

# Quality of Musical Instrument Sound Identification for Various Levels of Accompanying Sounds

Alicja Wieczorkowska<sup>1</sup> and Elżbieta Kolczyńska<sup>2</sup>

<sup>1</sup> Polish-Japanese Institute of Information Technology,  
Koszykowa 86, 02-008 Warsaw, Poland

`alicja@pjwstk.edu.pl`,

<sup>2</sup> Agricultural University in Lublin  
Akademicka 13, 20-950 Lublin, Poland  
`elzbieta.kolczynska@ar.lublin.pl`

**Abstract.** Research on automatic identification of musical instrument sounds has already been performed through last years, but mainly for monophonic singular sounds. In this paper we work on identification of musical instrument in polyphonic environment, with added accompanying orchestral sounds for the training purposes, and using mixes of two instrument sounds for testing. Four instruments of definite pitch has been used. For training purposes, these sounds were mixed with orchestral recordings of various levels, diminished with respect to the original recording level. The level of sounds added for testing purposes was also diminished with respect to the original recording level, in order to assure that the investigated instrument actually produced the sound dominating in the recording. The experiments have been performed using WEKA classification software.

## 1 Introduction

Recognition of musical instrument sound from audio files is not a new topic and research in this area has already been performed worldwide by various groups of scientists, see for example [2], [3], [4], [6], [7], [9], [11], [14], [21]. This research was mainly performed on singular monophonic sounds, and in this case the recognition is quite successful, with the accuracy level at about 70% for a dozen or more instruments, and exceeding 90% for a few instruments (up to 100% correctness). However, the recognition of instrument, or instruments, in polyphonic recording is much more difficult, especially when no spatial clues are used to locate the sound source and thus facilitate the task [20]. The research has already been performed to separate instruments from polyphonic, poly-tymbral recordings, and to recognize instruments in a noisy environment [6], [13]. The noises added to the recording included noises recorded in the museum (footsteps, talking and clatter), wind gusts, old air-conditioner, and steam factory engine. Therefore, some of the noises were rather unnatural to meet in real recordings. Our idea

was to imitate the sounds found in real recordings, so we decided to use sounds of other musical instruments, or of the orchestra.

In our paper we aim at training classifiers for the purpose of recognition of predominant musical instrument sound, using various sets of training data. We believe that using for training not only clean singular monophonic musical instrument sound samples, but also the sounds with added other accompanying sounds, may improve classification quality. We are interested in checking how various levels of accompanying sounds (distorting the original sound waves) influence correctness of classification of predominant (louder) instrument in mixes containing two instrumental sounds.

## 2 Sound Parameterization

Audio data, for example files of .wav or .snd type, represent a sequence of samples for each recorded channel, where each sample is a digital representation of amplitude of digitized sound. Such sound data are usually parameterized for sound classification purposes, using various features describing temporal, spectral, and spectral-temporal properties of sounds. Features implemented in the worldwide research on musical instrument sound recognition so far include parameters based on DFT, wavelet analysis, MFCC (Mel-Frequency Cepstral Coefficients), MSA (Multidimensional Analysis Scaling) trajectories, and so on [3], [4], [9], [11], [14], [21]. Also, MPEG-7 sound descriptors can be applied [10], although these parameters are not dedicated to recognition of particular instruments in recordings.

We are aware that the choice of the feature vector is important for the success of classification process, and that the results may vary if a different feature vector is used for the training of any classifier. Therefore, we decided to use the feature vector already used in a similar research, which yielded good results for musical instrument identification for monophonic sounds [23]. We applied the following 219 parameters, based mainly on MPEG-7 audio descriptors, and also other parameters used for musical instrument sound identification purposes [23]:

- MPEG-7 audio descriptors [10], [16]:
  - *AudioSpectrumSpread* - a RMS value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame; averaged over all analyzed frames for a given sound;
  - *AudioSpectrumFlatness*,  $flat_1, \dots, flat_{25}$  - describes the flatness property of the power spectrum within a frequency bin; 25 out of 32 frequency bands were used to calculate these parameters for each frame; averaged over all frames for a given sound;
  - *AudioSpectrumCentroid* - computed as power weighted average of the frequency bins in the power spectrum of all the frames in a sound with a Welch method;
  - *AudioSpectrumBasis*:  $basis_1, \dots, basis_{165}$  - spectrum basis function is used to reduce the dimensionality by projecting the spectrum of the analyzed frame from high dimensional space to low dimensional space with

compact salient statistical information; results averaged over all frames of the sound. Spectral basis parameters are calculated for the spectrum basis functions, where 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;

- *HarmonicSpectralCentroid* - the average over the sound duration of the instantaneous Harmonic Centroid within a frame. The instantaneous Harmonic Spectral Centroid is computed as the amplitude (in linear scale) weighted mean of the harmonic peak of the spectrum;
  - *HarmonicSpectralSpread* - the average over the sound duration of the instantaneous harmonic spectral spread of a frame, i.e. the amplitude weighted standard deviation of the harmonic peaks of the spectrum with respect to the instantaneous harmonic spectral centroid;
  - *HarmonicSpectralVariation* - mean value over the sound duration of the instantaneous harmonic spectral variation, i.e. the normalized correlation between the amplitude of the harmonic peaks of two adjacent frames;
  - *HarmonicSpectralDeviation* - the average over the sound duration of the instantaneous harmonic spectral deviation in each frame, i.e. the spectral deviation of the log amplitude components from a global spectral envelope;
  - *LogAttackTime* - the decimal logarithm of the duration from the beginning of the signal to the time when it reaches its maximum or its sustained part, whichever comes first;
  - *TemporalCentroid* - energy weighted mean of the duration of the sound - represents where in time the energy of the sound is focused;
- other audio descriptors:
- *Energy* - average energy of spectrum in the entire sound;
  - *MFCC* - min, max, mean, distance, and standard deviation of the MFCC vector; averaged over all frames of the sound;
  - *ZeroCrossingDensity*, averaged through all frames of the sound;
  - *RollOff* - measure of spectral shape, used in the speech recognition, where it is used to distinguish between voiced and unvoiced speech. The roll-off is defined as the frequency below which an experimentally chosen percentage of the accumulated magnitudes of the spectrum is concentrated; averaged over all frames of the sound;
  - *Flux* - the difference between the magnitude of the FFT points in a given frame and its successive frame (value multiplied by  $10^7$  to comply with WEKA requirements); value averaged over all frames of the sound;
  - *AverageFundamentalFrequency*;
  - *TristimulusParameters*:  $tris_1, \dots, tris_{11}$  - describe the ratio of the amplitude of a harmonic partial to the total harmonic partials (average for the entire sound); attributes based on [17].

Frame-based parameters are represented as average value of the attribute calculated 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 the described research, data from the left channel of stereo sounds were taken for parameterization.

The parameters presented above describe basic spectral, timbral spectral and temporal audio properties, incorporated into the MPEG-7 standard. Also, spectral basis descriptor from MPEG-7 was used. This attribute is actually a non-scalar one - spectral basis is a series of basis functions derived from the singular value decomposition of a normalized power spectrum. Therefore, a few other features were derived from the spectral basis attribute, to avoid too high dimensionality of the feature vector. Other attributes include time-domain and spectrum-domain properties of sound, commonly used in audio research, especially for music data.

### 3 Experiments

The goal of our research was to check how modification (i.e. sound mixing) of the initial audio data, representing musical instrument sounds, influences the quality of classifiers trained to recognize these instruments. The initial data were taken from McGill University CDs, used worldwide in research on music instrument sounds [15]. The sounds were recorded stereo with 44.1 kHz sampling rate, and 16 bit resolution. We have chosen 188 sounds of the following instruments (i.e. representing 4 classes):

1. B-flat clarinet - 37 sound objects,
2. C-trumpet (also trumpet muted, mute Harmon with stem out) - 65 objects,
3. violin vibrato - 42 objects
4. cello vibrato - 43 objects.

The sounds were parameterized as described in the previous section, thus yielding the clean data for further work. Next, the clean data were distorted in such a way that an excerpt from orchestral recording was added. We initially planned to use the recordings of the chords constant in time (for a few seconds, i.e. as long as the singular sounds from the MUMS recordings), but it is not so easy and fast to find such chords. Finally, we decided to use Adagio from Symphony No. 6 in B minor, Op. 74, Pathétique by P. Tchaikovsky for this purpose. Four short excerpts from this symphony were diminished to 10%, 20%, 30%, 40% and 50% of original amplitude, and added to the initial sound data, thus yielding 5 versions of distorted data, used for training of classifiers. Those disturbing data were changing in time, but since the parameterization was performed applying short analysis window, we did not decide to search through numerous recordings for excerpts with stable spectra (i.e. long lasting chords), especially that the main harmonic contents was relatively stable in the chosen excerpts.

For testing purposes, all clean singular sound objects were mixed with the following 4 sounds:

1. C4 sound of c-trumpet,
2. A4 sound of clarinet,
3. D5 sound of violin vibrato
4. G3 sound of cello vibrato,

where A4 = 440 Hz (i.e. MIDI notation is used for pitch). The added sounds represent various pitches and octaves, and various groups of musical instruments of definite pitch: brass, woodwinds, and stringed instruments producing both low and high pitched sounds. The amplitude of these added 4 sounds was diminished to 10% of the original level, to make sure that the recognized instrument is actually the main, dominating sound in the mixed recording.

As one can see, we decided to use different data for training and for recognition purposes. Therefore, we could check how classifiers perform on unseen data.

The experiments performed worldwide on musical instrument sounds, so far, mainly focused on monophonic sounds, and numerous classifiers were used for this purpose. The applied classifiers include Bayes decision rules, K-Nearest Neighbor (k-NN) algorithm, statistical pattern-recognition techniques, neural networks, decision trees, rough set based algorithms, Hidden Markov Models (HMM) and Support Vector Machines (SVM) [1], [4], [5], [7], [8], [9], [11], [12], [14], [19], [22]. The research on musical instrument recognition in polyphonic environment (without spacial clues) is more recent, and so far just a few classifiers were used for the identification of instruments (or separation) from poly-timbral recordings, including Bayesian, decision trees, artificial neural networks and some others [6], [13]. In our experiments, we decided to use WEKA (Waikato Environment for Knowledge Analysis) software for classification purposes, with the following classifiers: Bayesian Network, decision trees (Tree J48), Logistic Regression Model (LRM), and Locally Weighted Learning (LWL) [18]. Standard settings of the classifiers were used. The training of each classifier was performed three-fold, separately for each level of the accompanying orchestral sound (i.e. for 10%, 20%, 30%, 40%, and 50%):

- on clean singular sound data only (singular instrument sounds)
- on both singular and accompanied sound data (i.e. mixed with orchestral recording)
- on accompanied sound data only

In each case, the testing was performed on the data obtained via mixing of the initial clean data with other instrument sound (diminished to 10% of original amplitude), as described above.

Summary of results for all these experiments is presented in tables 1–4.

The improvement of correctness for each classifier, trained on both clean singular sound and accompanied sound data, in comparison with the training on clean singular sound data only, is presented in Figure 1. Negative values indicate

**Table 1.** Results of experiments for Bayesian network

Classifier	Added sound level	Training on data:	Correctness %
BayesNet	10%	Singular sounds only	73,14%
		Both singular and accompanied sounds	81,91%
		Accompanied sounds only	77,53%
	20%	Singular sounds only	73,14%
		Both singular and accompanied sounds	76,20%
		Accompanied sounds only	69,41%
	30%	Singular sounds only	73,14%
		Both singular and accompanied sounds	77,39%
		Accompanied sounds only	63,56%
	40%	Singular sounds only	73,14%
		Both singular and accompanied sounds	75,40%
		Accompanied sounds only	60,77%
50%	Singular sounds only	73,14%	
	Both singular and accompanied sounds	75,93%	
	Accompanied sounds only	55,98%	

**Table 2.** Results of experiments for decision trees (Tree J48)

Classifier	Added sound level	Training on data:	Correctness %
TreeJ48	10%	Singular sounds only	81,65%
		Both singular and accompanied sounds	80,19%
		Accompanied sounds only	74,47%
	20%	Singular sounds only	81,65%
		Both singular and accompanied sounds	82,05%
		Accompanied sounds only	58,78%
	30%	Singular sounds only	81,65%
		Both singular and accompanied sounds	83,64%
		Accompanied sounds only	64,63%
	40%	Singular sounds only	81,65%
		Both singular and accompanied sounds	66,49%
		Accompanied sounds only	62,23%
50%	Singular sounds only	81,65%	
	Both singular and accompanied sounds	76,86%	
	Accompanied sounds only	50,53%	

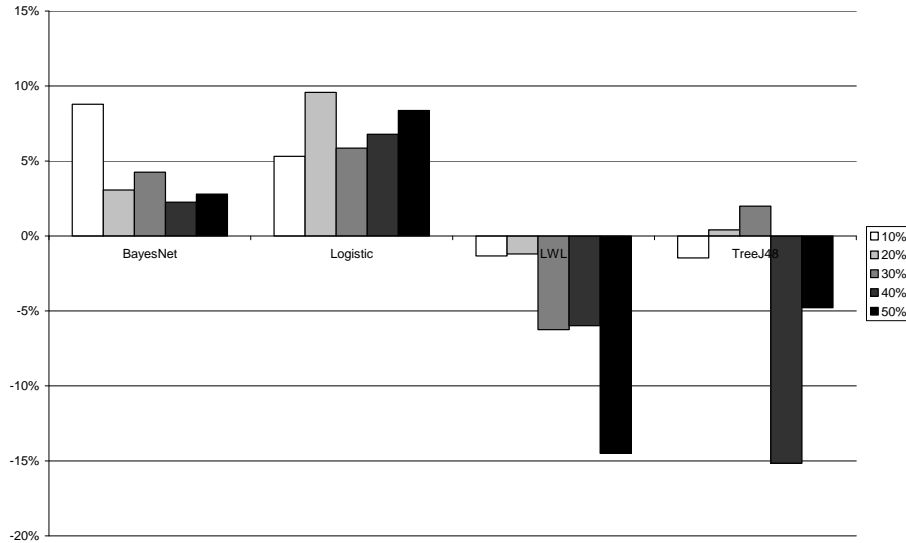
**Table 3.** Results of experiments for Logistic Regression Model

Classifier	Added sound level	Training on data:	Correctness %
Logistic	10%	Singular sounds only	78,99%
		Both singular and accompanied sounds	84,31%
		Accompanied sounds only	67,95%
	20%	Singular sounds only	78,99%
		Both singular and accompanied sounds	88,56%
		Accompanied sounds only	64,23%
	30%	Singular sounds only	78,99%
		Both singular and accompanied sounds	84,84%
		Accompanied sounds only	63,16%
	40%	Singular sounds only	78,99%
		Both singular and accompanied sounds	85,77%
		Accompanied sounds only	53,06%
50%	Singular sounds only	78,99%	
	Both singular and accompanied sounds	87,37%	
	Accompanied sounds only	49,20%	

**Table 4.** Results of experiments for Locally Weighted Learning

Classifier	Added sound level	Training on data:	Correctness %
LWL	10%	Singular sounds only	68,35%
		Both singular and accompanied sounds	67,02%
		Accompanied sounds only	66,62%
	20%	Singular sounds only	68,35%
		Both singular and accompanied sounds	67,15%
		Accompanied sounds only	67,55%
	30%	Singular sounds only	68,35%
		Both singular and accompanied sounds	62,10%
		Accompanied sounds only	62,37%
	40%	Singular sounds only	68,35%
		Both singular and accompanied sounds	62,37%
		Accompanied sounds only	61,70%
50%	Singular sounds only	68,35%	
	Both singular and accompanied sounds	53,86%	
	Accompanied sounds only	53,86%	

decrease of correctness, when the mixes with accompanied sounds were added to the training set.



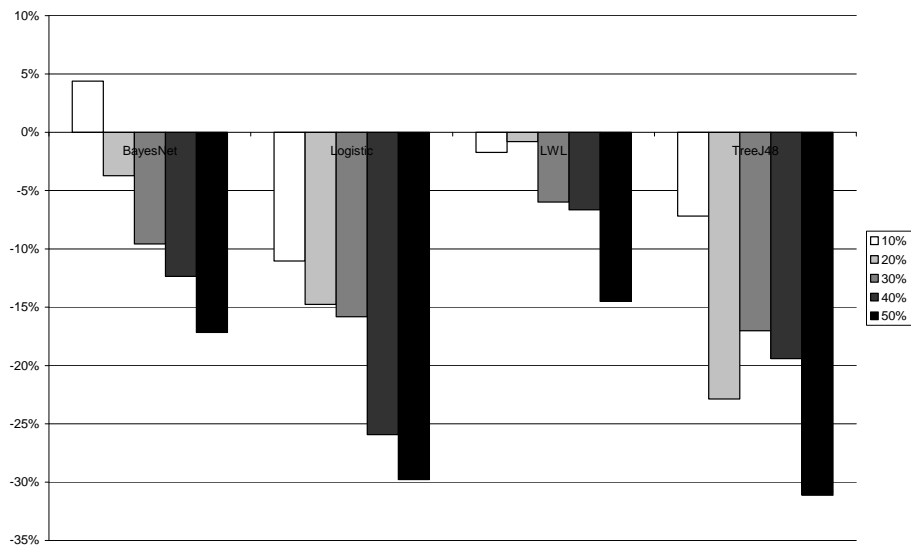
**Fig. 1.** Change of correctness of musical instrument sound recognition for classifiers built on both clean singular musical instrument sound and accompanied sound data, i.e. with added (mixed) orchestral excerpt of various levels (10%, 20%, 30%, 40%, 50% of original amplitude), and tested on the data distorted through adding other instrument sound to the initial clean sound data. Comparison is made with respect to the results obtained for classifiers trained on clean singular sound data only.

As we can see, for LWL classifier adding mixes with the accompanying sounds to the training data always caused decrease of the correctness of the instrument recognition. However, in most other cases (apart from decision trees) we observe improvement of classification correctness, when mixed sound data are added to the training set.

The improvement of correctness for our classifiers, but trained on mixed sound data only, in comparison with the training on clean singular sound data only, is presented in Figure 2.

As we can see, in this case the accuracy of classification almost always decreases, and we only have improvement of correctness for low levels of accompanying sounds for the Bayesian network. Therefore we can conclude that clean singular sound data are rather necessary to train classifiers for instrument recognition purposes.

We are aware that the results may depend on the instruments used, and a different choice of instruments may produce different results. Additionally, the loudness of each sounds (both the sounds of interest and accompanying sounds) changes in time, so it may also obfuscate the results of experiments. Also, decision



**Fig. 2.** Change of correctness of musical instrument sound recognition for classifiers built on the mixed sounds, i.e. of singular instruments with added orchestral excerpt of various levels (10%, 20%, 30%, 40%, 50% of original amplitude), and tested on the data with added other instrument sound to the main instrument sound. Comparison is made with respect to the results obtained for classifiers trained on clean singular sound data only.

trees are not immune to noise in the data, and since the addition of other sounds can be considered as adding noise to the data, the results are not as good as in case of clean monophonic sounds.

When starting these experiments, we hoped to observe some dependencies between the added disturbances (i.e. accompanying sounds) to the training sound data, the level of the disturbance, and change of the classification correctness. As we can see, there are no such clear linear dependencies. On the other hand, the type of the disturbance/accompaniment (for example, its harmonic contents, and how it overlaps with the initial sound) may also influence the results. Also, when sound mixes were produced, both sounds in any mix were changing in time, what is natural and unavoidable in case of music. Therefore, in some frames the disturbing, accompanying sounds could be louder than the sound of interest, so mistakes regarding identification of the dominant instrument in the mix also may happen as well.

## 4 Summary and Conclusions

The experiments described in this paper aimed at observing if (and how) adding disturbance (i.e. accompanying sound added) to the clean musical instrument sound data influences correctness of classification of the instrument, dominating in the polyphonic recording. The clean data represented singular musical instrument sounds of definite pitch and harmonic spectrum. The disturbances added represented various levels of orchestral recordings, added to singular monophonic musical instrument sounds. Tests performed on pairs of instruments sounds have shown that in most cases the use of disturbed (mixed) data, together with initial clean singular sound data, increases the correctness of the classifier, thus increasing its quality. However, no clear linear relationships can be observed. The results for using only distorted data for training showed that clean data are necessary for training purposes.

We plan to continue our experiments, with using various levels of added orchestral sounds for training and for testing the classifiers. Also, since the set of sound features is very important for the correct classification, we plan to check how changes in the feature set influence the quality of classification for distorted in such a way data set.

## 5 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 help with preparing the initial data.

## References

1. Agostini, G., Longari, M., and Pollastri, E.: Musical Instrument Timbres Classification with Spectral Features. *EURASIP Journal on Applied Signal Processing* 1 (2003), 1–11
2. Ando, S. and Yamaguchi, K.: Statistical Study of Spectral Parameters in Musical Instrument Tones. *Journal of the Acoustical Society of America* **94**(1), 1993, 37–45
3. Aniola, P., Lukasik, E.: JAVA Library for Automatic Musical Instruments Recognition. AES 122 Convention, Vienna, Austria, May 2007
4. 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
5. Cosi, P., De Poli, G., and Lauzzana, G.: Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification. *Journal of New Music Research* **23** (1994) 71–98
6. Dziubinski, M., Dalka, P. and Kostek, B.: Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks. *Journal of Intelligent Information Systems*, **24**:2/3 (2005) 133–157
7. Eronen, A.: Comparison of features for musical instrument recognition. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA 2001*
8. Fujinaga, I. and McMillan, K.: Realtime recognition of orchestral instruments. *Proceedings of the International Computer Music Conference, Berlin, Germany, August 2000*, 141–143
9. Herrera, P., Amatriain, X., Batlle, E., and 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
10. ISO/IEC JTC1/SC29/WG11: MPEG-7 Overview. (2004) Available at <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>
11. 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
12. Kostek, B. and Czyzewski, A.: Representing Musical Instrument Sounds for Their Automatic Classification. *Journal of the Audio Engineering Society* **49**(9), 2001, 768–785
13. Lewis, R. A., Zhang, X., and Raś, Z. W.: Blind Signal Separation of Similar Pitches and Instruments in a Noisy Polyphonic Domain. In: F. Esposito, Z. W. Ras, D. Malerba, G. Semeraro (Eds.), *Foundations of Intelligent Systems. 16th International Symposium, ISMIS 2006, Bari, Italy, September 2006, Proceedings. LNAI 4203*, Springer 2006, 228–237
14. Martin, K. D., and Kim, Y. E.: Musical instrument identification: A pattern-recognition approach. 136-th meeting of the Acoustical Society of America, Norfolk, VA (1998)
15. Opolko, F., Wapnick, J.: MUMS - McGill University Master Samples. CD's (1987)
16. Peeters, G., McAdams, S., and Herrera, P.: Instrument Sound Description in the Context of MPEG-7. *Proceedings of the International Computer Music Conference ICMC'2000, Berlin, Germany, August 2000*
17. Pollard, H. F. and Jansson, E. V.: A Tristimulus Method for the Specification of Musical Timbre. *Acustica* **51**, 1982, 162–171

18. The University of Waikato: Weka Machine Learning Project. Internet, 2007. Available at <http://www.cs.waikato.ac.nz/ml/>
19. Toivainen, P.: Optimizing Self-Organizing Timbre Maps: Two Approaches. Joint International Conference, II International Conference on Cognitive Musicology, College of Europe at Brugge, Belgium, 1996, 264–271
20. Viste, H. and Evangelista, G.: Separation of Harmonic Instruments with Overlapping Partial in Multi-Channel Mixtures. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA-03, New Paltz, NY, USA, October 2003
21. Wierzchowska, A.: Towards Musical Data Classification via Wavelet Analysis. In: Foundations of Intelligent Systems, (Eds. Z. W. Ras, S. Ohsuga), Proceedings of ISMIS'00, Charlotte, NC, USA, LNCS/LNAI, No. 1932, Springer-Verlag 2000, 292–300
22. Wierzchowska, A., Wróblewski, J., Synak, P., and Slezak, D.: Application of Temporal Descriptors to Musical Instrument Sound Recognition. *Journal of Intelligent Information Systems* **21**:1 (2003) 71-93
23. Zhang, X. and Ras, Z. W.: Analysis of Sound Features for Music Timbre Recognition. International Conference on Multimedia and Ubiquitous Engineering MUE 2007, 26-28 April 2007, Seoul, Korea. Edited by S. Kim, J. H. Park, N. Pissinou, T. Kim, W. C. Fang, D. Slezak, H. Arabnia, D. Howard. IEEE Computer Society, 3–8