data were taken from

Table 1. Number of pieces in RWC Classical Music Database with the selected instruments playing together

	clarinet	cello	dBass	flute	fHorn	piano	trbone	viola	violin	oboe
clarinet	0	8	7	5	6	1	3	8	8	5
cello	8	0	13	9	9	4	3	17	20	8
doublebass	7	13	0	9	9	2	3	13	13	8
flute	5	9	9	1	7	1	2	9	9	6
frenchhorn	6	9	9	7	3	4	4	9	11	8
piano	1	4	2	1	4	0	0	2	9	0
trombone	3	3	3	2	4	0	0	3	3	3
viola	8	17	13	9	9	2	3	0	17	8
violin	8	20	13	9	11	9	3	17	18	8
oboe	5	8	8	6	8	0	3	8	8	2

MUMS – McGill University Master Samples CDs [22] and The University of IOWA Musical Instrument Samples [26]. Both isolated single sounds and artificially generated mixes were used as training data. The mixes were generated using 3 sounds. Pitches of composing sounds were chosen in such a way that the mix constitutes a minor or major chord, or its part (2 different pitches), or even a unison. The probability of choosing instruments is based on statistics drawn from RWC Classical Music Database [6], describing in how many pieces these instruments play together in the recordings (see Table 1). The mixes were created in such a way that for a given sound, chosen as the first one, two other sounds were chosen. These two other sounds represent two different instruments, but one of them can also represent the instrument selected as the first sound. Therefore, the mixes of 3 sounds may represent only 2 instruments.

Since testing was already performed on mixes in our previous works, the results reported here describe tests on real recordings only, not based on sounds from the training set. Test data were taken from RWC Classical Music Database [6]. Sounds of length of at least 150 ms were used. For our tests we selected available sounds representing the 10 instruments used in training, playing in chords of at least 2 and no more than 6 instruments. The sound segments were manually selected and labeled (also comparing with available MIDI data) in order to prepare ground-truth information for testing.

3 Classification Methodology

So far, we applied various classifiers for the instrument identification purposes, including support vector machines (SVM, see e.g. [10]) and random forests (RF, [2]). The results obtained using RF for identification of instruments in mixes outperformed the results obtained via SVM by an order of magnitude. Therefore, the classification performed in the reported experiments was based on RF technique, using WEKA package [27].

Random forest is an ensemble of decision trees. The classifier is constructed using procedure minimizing bias and correlations between individual trees,

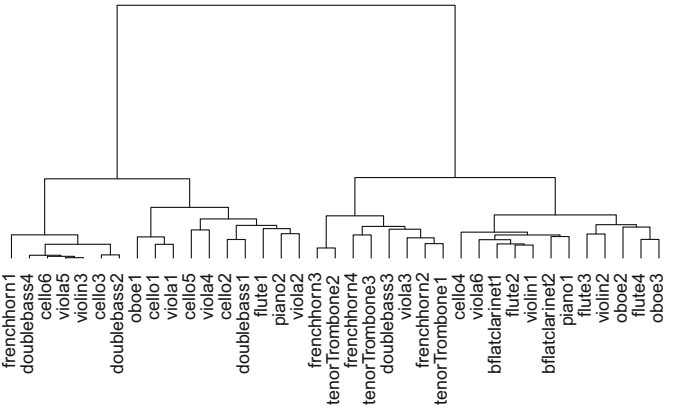

Fig. 1. Hierarchical classification of musical instrument sounds for the 10 investigated instruments

according to the following procedure [17]. Each tree is built using different N-element bootstrap sample of the training N-element set; the elements of the sample are drawn with replacement from the original set. At each stage of tree building, i.e. for each node of any particular tree in the random forest, p attributes out of all P attributes are randomly chosen ($p \ll P$, often $p = \sqrt{P}$). The best split on these p attributes is used to split the data in the node. Each tree is grown to the largest extent possible - no pruning is applied. By repeating this randomized procedure M times one obtains a collection of M trees – a random forest. Classification of each object is made by simple voting of all trees.

Because of similarities between timbres of musical instruments, both from psychoacoustic and sound-analysis point of view, hierarchical clustering of instrument sounds was performed using R – an environment for statistical computing [25]. Each cluster in the obtained tree represents sounds of one instrument (see Figure 1). More than one cluster may be obtained for each instrument; sounds representing similar pitch usually are placed in one cluster, so various pitch ranges are basically assigned to different clusters. To each leaf a classifier is assigned, trained to identify a given instrument. When the threshold of 50% is exceeded for this particular classifier alone, the corresponding instrument is identified.

We also performed node-based classification in additional experiments, i.e. when any node exceeded the threshold, but no its children did, then the instruments represented in this node were returned as a result. The instruments from this node can be considered similar, and they give a general idea on what sort of timbre was recognized in the investigated chord.

Data cleaning. When this tree was built, pruning was performed and the leaves representing less than 5% of sounds of a given instruments were removed, and these sounds were removed from the training set. As a result, the training data

in case of 63-element feature vector consisted of 1570 isolated single sounds, and the same number of mixes. For the extended feature vector (with temporal parameters added), 1551 isolated sounds and the same number of mixes was used. The difference in number is caused by different pruning for the different hierarchical classification tree, built for the extended feature vector. Testing data set included 100 chords.

Since we are recognizing instruments in chords, we are dealing with multi-label data. The use of multi-label data makes reporting of results more complicated, and the results depend on the way of counting the number of correctly identified instruments, omissions and false recognitions [18], [34]. We are aware of influence of these factors on the precision and recall of the performed classification. Therefore, we think the best way to present the results is to show average values of precision and recall for all chords in the test set, and f-measures calculated from these average results.

4 Experiments and Results

General results of our experiments are shown in Table 2, for various experimental settings regarding training data, classification methodology, and feature vector applied. As we can see, the classification quality is not as good as in case of our previous research, thus showing the increased level of difficulty in case of our current research.

The presented experiments were performed for various sets of training data, i.e. for isolated musical instrumental sounds only, and for mixes added to the training set. Classification was basically performed aiming at identification of each instrument (i.e. down to the leaves of hierarchical classification), but we also performed classification using information from nodes of the hierarchical tree, as described in Section 3. Experiments was performed for 2 versions of feature vector, including 63 parameters describing average values of sound features

Table 2. General results of recognition of 10 selected musical instruments playing in chords taken from real audio recording from RWC Classical Music Database [6]

Training data	Classification	Feature vector	Precision	Recall	F-measure
Isolated sounds + mixes	Leaves + nodes	Averages only	63.06%	49.52%	0.5547
Isolated sounds + mixes	Leaves only	Averages only	62.73%	45.02%	0.5242
Isolated sounds only	Leaves + nodes	Averages only	**74.10%**	32.12%	0.4481
Isolated sounds only	Leaves only	Averages only	71.26%	18.20%	0.2899
Isolated sounds + mixes	Leaves + nodes	Averages + temporal	57.00%	**59.22%**	**0.5808**
Isolated sounds + mixes	Leaves only	Averages + temporal	57.45%	53.07%	0.5517
Isolated sounds only	Leaves + nodes	Averages + temporal	51.65%	25.87%	0.3447
Isolated sounds only	Leaves only	Averages + temporal	54.65%	18.00%	0.2708

Table 3. Results of recognition of 10 selected musical instruments playing in chords taken from real audio recording from RWC Classical Music Database [6] - the results for best settings for each instruments are shown

	precision	recall	f-measure
bflatclarinet	50.00%	16.22%	0.2449
cello	69.23%	77.59%	0.7317
doublebass	40.00%	61.54%	0.4848
flute	31.58%	33.33%	0.3243
frenchhorn	20.00%	47.37%	0.2813
oboe	16.67%	11.11%	0.1333
piano	14.29%	16.67%	0.1538
tenorTrombone	25.00%	25.00%	0.2500
viola	63.24%	72.88%	0.6772
violin	89.29%	86.21%	0.8772

calculated through the entire sound in the first version of the feature vector, and additionally temporal parameters describing the evolution of these features in time in the second version. Precision and recall for these settings, as well as F-measure, are shown in Table 2.

As we can see, when training is performed on isolated sound only, the obtained recall is rather low, and it is increased when mixes are added to the training set. On the other hand, when training is performed on isolated sound only, the highest precision is obtained. This is not surprising, as illustrating a usual trade-off between precision and recall. The highest recall is obtained when information from nodes of hierarchical classification is taken into account. This was also expected; when the user is more interested in high recall than in high precision, then such a way of classification should be followed. Adding temporal descriptors to the feature vector does not make such a clear influence on the obtained precision and recall, but it increases recall when mixes are present in the training set.

One might be also interested in inspecting the results for each instrument. These results are shown in Table 3, for best settings of the classifiers used. As we can see, some string instruments (violin, viola and cello) are relatively easy to recognize, both in terms of precision and recall. Oboe, piano and trombone are difficult to be identified, both in terms of precision and recall. For double bass recall is much better than precision, whereas for clarinet the obtained precision is better than recall. Some results are not very good, but we must remember that correct identification of all instruments playing in a chord is generally a difficult task, even for humans.

It might be interesting to see which instruments are confused with which ones, and this is illustrated in confusion matrices. As we mentioned before, omissions and false positives can be considered in various ways, thus we can present different confusion matrices, depending on how the errors are counted. In Table 4 we presents the results when $1/n$ is added in each cell when identification happens (n represents the number of instruments actually playing in the mix).

Table 4. Confusion matrix for the recognition of 10 selected musical instruments playing in chords taken from real audio recording from RWC Classical Music Database [6]. When n instruments are actually playing in the recording, $1/n$ is added in case of each identification.

Classified as / Instrument	clarinet	cello	dBass	flute	fHorn	oboe	piano	trombone	viola	violin
clarinet	6	2	1	3.08	4.42	1.75	2.42	0.75	4.92	0.58
cello	2	45	4.67	0.75	8.15	1.95	3.2	1.08	1.5	0.58
dBass	0	0.25	16	0.5	2.23	0.45	1.12	0	0.5	0.25
flute	0.67	0.58	1.17	6	1.78	1.37	0.95	0	0.58	0.5
fHorn	0	4.33	1.83	0.17	9	0	0.33	0	4.83	3
oboe	0	0.67	0.33	1.33	1.67	2	1.5	0.33	0	0.5
piano	0	4.83	2.83	0	0	0	3	0	4.83	3
trombone	0	0	0	0.17	0.53	0	0.92	2	0.58	0.58
viola	1.33	1.75	4.5	2.25	7.32	1.03	3.28	1.92	43	0
violin	2	5.58	7.67	4.75	9.9	3.45	4.28	1.92	7.25	75

Table 5. Confusion matrix for the recognition of 10 selected musical instruments playing in chords taken from real audio recording from RWC Classical Music Database [6]. In case of each identification, 1 is added in a given cell.

Classified as / Instrument	clarinet	cello	dBass	flute	fHorn	oboe	piano	trombone	viola	violin
clarinet	6	4	2	8	17	4	8	3	11	2
cello	6	45	14	4	31	7	13	4	5	2
dBass	0	1	16	3	12	2	6	0	2	1
flute	2	2	4	6	7	5	3	0	2	1
fHorn	0	10	4	1	9	0	2	0	12	6
oboe	0	2	1	5	9	2	5	1	0	1
piano	0	11	6	0	0	0	3	0	12	6
trombone	0	0	0	1	2	0	4	2	2	2
viola	4	5	14	8	29	4	13	6	43	0
violin	6	14	21	13	35	10	15	6	18	75

To compare with, the confusion matrix is also shown when each identification is counted as 1 instead (Table 5). We believe that Table 4 more properly describes the classification results than Table 5, although the latter is more clear to look at. We can observe from both tables which instruments are confused with which ones, but we must remember that we are aiming at identifying actually a group of instruments, and our output also represents a group. Therefore, concluding about confusion between particular instruments is not so simple and straightforward, because we do not know exactly which instrument caused which confusion.

5 Summary and Conclusions

The investigations presented in this paper aimed at identification of instruments in real audio polytimbral (multi-instrumental) recordings. The parameterization included temporal descriptors, which improved recall when training was performed on both single isolated sounds and mixes. The use of real recordings not included in training set posed high level of difficulties for the classifiers; not only the sounds of instruments originated from different audio sets, but also the recording conditions were different. Taking this into account, we can conclude that the results were not bad, especially that some sounds were soft, and still several instruments were quite well recognized (certainly higher than random choice). In order to improve classification, we can take into account usual settings of instrumentation and the probability of use of particular instruments and instrument groups playing together. The classifiers adjusted specifically to given genres and sub-genres may yield much higher results, further improved by taking into account cleaning of results (removal of spurious single indications in the context of neighboring recognized sounds). Basing on the results of other research [20], we also believe that adjusting the feature set and performing feature selection in each node should improve our results. Finally, adjusting thresholds of firing of the classifiers may improve the results.

Acknowledgments. This project was partially supported by the Research Center of PJIIT, supported by the Polish National Committee for Scientific Research (KBN) and also by the National Science Foundation under Grant Number *IIS 0968647*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Birmingham, W.P., Dannenberg, R.D., Wakefield, G.H., Bartsch, M.A., Bykowski, D., Mazzoni, D., Meek, C., Mellody, M., Rand, B.: MUSART: Music retrieval via aural queries. In: Proceedings of ISMIR 2001, 2nd Annual International Symposium on Music Information Retrieval, Bloomington, Indiana, pp. 73–81 (2001)
2. Breiman, L., Cutler, A.: Random Forests, `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`
3. Dziubinski, M., Dalka, P., Kostek, B.: Estimation of musical sound separation algorithm effectiveness employing neural networks. J. Intel. Inf. Syst. 24(2-3), 133–157 (2005)
4. Downie, J.S.: Wither music information retrieval: ten suggestions to strengthen the MIR research community. In: Downie, J.S., Bainbridge, D. (eds.) Proceedings of the Second Annual International Symposium on Music Information Retrieval: ISMIR 2001, pp. 219–222. Bloomington, Indiana (2001)
5. Foote, J., Uchihashi, S.: The Beat Spectrum: A New Approach to Rhythm Analysis. In: Proceedings of the International Conference on Multimedia and Expo ICME 2001, Tokyo, Japan, pp. 1088–1091 (2001)

6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases. In: Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002), pp. 287–288 (2002)
7. Guaus, E., Herrera, P.: Music Genre Categorization in Humans and Machines, AES 121st Convention, San Francisco (2006)
8. Heittola, T., Klapuri, A., Virtanen, T.: Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In: 10th ISMIR, pp. 327–332 (2009)
9. Herrera, P., Amatriain, X., Batlle, E., Serra, X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In: International Symposium on Music Information Retrieval ISMIR (2000)
10. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
11. ISO: MPEG-7 Overview, http://www.chiariglione.org/mpeg/
12. Itoyama, K., Goto, M., Komatani, K., Ogata, T., Okuno, H.G.: Instrument Equalizer for Query-By-Example Retrieval: Improving Sound Source Separation Based on Integrated Harmonic and Inharmonic Models. In: 9th ISMIR (2008)
13. Jiang, W.: Polyphonic Music Information Retrieval Based on Multi-Label Cascade Classification System. Ph.D thesis, Univ. North Carolina, Charlotte (2009)
14. Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.: Instrogram: Probablilistic Representation of Instrument Existence for Polyphonic Music. IPSJ Journal 48(1), 214–226 (2007)
15. Klapuri, A.: Signal processing methods for the automatic transcription of music. Ph.D. thesis, Tampere University of Technology, Finland (2004)
16. Kursa, M.B., Kubera, E., Rudnicki, W.R., Wieczorkowska, A.A.: Random Musical Bands Playing in Random Forests. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 580–589. Springer, Heidelberg (2010)
17. Kursa, M., Rudnicki, W., Wieczorkowska, A., Kubera, E., Kubik-Komar, A.: Musical Instruments in Random Forest. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) Foundations of Intelligent Systems. LNCS, vol. 5722, pp. 281–290. Springer, Heidelberg (2009)
18. Lauser, B., Hotho, A.: Automatic multi-label subject indexing in a multilingual environment. FAO, Agricultural Information and Knowledge Management Papers (2003)
19. Little, D., Pardo, B.: Learning Musical Instruments from Mixtures of Audio with Weak Labels. In: 9th ISMIR (2008)
20. Mierswa, I., Morik, K., Wurst, M.: Collaborative Use of Features in a Distributed System for the Organization of Music Collections. In: Shen, J., Shephard, J., Cui, B., Liu, L. (eds.) Intelligent Music Information Systems: Tools and Methodologies, pp. 147–176. IGI Global (2008)
21. Niewiadomy, D., Pelikant, A.: Implementation of MFCC vector generation in classification context. Journal of Applied Computer Science 16(2), 55–65 (2008)
22. Opolko, F., Wapnick, J.: MUMS – McGill University Master Samples. CD's (1987)
23. Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T.: Computational Auditory Scene Recognition. In: International Conference on Acoustics Speech and Signal Processing, Orlando, Florida (2002)
24. Raś, Z.W., Wieczorkowska, A.A. (eds.): Advances in Music Information Retrieval. Studies in Computational Intelligence, vol. 274. Springer, Heidelberg (2010)

25. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2009)
26. The University of IOWA Electronic Music Studios: Musical Instrument Samples, http://theremin.music.uiowa.edu/MIS.html
27. The University of Waikato: Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/
28. Miotto, R., Montecchio, N., Orio, N.: Statistical Music Modeling Aimed at Identification and Alignment. In: Raś, Z.W., Wieczorkowska, A.A. (eds.) Advances in Music Information Retrieval. SCI, vol. 274, pp. 187–212. Springer, Heidelberg (2010)
29. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. Organized Sound 4(3), 169–175 (2000)
30. Viste, H., Evangelista, G.: Separation of Harmonic Instruments with Overlapping Partials in Multi-Channel Mixtures. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA 2003, New Paltz, NY (2003)
31. Wieczorkowska, A.A., Kubera, E.: Identification of a dominating instrument in polytimbral same-pitch mixes using SVM classifiers with non-linear kernel. J. Intell. Inf. Syst. (2009), doi: 10.1007/s10844-009-0098-3
32. Wieczorkowska, A., Kubera, E., Kubik-Komar, A.: Analysis of Recognition of a Musical Instrument in Sound Mixes Using Support Vector Machines. In: Nguyen, H.S. (ed.) SCKT 2008 Hanoi, Vietnam (PRICAI), pp. 110–121 (2008)
33. Wieczorkowska, A.A.: Music Information Retrieval. In: Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining, 2nd edn., pp. 1396–1402. IGI Global (2009)
34. Wieczorkowska, A., Synak, P.: Quality Assessment of k-NN Multi-Label Classification for Music Data. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 389–398. Springer, Heidelberg (2006)
35. Zhang, X.: Cooperative Music Retrieval Based on Automatic Indexing of Music by Instruments and Their Types. Ph.D thesis, Univ. North Carolina, Charlotte (2007)
36. Zhang, X., Marasek, K., Raś, Z.W.: Maximum Likelihood Study for Sound Pattern Separation and Recognition. In: 2007 International Conference on Multimedia and Ubiquitous Engineering MUE 2007, pp. 807–812. IEEE, Los Alamitos (2007)