

Identification of Dominating Instrument in Mixes of Sounds of the Same Pitch

Alicja Wieczorkowska¹ and Elżbieta Kolczyńska²

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

Abstract. In this paper we deal with the problem of identification of the dominating instrument in the recording containing simultaneous sounds of the same pitch. Sustained harmonic sounds from one octave of eight instruments were considered. The training data set contains sounds of singular instruments, as well as the same sounds with added artificial harmonic and noise sounds of lower amplitude. The test data set contains mixes of musical instrument sounds. SVM classifier from WEKA was used for training and testing experiments. Results of these experiments are presented and discussed in the paper.

Key words: music information retrieval, sound recognition

1 Introduction

Automatic recognition of the dominating musical instrument in sound mixes, when spectra overlap, is quite a difficult issue, most difficult when the interfering sounds are of the same pitch. The number of possible combinations is very high because of the number of existing instruments and sounds within their scale ranges. Therefore, it would be desirable to obtain a classifier performing such recognition, and also to train the classifier on a limited data set. The motivation for this paper was to perform experiments on selected instrument sounds, and use added artificial sounds with broadband spectrum, overlapping with the sounds under consideration, in order to check if classifiers trained this way would work for sounds mixes of real instruments. In other words, our goal was to check if using a limited number of added artificial sounds can be sufficient to train a classifier to recognize dominating musical instrument in polytimbral mix of one pitch. The main focus was on construction of the training and testing data, because if this set up is successful, we have a starting point for further experiments with other musical instrument sound mixes.

In this research, we decided to choose 8 instruments producing sustained harmonic sounds (of definite pitch), and limit the range to the octave no. 4 in MIDI notation. The sounds added to the original sounds in the training set

include noises and artificial sound waves of harmonic spectrum. The test set contains the original sounds mixed with sounds of other instruments, always of the same pitch. The level of the added sounds was processed to make sure that the sound of the main instrument is louder than the other sound all the time.

We have already performed experiments on training classifiers recognizing dominating musical instrument in polytimbral environment, i.e. when the sound of other instrument is accompanying the main sound. However, the pitch of the accompanying sound or sounds was usually different than the pitch of the main sound. Additionally, the added sounds were diminished in amplitude in a very simple way, i.e. by re-scaling their amplitude. Since the main and added sounds were not edited in any other way, in many cases the added sounds were starting earlier or ending later than the main ones, thus being actually louder in some parts, and obscuring the results [12]. Therefore, we decided to change the experiment set up to make sure that the main sound is actually louder all the time, and thus that the classifiers are trained properly.

2 Data for Training and Testing

The data for training and testing consist of musical instrument sounds of sustained harmonic type, also with addition of other sounds. The added sounds include artificially generated noises, white and pink, and also artificially generated harmonic waves, of triangular and saw-tooth shape, always of the same pitch as the main sound.

In our previous research we focused on sounds of musical instruments. However, the choice of the accompanying instrument sounds in the training and testing sets was arbitral, and the spectra were not overlapping in most cases. Also, the length and the level of accompanying sound was not normalized. As a result, a clear dependency between the quality of the classifiers and the level of accompanying sound was not observed. This is why we decided to normalize the sound length and level of the added sounds with respect to the main sounds, and make sure that the level of added sounds does not exceed the level of the main sounds at any time.

2.1 Parameterization

The audio data we deal with represent series of samples, saved in .snd format. Each sample is a discrete value, representing the instantaneous sound amplitude of digitized sound, recorded separately for each channel. If the sound is recorded stereo with 44,1 kHz sampling rate (i.e. 44100 samples per second per channel) and with 16-bit resolution (i.e. 16 bits per sample), then 1 second of such a recording consists of 176.4 kB, and even the smallest variation of the sound wave causes substantial changes to the amplitude values. Such data would be quite inconvenient for classification purposes. Therefore, audio data are usually parameterized before classification.

Features used for parameterization of musical audio data may describe temporal, spectral, and spectral-temporal properties of sounds. The research on musical instrument sound recognition conducted worldwide is based on various parameters, 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], [4], [6], [7], [11]. MPEG-7 sound descriptors can also be applied for musical sound parameterization [5], but these parameters are not dedicated to recognition of particular instruments in recordings.

The construction of feature set is an important part of creating the database for classification purposes, and the results may vary depending on the feature vector applied for the training and then testing of a classifier. In our research, we decided to use the feature vector already applied for the recognition of musical instruments in polyphonic (polytimbral) environment [13]. The feature set we have chosen consists of 219 parameters, based mainly on MPEG-7 audio descriptors, and on other parameters in similar research. Most of these parameters represent average value of frame-based attributes, calculated for consecutive frames of a singular sound using sliding analysis window, moved through the entire sound. The calculations were performed using 120 ms analyzing frame with Hamming window and hop size 40 ms; such a long analyzing frame allows analysis even of the lowest sounds. In this research, data from the left channel of stereo sounds were taken for parameterization, and the following features were used [13]:

- MPEG-7 audio descriptors [5], [9], [12]:
 - *AudioSpectrumSpread* - a RMS value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame; the value was averaged through frames for the entire sound;
 - *AudioSpectrumFlatness*, $flat_1, \dots, flat_{25}$ - parameter describing the flatness property of the power spectrum within a frequency bin for selected bins; 25 out of 32 frequency bands were used to calculate these parameters for a given frame, and the value was averaged for the entire sound;
 - *AudioSpectrumCentroid* - power weighted average of the frequency bins in the power spectrum of all the frames in a sound segment, calculated with a Welch method;
 - *AudioSpectrumBasis*: $basis_1, \dots, basis_{165}$; spectral basis parameters are calculated for the spectrum basis functions. In our case, the total number of sub-spaces in basis function is 33, and for each sub-space, minimum/maximum/mean/distance/standard deviation are extracted to flat the vector data. Distance is calculated as the summation of dissimilarity (absolute difference of values) of every pair of coordinates in the vector. Spectrum basis function is used to reduce the dimensionality by projecting the spectrum (for each frame) from high dimensional space to low dimensional space with compact salient statistical information. The calculated values were averaged over all analyzed frames of the sound;
 - *HarmonicSpectralCentroid* - the average (over the entire sound) of the instantaneous Harmonic Centroid, calculated for each analyzing frame.

The instantaneous Harmonic Spectral Centroid is calculated as the mean of the harmonic peaks of the spectrum, weighted by the amplitude in linear scale;

- *HarmonicSpectralSpread* - the average over the entire sound of the instantaneous harmonic spectral spread, calculated for each frame. Instantaneous harmonic spectral spread represents the standard deviation of the harmonic peaks of the spectrum with respect to the instantaneous harmonic spectral centroid, weighted by the amplitude;
 - *HarmonicSpectralVariation* - mean value over the entire sound of the instantaneous harmonic spectral variation, i.e. of the normalized correlation between the amplitude of the harmonic peaks of each two adjacent frames;
 - *HarmonicSpectralDeviation* - the average over the entire sound of the instantaneous harmonic spectral deviation, calculated for each frame, where the instantaneous harmonic spectral deviation represents the spectral deviation of the log amplitude components from a global spectral envelope;
 - *LogAttackTime* - the decimal logarithm of the duration from the time when the signal starts to the time when it reaches its maximum value, or when it reaches its sustained part, whichever comes first
 - *TemporalCentroid* - energy weighted mean of the sound duration; this parameter shows where in time the energy of the sound is focused;
- other audio descriptors:
- *Energy* - average energy of spectrum in the parameterized sound;
 - *MFCC* - min, max, mean, distance, and standard deviation of the MFCC vector, through the entire sound;
 - *ZeroCrossingDensity*, averaged through all frames for a given sound;
 - *RollOff* - the frequency below which an experimentally chosen percentage of the accumulated magnitudes of the spectrum is concentrated (averaged over all frames). It is a measure of spectral shape, used in the 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, averaged through the entire sound. This value multiplied by 10^7 to comply with the requirements of the classifier applied in our research;
 - *AverageFundamentalFrequency*;
 - *Ratio* r_1, \dots, r_{11} - parameters describing the ratio of the amplitude of a harmonic partial to the total harmonic partials.

These parameters describe basic spectral, timbral spectral and temporal audio properties, and also spectral basis descriptor, as described in the MPEG-7 standard. The spectral basis descriptor is a series of basis functions derived from the Singular Value Decomposition (SVD) of a normalized power spectrum. In order to avoid too high dimensionality of the feature vector, a few other features were derived from the spectral basis attribute. Other audio descriptors used in our feature vector include time-domain and spectrum-domain properties of sound, used in research on audio data classification.

2.2 Training Data

The training data contain singular sounds of musical instruments, and also the same sounds with added other sounds. The audio recordings from McGill University Master Samples CDs have been used as a source of these sounds [8]. These CSs are commonly used worldwide for research on musical instrument sounds. The following instruments have been chosen:

1. B-flat clarinet,
2. cello - bowed, played vibrato,
3. trumpet,
4. flute played vibrato,
5. oboe,
6. tenor trombone,
7. viola - bowed, played vibrato,
8. violin - bowed, played vibrato.

All these instruments produce sounds of definite pitch, and their spectra are of harmonic type. Only sustained sounds were considered. We decided to use only sounds from the octave no. 4 (in MIDI notation). We also prepared the mixes of pairs of sounds, i.e. the instrumental sounds mentioned above and the following sounds:

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

All added sounds have broadband spectra, continuous in case of noises and harmonic in case of triangular and saw-tooth wave. Again, harmonic sounds were prepared for the frequencies from the octave no. 4. These sounds were produced using Adobe Audition [1], where only integer values were allowed. Therefore, the frequency values of the generated harmonic waves were rounded to the nearest integers, as below (standard values for A4=440 Hz shown in parentheses):

- C4 - 262 Hz (261.6),
- C#4 - 277 Hz (277.2),
- D4 - 294 Hz (293.7),
- D#4 - 311 Hz (311.1),
- E4 - 330 Hz (329.6),
- F4 - 349 Hz (349.2),
- F#4 - 370 Hz (370.0),
- G4 - 392 Hz (392.0),
- G#4 - 415 Hz (415.3),
- A4 - 440 Hz (440),
- A#4 - 466 Hz (466.2),
- B4 - 494 Hz (493.9).

Eight-second long sounds were prepared, since the longest musical instrument sounds was below 8 second of length. The mixes were prepared in such a way that for each pair the length of the added sound was truncated to the length of the main sound, and 0.1 s of silence replaced the beginning and the end of the added sound. Next, from the end of the silence at the beginning till 1/3 of the sound length the fade in effect was applied; similarly, fade out was applied from 2/3 of the sound. During mixing, for each instrumental sound chosen to dominate in the mix, the level of the added sound was first re-scaled to match the RMS of the main sound. Thus we assure that the main sound is louder even during transients. Next, three versions of mixes were prepared:

1. with the level of added sounds diminished to 12.5 % of the main sounds,
2. with the level of added sounds diminished to 25 % of the main sounds,
3. with the level of added sounds diminished to 50 % of the main sounds.

In each case, the mix was prepared in such as the average of the added sounds.

Altogether, the training set consisted of 96 singular sounds of musical instruments, and also these same sounds with added noises and harmonic sounds in 3 level versions as described above, i.e. 1152 mixes.

2.3 Testing Data

The data for testing consisted of mixes of instrument sounds only. As in case of the training data, the testing data were prepared in 3 versions, i.e. for the same 3 levels of added sounds. For each subset, the added sound was of the same pitch as the main sound, and was created as the average of the 7 remaining instruments from the training set, modified in amplitude as the sounds added in the training set, and diminished to the desired level. Therefore, we had 3 subsets of the test set, each one consisting of 96 sounds.

3 Experiments and Results

The classification experiments were performed using WEKA software [10]; we chose Support Vector Machine (SMO) classifier, since we have multi-dimensional data for which SVM is suitable, as it aims at finding the hyperplane that best separates observations belonging to different classes in multi-dimensional feature space. Also, SVM classifier was already reported successful in case of musical instrument sound identification [4].

General results of all experiments described in Section 2 are shown in Fig. 1. Detailed confusion matrices for the subsets are shown in Fig. 2, 3 and 4.

The results for training on singular musical instrument sounds are not very high, but after adding mixes to the training sets the results usually improve significantly. When comparing results for various levels of added sounds, we can observe that adding sounds of relatively low level (12.5%) yields significant improvement, but worsens the recognition of violin. For 25.5% no improvement

Training Set	Test Set	Classification Correctness
Singular instrument sounds	Instrument sounds mixes with the level of added sounds diminished to 12,5% volume of the main instrument	66.6667 %
Singular instrument sounds with added artificial sound waves of 12,5% volume		81.25 %
Singular instrument sounds	Instrument sounds mixes with the level of added sounds diminished to 25% volume of the main instrument	88.5417 %
Singular instrument sounds with added artificial sound waves of 25% volume		88.5417 %
Singular instrument sounds	Instrument sounds mixes with the level of added sounds diminished to 50% of the main instrument	66.6667 %
Singular instrument sounds with added artificial sound waves of 50% volume		81.25 %
Singular instrument sounds	All 3 subsets together	73.9583 %
All 3 training subsets together		82.2917 %

Fig. 1. Results of experiments for all training and testing data.

is observed. In case of higher level of added sounds (50%), the improvement of accuracy is significant, and also the confusion between violin and viola, quite visible in all cases, is lower this time.

As we can observe, adding mixes to training sets usually improves the recognition accuracy. However, violin seems to be always more difficult to recognize after adding mixed sounds to each training set, and it is usually mistaken with viola (also when all available data were used - see Fig. 5). On the other hand, sounds of these instruments are very similar, so mistaking these instruments is not surprising. However, correctness of recognizing viola improves after adding mixes to the training set. Viola is also quite often mistaken with cello, but 4th octave is rather high for cello, so such mistakes could have been expected.

The obtained results show quite high capabilities of classifiers trained as shown, because the recognition of the main sound in a mix of sounds of the same pitch is one of the most difficult tasks to perform in multi-timbral environment.

4 Summary and Conclusions

The purpose of this research was to perform experiments on recognizing the dominating instrument in mixes of sustained sounds of the same pitch, assuming harmonic-type spectrum of the test data. This case is the most difficult for classification, since spectra of mixed sounds overlap to high extend. The set-up of experiments for training and testing of classifiers was designed in such a way that a relatively small training data set can be used for learning. Instead of testing all possible pairs for training, we used mixes with artificial sounds with spectra overlapping with the main sounds, i.e. of the same pitch in case

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	11	0	1	0	0	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	1	4	0	7	0	0	0	0
e = oboe	0	3	0	0	4	0	0	5
f = trombone	0	2	2	0	0	8	0	0
g = viola	0	6	0	0	0	0	4	2
h = violin	1	1	0	0	0	0	4	6

a) training on singular musical instrument sounds only

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	10	1	0	0	1	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	1	2	0	8	1	0	0	0
e = oboe	0	0	0	0	12	0	0	0
f = trombone	0	0	0	0	0	12	0	0
g = viola	0	4	0	0	0	0	8	0
h = violin	1	0	0	0	1	0	6	4

b) training on both singular and mixed sounds

Fig. 2. Results of experiments with testing on mixes with added sounds diminished in level to 12.5% of the main sound

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	12	0	0	0	0	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	0	0	0	11	0	0	1	0
e = oboe	0	0	0	0	11	0	1	0
f = trombone	0	0	0	0	0	12	0	0
g = viola	0	4	0	0	0	0	8	0
h = violin	2	0	0	0	0	0	3	7

a) training on singular musical instrument sounds only

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	12	0	0	0	0	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	0	0	0	12	0	0	0	0
e = oboe	0	0	0	0	12	0	0	0
f = trombone	0	0	0	0	0	12	0	0
g = viola	0	3	0	0	0	0	9	0
h = violin	0	0	0	0	1	0	7	4

b) training on both singular and mixed sounds

Fig. 3. Results of experiments with testing on mixes with added sounds diminished in level to 25% of the main sound

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	11	0	1	0	0	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	1	4	0	7	0	0	0	0
e = oboe	0	3	0	0	4	0	0	5
f = trombone	0	2	2	0	0	8	0	0
g = viola	0	6	0	0	0	0	4	2
h = violin	1	1	0	0	0	0	4	6

a) training on singular musical instrument sounds only

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	10	1	0	0	1	0	0	0
b = cello	0	12	0	0	0	0	0	0
c = trumpet	0	0	12	0	0	0	0	0
d = flute	1	2	0	8	1	0	0	0
e = oboe	0	0	0	0	12	0	0	0
f = trombone	0	0	0	0	0	12	0	0
g = viola	0	4	0	0	0	0	8	0
h = violin	1	0	0	0	1	0	6	4

b) training on both singular and mixed sounds

Fig. 4. Results of experiments with testing on mixes with added sounds diminished in level to 50% of the main sound

of added harmonic sounds, or noises. The results show that adding mixes to the training set may yield significant improvement in classification accuracy, although stringed instruments cause difficulties and cello/viola or viola/violin is most common mistake. The recognition of other instruments in most cases improves after adding mixes to the training set.

We plan to continue our research using also percussive instruments; this is one of the reasons why we chose noises for mixes. Also, experiments with training on mixes with artificial sounds are planned for sounds of different pitch, and testing on instrument sounds from outside the training set as well.

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).

The authors would like to express thanks to Xin Zhang from the University of North Carolina at Charlotte for her help with data parameterization. We are also grateful to Zbigniew W. Ras from UNC-Charlotte for fruitful discussions.

References

1. Adobe Systems Incorporated: Adobe Audition 1.0, 2003
2. Aniola, P., Lukasik, E.: JAVA Library for Automatic Musical Instruments Recognition. AES 122 Convention, Vienna, Austria, May 2007

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	34	0	2	0	0	0	0	0
b = cello	0	36	0	0	0	0	0	0
c = trumpet	0	0	36	0	0	0	0	0
d = flute	2	8	0	25	0	0	1	0
e = oboe	0	6	0	0	19	0	1	10
f = trombone	0	4	4	0	0	28	0	0
g = viola	0	16	0	0	0	0	16	4
h = violin	4	2	0	0	0	0	11	19

a) training on singular musical instrument sounds only

Classified as ->	a	b	c	d	e	f	g	h
a = clarinet	32	2	0	0	2	0	0	0
b = cello	0	33	0	0	0	0	3	0
c = trumpet	0	0	36	0	0	0	0	0
d = flute	0	0	0	30	4	0	2	0
e = oboe	0	0	0	0	34	0	2	0
f = trombone	0	0	0	0	0	36	0	0
g = viola	0	8	0	0	0	0	28	0
h = violin	2	0	0	0	3	0	23	8

b) training on both singular and mixed sounds

Fig. 5. Results of experiments with testing on mixes with added sounds diminished in level to 12.5, 25 and 50% of the main sound

- 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
- Herrera, P., Amatriain, X., Batlle, E., Serra X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In *Proc. of International Symposium on Music Information Retrieval ISMIR 2000*, Plymouth, MA
- ISO/IEC JTC1/SC29/WG11: MPEG-7 Overview. (2004) Available at <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>
- 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
- 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)
- Opolko, F., Wapnick, J.: MUMS - McGill University Master Samples. CD's (1987)
- Peeters, G., McAdams, S., Herrera, P.: Instrument Sound Description in the Context of MPEG-7. *Proceedings of the International Computer Music Conference ICMC'2000*, Berlin, Germany, August 2000
- The University of Waikato: Weka Machine Learning Project. Internet, 2007. Available at <http://www.cs.waikato.ac.nz/ml/>
- Wiczorkowska, 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
- Wiczorkowska, A., Kolczyńska, E.: Quality of Musical Instrument Sound Identification for Various Levels of Accompanying Sounds. In: Ras, Z. W., Tsumoto, S., Zighed D. (eds.): *Mining Complex Data, Post-proceedings. LNCS/LNAI 2007*
- Zhang, X.: Cooperative Music Retrieval Based on Automatic Indexing of Music by Instruments and Their Types. Ph.D thesis, Univ. North Carolina, Charlotte 2007