Training of Classifiers for the Recognition of Musical Instrument Dominating in the Same-Pitch Mix

Alicja Wieczorkowska[1], Elżbieta Kolczyńska[2], and Zbigniew W. Raś[3,1]

[1] Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland
alicja@pjwstk.edu.pl
[2] Agricultural University in Lublin,
Akademicka 13, 20-950 Lublin, Poland
elzbieta.kolczynska@ar.lublin.pl
[3] University of North Carolina,
Department of Computer Science, Charlotte, NC 28223, USA
ras@uncc.edu

Abstract. Preparing a database to train classifiers for identification of musical instruments in audio files is very important, especially in a case of sounds of the same pitch, when a dominating instrument is most difficult to identify. Since it is infeasible to prepare a data set representing all possible ever recorded mixes, we had to reduce the number of sounds in our research to a reasonable size. In this paper, our data set represents sounds of selected instruments of the same octave, with additions of artificial sounds of broadband spectra for training, and additions of sounds of other instruments for testing purposes. We tested various levels of added sounds taking into consideration only equal steps in logarithmic scale which are more suitable for amplitude comparison than linear one. Additionally, since musical instruments can be classified hierarchically, experiments for groups of instruments representing particular nodes of such hierarchy have been also performed. The set-up of training and testing sets, as well as experiments on classification of the instrument dominating in the sound file, are presented and discussed in this paper.

1 Introduction

One of the main goals in Music Information Retrieval (MIR) area is a creation of a storage and retrieval system that can automatically index musical input into a database and answer queries requesting specific musical pieces. The system may search for a specified melody, but also it should be able to automatically identify the most dominating musical instruments associated with the musical segment. In knowledge discovery based approach, the system usually divides a musical waveform into segments of equal size right after receiving it as the input data. These segments have to be somehow compared with a very large number of sound objects in a large musical database and checked for their similarity. The problem

is greatly simplified by treating a database of singular monophonic sound objects as a training database for constructing classifiers for music automatic indexing by instruments. The precision and recall of the classifiers, trained on isolated sounds of singular instruments and then tested on polyphonic music, is very low. One way to handle the problem successfully, is trying to improve sound separation algorithm. Another approach is to enlarge the training database by a new set of tuples representing sounds coming from certain pairs (or groups) of instruments with a goal to train new classifiers for recognizing instruments in polyphonic music. The testing we have done on a small group of instruments supports the second approach.

In this paper, the experiments will focus on observing if (and how) mixing the clean musical instrument sound data with other sounds (i.e. adding accompanying sound) may influence the correctness of a classification of instruments, dominating in a polyphonic recording. The clean data represent singular musical instrument sounds of definite pitch and harmonic spectrum. The additions used for testing represent mixes of musical instrument sounds of various levels added to singular monophonic musical instrument sounds; the training additions represent artificial harmonic and noise type sound waves of broadband spectra. Our plan was to establish thresholds (if such thresholds exist) for a level of sounds added to singular instrument sounds which guarantee the highest confidence of classifiers for polyphonic recordings.

The categorization of musical instrument sounds can be used to assist extraction of information requested by users, who are browsing the database of music pieces and looking for the excerpts played by a desired instrument. However, the excerpt played by this instrument may not be available, so in such cases, a piece representing a similar category should be returned by the system.

There are many ways to categorize music instruments, such as by playing methods, by instrument type, or by other generalization concepts. Any categorization process can be represented as a hierarchical schema which is used by a cooperative query answering system to handle failing queries. By definition, a cooperative system is relaxing a failing query with a goal to find its smallest generalization which does not fail. Two different hierarchical schemas, used as models of a decision attribute have been already investigated by authors of this paper (see [10]): Hornbostel-Sachs classification of musical instruments [5], and classification of musical instruments by articulation, i.e. the way the sound is started, played and ended, with the following articulation methods (seen as attribute values): blown, bowed, bowed vibrato, concussive, hammered, lip-vibrated, martele, muted, muted vibrato, picked, pizzicato, rubbed, scraped and shaken. Each hierarchical classification represents a unique hierarchical decision attribute in MIRAI database (http://www.mir.uncc.edu), leading to a construction of new classifiers and the same to a different system for automatic indexing of music by instruments and their types.

In our previous research, we already tried to search for thresholds (describing the loudness level of added sounds) which are most effective for recognition of dominating instrument in sound mix, using 5 linear or 3 logarithmic levels [13].

No such thresholds have been found. This time, we decided to use denser steps for the loudness level, considering also experiments when the accuracy improves or worsens around maximum, to establish the threshold more precisely. Also, since the thresholds may differ for particular instruments or instrument groups, we decided to investigate hierarchical classifications as well. If threshold level can be found (at least for some nodes), the classifiers for each node of a hierarchy can be built basing on the sounds associated with this node, with added sounds from outside the identified class and with loudness level determined by this threshold. We also believe that hierarchical classifiers may outperform the general classifiers at each level of classification, comparing the performance for the groups of classes corresponding to a given level of a hierarchical classifier.

The paper is organized as follows: in Section 2, sound parameterization for instrument identification is briefly described. Section 3 presents sound data used for training and testing in our experiments. Classification results are shown in Section 4. Section 5 summarizes and concludes the paper.

2 Data Parameterization

Digital music (audio) data represent series of samples, where each sample represents instantaneous amplitude value of the recorded sound wave. Therefore, parameterization is needed to perform classification experiments on such sounds, since even the slightest change in the sound wave causes considerable changes in the recorded amplitude samples. Parameterization may describe temporal, spectral, and spectral-temporal properties of the sound. Numerous parameters have been used so far in research on musical instrument sound recognition, including features describing properties of DFT spectrum, wavelet analysis coefficients, MFCC (Mel-Frequency Cepstral Coefficients), MSA (Multidimensional Analysis Scaling) trajectories, and so on [2], [3], [7], [8], [12]. General overview of parameterization and classification of musical instrument sounds is given in [4]. Also, MPEG-7 sound descriptors can be applied for sound parameterization purposes [6]. However, these parameters are not dedicated to recognition of particular instruments in recordings. The choice of parameters for feature vector is an important part of obtaining the data for classification purposes, and the results may vary depending on the parameters and classifiers chosen.

In this research, the feature vector consists of 219 parameters, based on MPEG-7 and other parameters already applied for the recognition of musical instruments from singular sounds, and also in polyphonic (polytimbral) environment [14]. The parameterization was performed for the investigated sounds using 120 ms analyzing frame, sliding along the entire sound, with Hamming window and hop size 40 ms. Long frame was used in order to parameterize low sounds if needed. Most of the calculated parameters represent average value of parameters calculated for consecutive frames of a sound; some of the descriptors are multi-dimensional. This set of features has been already used in previous research [13], [14]; it consists of the following parameters:

- MPEG-7 audio descriptors: $AudioSpectrumSpread$, $AudioSpectrumFlatness$, $AudioSpectrumCentroid$, $AudioSpectrumBasis$, $HarmonicSpectralCentroid$, $HarmonicSpectralSpread$, $HarmonicSpectralVariation$, $HarmonicSpectralDeviation$, $LogAttackTime$, $TemporalCentroid$;
- other descriptors: $Energy$, $MFCC$, $ZeroCrossingDensity$, $RollOff$, $Flux$, $AverageFundamentalFrequency$, $Ratio$.

3 Training and Testing Data

In our experiments, we decided to choose 8 instruments, representing aerophones and chordophones in Hornbostel-Sachs classification:

- B-flat clarinet (aerophone),
- cello - bowed, played vibrato (chordophone),
- trumpet (aerophone),
- flute played vibrato (aerophone),
- oboe (aerophone),
- tenor trombone (aerophone),
- viola - bowed, played vibrato (chordophone),
- violin - bowed, played vibrato (chordophone).

These instruments produce sounds of definite pitch, with spectra of harmonic type. We have chosen sustained sounds from the octave no. 4 (in MIDI notation) of these instruments, i.e. 12 sounds for each instrument. These sounds represent instruments to be used for training classifiers. The sounds were taken from McGill University Master Samples CDs [9]. Sampling rate 44.1 kHz and 16-bit resolution was chosen to prepare digital audio samples. Sounds were recorded in .snd format, and left channel of stereo recordings was used.

Apart from singular sounds, mixes with other sounds were used, both for training and testing purposes. The added sounds were diminished in level. After rescaling the amplitude of the added sounds, to match the RMS of the main sound, the level of added sounds was diminished with respect to the main sound with scaling factor $\sqrt{2}$. The following levels were used:

- 50%,
- $50/\sqrt{2} \approx 35.355339059327376220042218105242\%$,
- 25%,
- $25/\sqrt{2} \approx 17.677669529663688110021109052621\%$
- 12.5%,
- $12.5/\sqrt{2} \approx 8.838834764831844055010554526310 6\%$,
- 6.25%

These levels represent logarithmic diminishing of amplitude, and are perceived by human hearing system as uniform, since human perception is logarithmic with respect to changes of stimulus of any type.

Since the amplitude of any musical instrument sound changes in time, the added sounds were truncated (if needed) to the length of the main sound. Also,

we replaced 0.1 s of the beginning and ending of added sound with silence. Next, fade-in effect was applied from the end of the silence at the beginning till 1/3 of the sound length, and similarly fade-out from 2/3 of the sound. Thus we ensure that even during transients the main sound is still dominating in the mix.

The training was performed on singular sounds of musical instruments, as mentioned above, and also on the same sounds with added artificial harmonic sound waves of the same pitch and noises, generated using Adobe Audition [1]:

- white noise,
- pink noise,
- triangular wave,
- saw-tooth wave.

All these sounds have broadband spectra, continuous (noises) or harmonic (triangular and saw-tooth wave), strongly overlapping with the main sounds. The frequency values of the generated harmonic waves were rounded to the nearest integers, as below:

- C4 - 262 Hz,
- C#4 - 277 Hz,
- D4 - 294 Hz,
- D#4 - 311 Hz,
- E4 - 330 Hz,
- F4 - 349 Hz,
- F#4 - 370 Hz,
- G4 - 392 Hz,
- G#4 - 415 Hz,
- A4 - 440 Hz,
- A#4 - 466 Hz,
- B4 - 494 Hz.

Eight-second long sounds were generated, since the longest instrumental sound was less than 8 s long.

The testing was performed on the musical instrument sounds mixed with other instruments, of level diminished as described above. Singular sounds were not used for tests, since we know from other experiments that the results in this case are very high, and this was not our subject of experiments.

Additionally, hierarchical classification was performed. In this case, the binary classifier to distinguish between aerophones and chordophones was trained first. Aerophones can be further divided into subclasses: single reed (clarinet), double reed (oboe), lip vibrated (trumpet, trombone), and side blown flute (flute). Chordophones used in our experiments represent one subclass in Hornbostel-Sachs classification. Therefore, we decided to investigate chordophones from our data set, to see if the classification improves, and if thresholds for levels of added sounds can be found. The classifier for chordophones in our research was trained to identify violin, viola, or cello, when one of these instruments dominates in the recording. Obviously, similar classification can be performed for aerophones,

but then more instruments should rather be used. The problem which needs to taken into account is that in some cases, if the binary classifier distinguishing aerophones and chordophones yields erroneous results, the next classifier (for the subclass - chordophones in our research) may only yield random results. Therefore, the general classifier (for all classes) can perform better in such cases.

3.1 Training Data

The main training data set consists of 96 singular sounds, representing 4th octave of 8 instruments, as mentioned before. Also, another version of the training set was used, containing both singular sounds, and the same sounds with added noises and artificial harmonic sounds, as described above. These data were used to train classifiers for the recognition of one of these 8 instruments.

The training data for the binary classifier, distinguishing between aerophones and chordopnones, represented instruments only from these 2 classes: clarinet, oboe, trumpet, trombone, and flute sounds (singular sounds and in mixes) represented aerophones, whereas viola, violin and cello sounds (singular and in mixes) represented chordophones class. The training data for the classifiers dedicated for chordophones was trained on violin, viola, and cello sounds, singular and in mixes.

3.2 Testing Data

The testing data for the general classifier identifying one of our 8 instruments, as described above, consisted of the sounds of these instruments with added sounds of other 7 instruments. More precisely, for each sound of each instrument, the mix with sounds of the same pitch representing the remaining 7 instruments was prepared. The level of the added sound was calculated as average of this sum, and then modified as in the case of a training set. The same level was used for training and testing (if mixes were used in training set).

In a case of binary classifier distinguishing between aerophones and chordophones, the sounds added to any aerophone sound represented the sum of chordophones for the same pitch (averaged and modified in level, as before). Similarly, the sounds added to any chordophone represented analogous sum of aerophones.

For testing a classifier dedicated to recognize particular chordophones in our experiments, the test data for violin represented mixes with viola and cello, the test data for viola represented mixes with violin and cello, and the test data for cello represented mixes with viola and violin. The sounds added in mixes were modified in loudness level, as before.

4 Experiments and Results

Classification experiments were performed using WEKA software [11]. Support Vector Machine (SMO) classifier was chosen, as appropriate for multi-

dimensional data, and already used in similar research. Standard settings of this classifier were used.

General results of correctness of instrument identification for various levels of sounds added to the main instrument are shown in Fig. 1; as we can see, classifier yields higher results if trained on both singular and mixed sounds. Confusion matrices for the training on singular sounds only, and for the training on both singular and mixed sounds (for the same level of added sounds both in the training and test set) is shown in Fig. 2. As we can see, viola was often classified as cello, both in case of training on singular sound only and in training on both singular sounds and mixes. However, the addition of mixed sounds to the training set improves identification of oboe and trombone to 100% for all levels of added sounds, and also improves recognition of flute. However, violin and viola are always difficult to discern for this classifier. Therefore, we expect that the dedicated classifier for chordophones may help to improve this.

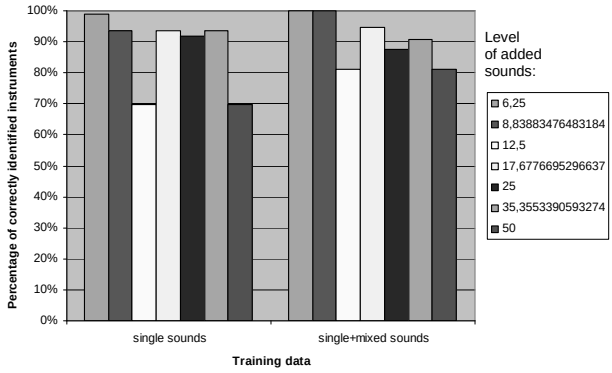

Fig. 1. Correctness of identification of dominating instruments for various types of training data (singular sounds only, or singular and mixed sounds), with testing on sound mixes with various levels of added sounds

The hierarchical classification required the training of a binary classifier first; this classifier should discern between aerophones and chordophones. The contingency tables for this classifier are shown in Fig. 3. As we mentioned before, the next level of classification was performed for chordophones, with 3 instruments to identify: violin, viola, and cello. Contingency tables for this classifier are shown in Fig. 4. Percentage of correctness for both classifiers is shown in Fig. 5. Final correctness of classification can be calculated by multiplying the appropriate correctness values of both classifiers.

Unfortunately, no thresholds for the best performance of classifiers were observed, but generally the lowest levels of added sounds yield best results.

Classified as ->	clarinet		cello		trumpet		flute		oboe		trombone		viola		violin	
clarinet+6.25%	12	12														
clarinet+8.49%	12	12														
clarinet+12.5%	11	10		1						1			1			
clarinet+17.68%	12	12														
clarinet+25%	12	12														
clarinet+35.36%	12	12														
clarinet+50%	11	10		1						1			1			
cello+6.25%			12	12												
cello+8.49%			12	12												
cello+12.5%			12	12												
cello+17.68%			12	12												
cello+25%			12	12												
cello+35.36%			12	12												
cello+50%			12	12												
trumpet+6.25%					12	12										
trumpet+8.49%					12	12										
trumpet+12.5%					12	12										
trumpet+17.68%					12	12										
trumpet+25%					12	12										
trumpet+35.36%					12	12										
trumpet+50%					12	12										
flute+6.25%							12	12								
flute+8.49%							12	12								
flute+12.5%		1	4	1			8	8		2						
flute+17.68%							12	12								
flute+25%							11	12					1			
flute+35.36%							12	12								
flute+50%		1	4	1			8	8		2						
oboe+6.25%									11	12			1			
oboe+8.49%			1						11	12						
oboe+12.5%			3						4	12			1		4	
oboe+17.68%			1						11	12						
oboe+25%									11	12			1			
oboe+35.36%			1						11	12						
oboe+50%			3						4	12			1		4	
trombone+6.25%											12	12				
trombone+8.49%											12	12				
trombone+12.5%			1		2						8	12	1			
trombone+17.68%											12	12				
trombone+25%											12	12				
trombone+35.36%											12	12				
trombone+50%			1		2						8	12	1			
viola+6.25%													12	12		
viola+8.49%			2										10	12		
viola+12.5%			5	3									5	9	2	
viola+17.68%			2	2									10	10		
viola+25%			3	3									9	9		
viola+35.36%			2	2									10	10		
viola+50%			5	3									5	9	2	
violin+6.25%															12	12
violin+8.49%											1		2		9	12
violin+12.5%		1	1							1			4	7	7	3
violin+17.68%											1		2	3	9	9
violin+25%										1			3	8	9	3
violin+35.36%										2	1		2	5	9	5
violin+50%		1	1							1			4	7	7	3

Fig. 2. Contingency table for classifiers trained on singular sounds - left columns, and on both singular and mixed sounds - right columns

5 Conclusions

These experiments were performed on selected instruments, representing aerophones and chordophones. We consider continuation of our experiments using all instruments from these classes, since the selected instruments do not show the whole picture of the problem. In a case of a few instruments, the testing data can easily represent sounds of all other instruments, covered by a given classifier. In a case of numerous instruments, we should rather use artificial sounds, or other mixes, because it is rather unrealistic to prepare all possible mixes with other instruments. When collecting data, we may also consider sounds played with various articulation. They may represent one class (i.e. instrument), or separate classes for each articulation method. Also, we can perform similar experiments to identify articulation, since, for example, sounds of viola and violin played

A. Classified as ->	aerophone	chordophone
aerophone+6.25%	60	
aerophone+8.49%	59	1
aerophone+12.5%	60	
aerophone+17.68%	60	
aerophone+25%	60	
aerophone+35.36%	56	4
aerophone+50%	54	6
chordophone+6.25%	1	35
chordophone+8.49%		36
chordophone+12.5%		36
chordophone+17.68%	1	35
chordophone+25%	1	35
chordophone+35.36%	3	33
chordophone+50%	2	34

B. Classified as ->	aerophone	chordophone
aerophone+6.25%	60	
aerophone+8.49%	60	
aerophone+12.5%	60	
aerophone+17.68%	60	
aerophone+25%	60	
aerophone+35.36%	59	1
aerophone+50%	59	1
chordophone+6.25%		36
chordophone+8.49%		36
chordophone+12.5%	5	31
chordophone+17.68%	2	34
chordophone+25%		36
chordophone+35.36%	3	33
chordophone+50%	9	27

Fig. 3. Contingency tables for the aerophones/chordophones classifier for the training on singular sounds only (table A) and on both singular and mixed sounds (B)

A. Classified as ->	cello	viola	violin
cello+6.25%	12		
cello+8.49%	12		
cello+12.5%	12		
cello+17.68%	12		
cello+25%	12		
cello+35.36%	12		
cello+50%	12		
viola+6.25%		12	
viola+8.49%		12	
viola+12.5%		12	
viola+17.68%		12	
viola+25%	1	11	
viola+35.36%		11	1
viola+50%	1	10	1
violin+6.25%			12
violin+8.49%			12
violin+12.5%			12
violin+17.68%			12
violin+25%			12
violin+35.36%		2	10
violin+50%		3	9

B. Classified as ->	cello	viola	violin
cello+6.25%	12		
cello+8.49%	12		
cello+12.5%	11	1	
cello+17.68%	12		
cello+25%	11	1	
cello+35.36%	11	1	
cello+50%	11	1	
viola+6.25%		12	
viola+8.49%		12	
viola+12.5%		11	1
viola+17.68%		12	
viola+25%		12	
viola+35.36%		12	
viola+50%	1	11	
violin+6.25%			12
violin+8.49%			12
violin+12.5%		2	10
violin+17.68%		1	11
violin+25%		1	11
violin+35.36%		1	11
violin+50%		3	9

Fig. 4. Contingency tables for the chordophone classifier, for the training on singular sounds only (table A) and on both singular and mixed sounds (B)

pizzicato can be considered more similar that sound of violin played vibrato and pizzicato. Additionally, more attributes describing sound samples may be needed, especially for lower levels of hierarchical classification. It is possible that the most efficient levels of added sounds can be found for some nodes of hierarchical classifier, or maybe we can find new classes of instruments, for which we are able to find such thresholds, and thus introduce a new decision attribute to our database, and new classification.

Acknowledgments. This work was supported by the National Science Foundation under grant IIS-0414815, and also by the Research Center of PJIIT, supported by the Polish National Committee for Scientific Research (KBN).

References

1. Adobe Systems Incorporated: Adobe Audition 1.0, 2003

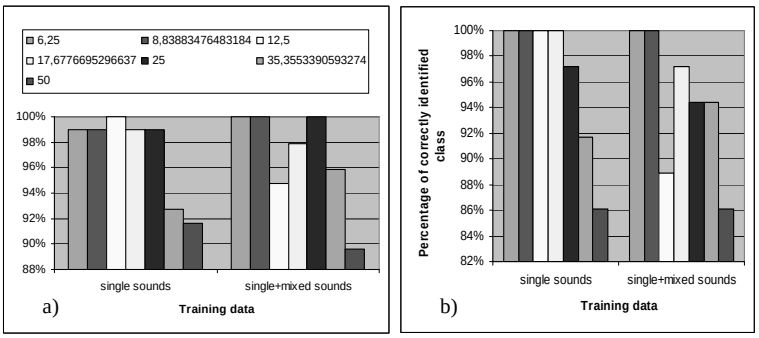

Fig. 5. Correctness of identification of dominating instruments for the aerophones/chordophones classifier (a), and for the chordophone classifier (b)

2. Aniola, P., Lukasik, E.: JAVA Library for Automatic Musical Instruments Recognition. AES 122 Convention, Vienna, Austria, May 2007
3. Brown, J. C.: Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. Journal of the Acoustical Society of America **105** (1999) 1933–1941
4. Herrera, P., Amatriain, X., Batlle, E., Serra X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. Int. Symp. on Music Information Retrieval ISMIR 2000, Plymouth, MA
5. Hornbostel, E. M. V., Sachs, C.: Systematik der Musikinstrumente. Ein Versuch. Zeitschrift fur Ethnologie, Vol. 46, No. 4-5, 1914, 553-90.
6. ISO/IEC JTC1/SC29/WG11: MPEG-7 Overview. (2004) Available at `http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm`
7. Kaminskyj, I.: Multi-feature Musical Instrument Sound Classifier w/user determined generalisation performance. Proceedings of the Australasian Computer Music Association Conference ACMC 2002, 53–62
8. Martin, K. D., Kim, Y. E.: Musical instrument identification: A pattern-recognition approach. 136-th meeting of the Acoustical Society of America, Norfolk, VA (1998)
9. Opolko, F., Wapnick, J.: MUMS - McGill University Master Samples. CD's (1987)
10. Ras, Z.W., Zhang, X., Lewis, R.: MIRAI: Multi-hierarchical, FS-Tree Based Music Information Retrieval System, Proc. International Conf. on Rough Sets and Intelligent Systems Paradigms, RSEISP 2007, LNAI, Vol. 4585, Springer (2007) 80–89
11. The University of Waikato: Weka Machine Learning Project. Internet, 2007. Available at `http://www.cs.waikato.ac.nz/~ml/`
12. Wieczorkowska, A.: Towards Musical Data Classification via Wavelet Analysis. In: Ras, Z. W., Ohsuga, S. (eds.): Foundations of Intelligent Systems. Proc. ISMIS'00, Charlotte, NC, USA, LNCS/LNAI, Vol. 1932, Springer-Verlag (2000) 292–300
13. Wieczorkowska, A., Kolczyńska, E.: Quality of Musical Instrument Sound Identification for Various Levels of Accompanying Sounds. In: Ras, Z. W., Tsumoto, S., Zighead D. (eds.): Mining Complex Data, Post-proceedings. LNCS/LNAI 2007
14. Zhang, X.: Cooperative Music Retrieval Based on Automatic Indexing of Music by Instruments and Their Types. Ph.D thesis, Univ. North Carolina, Charlotte 2007