Towards instrument segmentation for music content description: a critical review of instrument classification techniques

Perfecto Herrera, Xavier Amatriain,
Eloi Batlle, Xavier Serra
Audiovisual Institute - Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
{perfecto.herrera, xavier.amatriain, eloi.batlle,
xavier.serra}@iua.upf.es

A system capable of describing the musical content of any kind of soundfile or soundstream, as it is supposed to be done in MPEG7-compliant applications, should provide an account of the different moments where a certain instrument can be listened to. This segmentation according to instrument taxonomies must be solved with different strategies than segmentation according to perceptual features. In this paper we concentrate on reviewing the different techniques that have been so far proposed for automatic classification of musical instruments. Although the ultimate goal should be the segmentation of complex sonic mixtures, it is still far from being solved. Therefore, the practical approach is to reduce the scope of the classification systems to only deal with isolated, and out-of-context, sounds. There is an obvious tradeoff in endorsing this strategy: we gain simplicity and tractability, but we lose contextual and time-dependent cues that can be exploited as relevant features for classifying the sounds.

Classification of monophonic sounds

The following table contains a summary of the techniques and papers we have reviewed.

Method	**Researcher(s)**	**Database size** (sounds/classes)	**Accuracy** (% of success)	**Comments**
K-Nearest Neighbors				Memory intensive/lack of generalization
	Martin & Kim (1998)	1023/14	61-79%	"family" decission previous to class decission
	Fujinaga (1998-2000)	1200/39	68%	real time recognition; GA enhanced technique
	Eronen & Klapuri (1999)	1498/30	75%	mixed architecture; +Gaussian classifier
Bayesian Classifiers				
	Martin & Kim (1998)	1023/14	71%	
	Brown (1999)	30/2	85%	
Discriminant Analysis				Fast computation/post-hoc feature selection
	Herrera (unpublished)	120/8	75%	quadratic discriminant functions
Binary Trees				Quantization of feature values required
	Jensen (1999)	150/5	n/a	
	Wieczorkowska (1999)	n.a./18	68%	
Support Vector Machines				Better generalization than other techniques
	Marques (1999)	estim. 5000/8	70-83%	
Artificial Neural Networks				Very slow training procedure
	Kaminskyj et al. (1995)	240/4	97%	
	Kostek (1995-2000)	n.a. (est. 120)/4	90%	
	Cemgil et al. (1997)	40/10	94-100%	
Higher Order Statistics				
	Dubnov et al. (1997)	n.a./18	n/a	Details not available
Rough Sets				Quantization of feature values needed
	Kostek (1998)	n.a. (est. 120)/4	80%	
	Wieczorkowska (1999)	n.a./18	90%	

Most of them provide success rates higher than 75%, although their processing and memory requirements are quite diverse. Otherwise, a direct comparison of the performance figures in the previous table can be misleading because accuracy rates are sensitive to the database size, to the number of sound classes, to the variability of the training examples and to the type of features used for the classification (although this last issue is not discussed in the paper). Anyway, it seems that there is no method that clearly outperforms the others, and none of them is able to achieve an almost-perfect level. As a consequence, more powerful strategies must be addressed.

Towards classification of sounds in more complex contexts

Although we have found that there are several techniques and features which provide a high percent of success when classifying isolated sounds, it is not clear that they can be applied directly and successfully to the more complex task of segmenting monophonic phrases or complex mixtures. Additionally, many of them would not accomplish the requirements for real-world sound-source recognition systems. Instead of assuming a preliminary source separation stage that facilitates the direct applicability of those algorithms, we are committed with an approach of signal *understanding without separation*. This means that with relatively simple signal-processing and pattern-classification techniques we elaborate judgments about the musical qualities of a signal (hence, describing content). Provided that desideratum, we advance some apparently useful strategies to complement the previously discussed methods, in order to build artificial systems capable of automatic instrument segmentation:

- Content awareness (use metadata when available)
- Context awareness (use local information)
- Use of synchronicities and asynchronicities (co-modulations are important)
- Use of spatial cues
- Use of partial or incomplete cues (a consequence of not needing source separation)
- Use of neglected features (instrument specificities, note transitions…)
- Combining different techniques
- Use of more powerful algorithms for representing sequences of states (for example Hidden Markov Models)

Suggested Readings

Kostek, B. 1999. *Soft computing in acoustics: applications of neural networks, fuzzy logic and rough sets to musical acoustics* Heidelberg: Physica Verlag.

Martin, K. D. 1999 "Sound-Source Recognition: A Theory and Computational Model." Ph.D. thesis. MIT. Cambridge, MA.

Scheirer, E. D. 2000. "Music-Listening Systems." Ph.D. thesis. MIT. Cambridge, MA.