Audio Engineering Society

Convention Paper

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

Automatic labeling of unpitched percussion sounds

Perfecto Herrera[1], Amaury Dehamel[2], Fabien Gouyon[1]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]École Nationale Supérieure des Télécommunications, Paris, France

ABSTRACT

We present a large-scale study on the automatic classification of sounds from percussion instruments. Different subsets of temporal and spectral descriptors (up to 208) are used as features that several learning systems exploit to learn class partitions. More than thirty different classes of acoustic and synthetic instruments and near two-thousand different isolated sounds (i.e. not mixed with other ones) have been tested with ten-fold or holdout cross-validation. The best performance can be achieved with Kernel Density estimation (15% of errors), although *boosted* rule systems yielded similar figures. Multidimensional scaling of the classes provides a graphical and conceptual representation of the relationships between sound classes, and facilitates the explanation of some types of errors. We also explore several options to expand the sound descriptors beyond the class label, as for example the manufacturer-model label and confirm the feasibility of doing that. We finally discuss methodological issues regarding the generalization capabilities of usual experiments that have been done in this area.

1. INTRODUCTION

Classification is one of the processes involved in audio content description. Audio signals can be classified according to miscellaneous criteria. A broad partition into speech, music, sound effects (or noises), and their binary and ternary combinations is used for video soundtrack descriptions. Sound category classification schemes for this type of materials have been recently developed [1] and facilities for describing sound effects have even been provided in the MPEG-7 standard [2]. Usually, music streams are broadly classified according to genre, player, mood, or instrumentation, and some research has been devoted in order to automatically assign some of these labels to sound files [3], [4]. Automatic labeling of instrument sounds has some obvious applications for enhancing the operating systems of sampling and synthesis devices, in order to help sound designers to categorize (or suggesting names for) new patches and samples. Automatic annotation of the instrumentation played in a given musical recording is one of the long-term goals for music content description systems. One shorter-term goal is that of describing drum loops at the rhythmic level because, in them, we find a combination of almost isolated sounds with simpler 2- or 3-sounds mixtures. This paper can be considered as a follow-up of [5] and [6], and also as an attempt to further explore our previous approach in order to handle more classes (and more difficult class boundaries too). It can be also considered as groundwork for dealing with automatic labeling of drum loops, which will be presented in a forthcoming report.

Previous research in automatic classification of sounds from music instruments has focused in instruments with definite pitch. Classification of string and wind instrument sounds has been attempted using different techniques and features, yielding to varying degrees of success (see [7] for an comprehensive review). Classification of percussive instruments, on the other hand, has attracted little interest from researchers. In one of those above cited studies with pitched sounds, Kaminskyj [8] included three pitched percussive categories (glockenspiel, xylophone and marimba) and obtained good classification results (ranging from 75% to 100%) with a K-NN algorithm. Schloss [9] classified the stroke type of congas using relative energy from selected portions of the spectrum. He was able to differentiate between high-low sounds and open, muffled, slap and bass sounds. Using a K-means clustering algorithm, Bilmes [10] also was able to differentiate between sounds of three different congas. McDonald [11] used spectral centroid trajectories as classificatory features of sounds from percussive instruments. Sillanpää [12] used a representation of spectral shape for identification of the basic five categories of a drum kit: bass drum, snares, toms, hi-hats, and cymbals. His research was oriented towards transcription of rhythm tracks and therefore he additionally considered the case of identification of several simultaneous sounds. A database of 128 sounds was identified with 87% of accuracy for the case of isolated sounds. Performance dramatically dropped when there were two or three simultaneous sounds (respectively 49% and 8% for complete identification, though at least one of the sounds in the mixture was correctly identified all the times). In a subsequent study [13], the classification method used energy, Bark-frequency and log-time resolution spectrograms, and a fuzzy clustering of the original feature vectors into four clusters for each sound class. Weighted RMS-error fitting and an iterative spectral subtraction of models was used to match the test sounds against learnt models. Unfortunately, no systematic evaluation was presented that time. Goto and Murakoa [14] also studied drum sound classification in the context of source separation and beat tracking [15]. They implemented an “energy profile”-based snare-kick discriminator, though no effectiveness evaluation was provided. More recently, Zils et al. [16], using a technique of analysis and incremental refinement of synthesis that was originally developed by Gouyon [17], reported very good performance rate of identifying kicks and snares in songs. As a general criticism, in the previous research there is a lack of systematic evaluation of the different factors involved in automatic classification, and the databases are small to draw robust conclusions.

From another area of studies, those focusing on characteristics of beaten objects, it seems that information about the way an object is hit is conveyed by the attack segment, whereas the decay or release segment conveys information about the shape and material of the beaten object [18]. Repp [19] found that different hand-clapping styles (palm-to-palm versus fingers-to-palm) correlated with different spectral envelope profiles. Freed [20] observed that the attack segment conveyed enough information for the subjects to evaluate the hardness of a mallet hit. Four features were identified as relevant for this information: energy, spectral slope, spectral centroid and the time-weighted average centroid of the spectrum. Kaltzky et al. [21] have got experimental results supporting the main importance of the decay part (specifically the decay rate) of a contact sound in order to identify the material of the beaten object.

Besides all the previous research, in the last two years we have witnessed some promising but still quite inconclusive works. Orife [22] has presented a tool for extracting rhythmic information from loops and songs using Independent Subspace Analysis (ISA), a technique for source separation [23]. Unfortunately no evaluation of the system has been included in the available report. Fitzgerald et al [24], also using ISA but with some additional some sub-band pre-processing, reported a success rate of 89.5 % when transcribing a database of 15 drum loops containing snare, kick and hi-hats. Nevertheless the identification is done on the de-mixed tracks by a human listener, not by an automatic classification algorithm. Finally, Riskedal [25] presented a system that combined a drumloop-adapted onset detection procedure (taken from Klapuri [26]) with an adaptation of an Independent Component Analysis [27], another source separation technique. Though the results seem promising, only a few examples are presented by the author. Jørgensen [28] attempted to use cross-correlation between sound templates extracted from isolated sound recordings and realistic drum-kit recordings. Using this technique only kicks and snares seem to be detected with some reliability. A very different motivation has been that of Kragtwijk et al. [29] who have presented a 3D virtual drummer that re-creates with synthetic images the playing movements of a real drummer after analyzing the audio input coming from a real performance. Unfortunately, the audio analysis part of the system is still under development.

In the next sections we will present the method and results of different but complementary studies on automatic identification of unpitched percussion sounds[1] that do not have a well-defined fundamental frequency. First we will discuss the features we initially selected for the task and the ways for obtaining the smallest set without compromising classification effectiveness. Then we will summarize the class induction techniques we have been experimenting with and finally, in the main section, four studies will be presented, covering the automatic labeling of a large amount of percussion instrument sounds, the identification of similarities, differences and relationships between sound classes, some preliminary attempts of describing sounds beyond the class label, and validation through holdout databases. As a conceptual guide, the following questions will be discussed:

- Which features are the most appropriate for the task, and what can be learned about the sounds from the information conveyed by the features?
- Is there any classification algorithm that excels in the task; which are the tradeoffs of selecting one or another?
- Is it possible to detect some specificities of sounds such as if they are electronic, or if they come from a given manufacturer of acoustic instruments?
- Is it possible to derive some structural representation expressing relationships of similarity between classes of sounds? Which classes can be considered as "inter-related"?

2. METHOD

2.1. Selection of sounds

We have setup different databases in order to properly tackle with the above mentioned problems. Databases were populated using sounds that were drawn from different commercial sample CD's and CD-ROMs and also downloaded from internet providers of sound samples[2]. We cared for disregarding highly processed sounds (i.e. with lots of compression, distortion, reverberation, flanging, etc.) in order to preserve the acoustic naturalness of the classes to be labeled. Different dynamics and different physical instruments (i.e. different recording conditions, different manufacturers and models) were also looked for. Databases sizes ranged from 500 to almost 2000 sounds.

2.2. Descriptors

We considered descriptors or features belonging to different categories: Mel-frequency Cepstral coefficients, Bark Energy Ratios (BER), descriptors derived from BER, spectral descriptors, temporal descriptors, and descriptors derived from log-transforming some of the previous ones. Up to 207 descriptors per soundfile have been tested in some of the studies that will be presented. The analyses used a window size of 512 samples and a Hamming window without overlap except for the BER computation where it was a Hanning with 50% overlap and for the MFCCs where the window was of 256 samples.

2.2.1. Mel-Frequency Cepstrum Coefficients

MFCC's have been usually used for speech processing applications, though they have shown usefulness in music applications too [30]. As they can be used as a compact and robust representation of the spectral envelope, their variance was also recorded in order to keep some time-varying information. 13 MFCC's were computed over the whole signal, and their means and variances were used as descriptors.

2.2.2. Bark Energy Ratios

Bark bands are approximations to the first 24 critical bands of human hearing. We have included an alteration of the original definition of them consisting in using two more bands in order to extend low-frequency resolution, as the original bandwidth of the first two Barks was not enough for good

[1] We have left out of this study those percussion instruments that produce "notes" (i.e. scales of sounds with well defined fundamental frequencies) such as marimba, vibraphone, xylophone, tubular bells, or celesta. They will be the matter for a forthcoming report.

[2] http://www.sonomic.com, http://www.primesounds.com

discrimination between some instruments. We have then computed the proportion of energy in each of these 26 bands, being the lowest four a result of half-splitting the bandwidth of original Barks 1 and 2. In addition to the energy proportions and the variances, another set of descriptors have been derived from them. Some of these descriptors, as far as we know, have not been previously used for instrument sound classification.

2.2.3. Bark-derived descriptors

Information in Bark Energy Ratios is a bit redundant as usually there are contiguous bands that have correlated values. Using the information that is present in the BERs raw data, we can compute some relational and summarizing descriptors. It is believed that they can reveal specificities that are not evident from direct inspection of raw BERs. We have computed the following: band with the maximum of energy, band with the minimum of energy, band with the maximum energy variance, band with the minimum energy variance, proportion of energy in the band with the maximum energy, proportion of energy in the band with the minimum energy, a ratio between high- and low-frequency bands, a ratio between mid- and low-frequency bands, a ratio between mid- and high-frequency bands, energy difference between the first and the 26th band, the overall standard deviation of the energy proportions and of the energy variances. Additionally, the energy profile and the variance profile across all the bands can be roughly adjusted to a linear function and therefore we have used their slopes as descriptors.

2.2.4. Spectral descriptors

We have included some descriptors that are computed after an FFT and which describe the spectral envelope in different ways than the MFCCs do. They are:

- Spectral flatness: the ratio between the geometrical mean and the arithmetical mean of the spectrum (this gives an idea of the shape of the spectrum, if it's flat, the sound is more "white-noise"-like; if flatness is low, it will be more "musical");
- Spectral centroid: the centre of gravity of the spectrum;
- Spectral skewness: the 3rd order central moment (it gives indication about the shape of the spectrum in the sense that asymmetrical spectra tend to have large skewness values).
- Spectral kurtosis: the 4th order central moment (it gives clues about the shape of the spectrum: "peaky" spectra have larger kurtosis than scattered or outlier-prone spectra);
- Strong peak: it is an indicator of the presence of a very pronounced peak. The thinner and the higher the maximum of the spectrum is, the higher value;
- Spectral crossings: an approximation of the number of prominent spectral peaks.

2.2.5. Temporal descriptors

- Zero crossing rate (ZCR): the times that the waveform changes from positive to negative values.
- Strong decay: a feature built from the non-linear combination of the energy and temporal centroid of a frame (a frame containing a temporal centroid near its left boundary and strong energy is said to have a "strong decay");
- Variances of the ZCR and of the spectral centroid.
- We should also consider as temporal descriptors the variances of the MFCCs and the variances of the BERs.

2.2.6. Transformed descriptors

We additionally are interested in studying the effects of nonlinear transformations to the original data, as we observed that most of our descriptors were distributed in non-Gaussian ways. For certain classification techniques this could not be important, but there are some of them that assume this distribution for the data (for example, discriminant analysis), and conclusions can be inappropriate when this requirement is not held. Usually a log transformation converts a peaky distribution into another that is more Gaussian-like. Other transformations such as xxx have been recommended but are not explored in this paper. To summarize, we have taken the natural logarithm of the BERs, of the spectral descriptors, and of the temporal descriptors except the MFCCs variances.

2.3. Selection of relevant descriptors

Using very large sets of features for building a classification model is usually to be discouraged. First, some of them can be redundant or irrelevant, and the computational cost for keeping them might be high. Second, some of them can be misleading or inconsistent regarding the task. In this case, the cost is not only in terms of computation time but also in terms of task performance (as the error rate will be higher). Besides all that, interpreting a model containing a large set of features can be very difficult or even impossible.

There are three different strategies in order to use an optimal or near-optimal number of features for a classification task:

- *Embedding* has the feature selection stage intertwined with the classification algorithm, as it is the case with decision trees, or with discriminant analysis
- *Filtering* decouples the feature selection process from the model learning process; there are lots of feature selection techniques that can be applied before trying to build a model for the sound classes. After some literature review and unsystematic testing, we have selected CFS as our main filter.
- *Wrapping* makes an evaluation of features that is in intimate connection with the induction process; theoretically this strategy should be the best [31], but the price to pay is the computation time needed.

In addition to the way of connecting the feature selection and the induction process, we must care about the evaluation criterion. There are different criteria, each one leading to a bunch of different selection techniques. In our previous study we compared CFS, ReliefF and an F statistic and concluded that CFS was consistently better.

Here we have compared the selection made by CFS alone against a combination of CFS plus a wrapper. As the wrapping was attempted after an important but reasonable reduction of the number of features (almost an order of three), the time that was needed by the wrapper was acceptable (less than 24 hours on a Pentium-III 800MHz. CPU).

2.4. Classification techniques

Deciding the way to train a system to learn the automatic assignment of labels to classes of sounds depends on several factors: the available techniques, their performance in terms of error or success rates, the understandability of the class representations, and the computational complexity (in terms of memory and processing requirements) that can be assumed. There are some situations where it seems reasonable to sacrifice some performance efficiency in order to achieve clear interpretations of the class decisions (for example, when doing basic research); on the other hand, if the system is intended to be implemented inside a commercial application, a tradeoff between computational complexity and performance should be achieved (without any consideration of model understandability). In the following studies we have worked with quite different techniques, from lazy leaming ones (k-Nearest Neighbors) to neural networks, with non-parametric ones (kernel density estimation) to parametric ones (canonical discriminant analysis). We have decided to skip some of them in order to keep the discussion better focused, but let's say some generalities on the ones that will appear in the results section.

The *K-Nearest Neighbors* (K-NN) technique is one of the most popular instance-based learning techniques, and there are several papers on musical instrument sound classification using K-NN [32], [33], [34], [8]. These techniques are sometimes called *lazy* because, instead of computing some abstract model for partitioning the observation space they store all the observations in memory, and decide the class of a new instance by looking around its neighbors.

Another technique that does not generate a truly abstract model of the data is that of *Kernel Density estimation*. It constructs an approximation to the distribution function of data by placing a "bump" function (sometimes it is a Gaussian, but can have other shapes) at each data point and then summing them up. The "bump" function, i.e. the Kernel has a bandwidth that affects the smoothness of the approximation. In the implementation that we have used it has been automatically selected.

Canonical discriminant analysis is a statistical modeling technique that classifies new examples after deriving a set of orthogonal linear functions that partition the observation space into regions with the class centroids separated as far as possible, but keeping the variance of the classes as low as possible. It can be considered like an ANOVA (or MANOVA) that instead of continuous to-be-predicted variables uses discrete (categorical) variables. After a successful discriminant function analysis, "important" variables can be detected because it embeds some relevance testing (F statistic). Discriminant analysis has been successfully used by [35] for classification of wind and string instruments.

C4.5 [36] is a decision tree technique that tries to focus on relevant features and ignores irrelevant ones in order to partition the original set of instances into subsets with a strong majority of one of the classes. Decision trees, in general, have been pervasively used for different machine learning and classification tasks. Jensen and Arnspang [37] or Wieczorkowska [38] have used decision trees for musical instrument classification. An interesting variant of C4.5, that we have also tested, is PART (acronym for *partial decision trees*) [39]. It yields association rules between descriptors and classes by recursively selecting a class and finding a rule that "covers" as many instances as possible of it. The models derived by PART usually contain fewer rules than those generated by C4.5, and are easier to interpret.

2.5. Cross-validation

For the forthcoming experiments the usual ten-fold procedure was followed: 10 subsets containing a 90% of the sounds were randomly selected for learning or building the models, and the remaining 10% was kept for testing them. Hit-rates presented below have been computed as the average value for the ten runs.

3. RESULTS

3.1. Study 1: Classification of unpitched percussion sounds

A database containing 1976 sounds from 33 different classes was used in this study. Categories included acoustic and synthetic (or "electronic") sounds. Some of this synthetic sounds were drum-machine specific (e.g. Roland TR-808 cymbal and kick, or Roland TR-909 kick). Table 1 details the distribution of sounds across classes.

	acoustic	synthetic	Total
Bongo	100	43	143
Clap	21	50	71
Clave	85	0	85
Conga	100	75	175
Cowbell	100	0	100
Crash	50	35	85
Cymbal	0	19	19
HiHat	100	100	200
Kick	50	100	150
Ride	50	20	70
Shaker	100	50	150
SideStick	85	0	85
Snare	50	50	100
Tabla	115	0	115
Tambourine	75	40	115
Timbale	100	0	100
TomHi	45	0	45
TomLo	48	0	48
TomMe	80	0	80
Triangle	40	0	40
Total	1394	582	1976

Table 1. Composition of the database of unpitched percussive sounds

Table 2 summarizes the main results. We have first tested a set of 89 descriptors that were also used in our previous studies. Regarding selection of descriptors, using CFS as a filter yielded a best set of 38 descriptors, instead of the initial 89. The original best performance was that of the Kernel Density estimator at 83.4. The performance of any algorithm has not been significantly affected after applying this important filtering. We can even improve it by using a combination of CFS-filtering followed by a wrapping refinement, as the feature set is further reduced to 31 descriptors but the hit rate remains at 83.2 for the best case, which is again the Kernel Density estimation.

	Orig. set (89)	CFS (38)	CFS Wrap (31)	Log (89)	CFS log (40)
Kernel Density	83.4	83.1	83.2	83.5	**85.7**
IB1	80.4	81.5	82	84.5	85.3
C4.5	70	70.7	67.7	68.1	68.7
PART	68.2	67.7	67.3	69.2	70.7
PART-ADABoost	82.1	82.8	n.a.		

Table 2. Hit rates for different learning algorithms (rows) and different feature selection strategies (columns)

An interesting result is the degraded performance that we have obtained using decision-tree related techniques such as C4.5 or PART compared to instance-based ones. As those techniques provide the clearest conceptual models (and in our previous work on drum sounds we had not observed such dramatic differences), we insisted a bit more before abandoning them. Boosting [40] is a technique that has proved very useful for improving the performance of weak learners, as it can be the case of our PART algorithm. It has even been used for audio classification by Guo et al. [41]. The key point of boosting is that of giving more weight to difficult cases and less weight to instances that are easy to be classified. A boosted system can be considered as an "ensemble of classifiers" which concentrates on the most difficult examples, trying to reduce their classification errors. As can be seen, in this case the ADABoostM1 algorithm [42] combined with the PART one achieved a performance that is comparable to the Kernel Density or 1-NN methods. Boosting nevertheless has some costs: the first one is an important increase in the time needed for building and validating a set of rules; the second one is the increasing complexity of the models, as several sets of rules, each one with a given weight, have to be stored for classification of new instances. After some experimentation, the optimum number of sets was found to be 30 (then amounting around 3000 simple rules, instead of the 140 found by the original PART model).

In this report we have introduced "new" descriptors (the BER related set –see section 2.2.3- and the log transformed set –see section 2.2.6-). As we have set up several conceptually different subsets of descriptors, it could be interesting to compare them, which is presented in table 3. First of all, we can see that a generic spectral descriptors subset (including those from section 2.2.4 plus BER and MFCCs, but not their variances) is more effective (7%) than the temporal subset. Secondly, MFCCs (including variances) perform at similar rates than the whole spectral subset (which includes the MFCCs but not their variances) and both achieve the best overall performance. A third interesting remark is that log-transformed BERs yield 10% better results than raw BERs (or even 15% better results if CFS feature selection is used). It is also worth to note that CFS filtering degrades around 2% the performance of all subsets but the MFCCs and the logBERs.

Given the available data, we have finally set up an "alternative" set of descriptors that swaps all of them but the MFCCs and MFCCs variances by the log-transformed ones. Performance for this new set is shown in the last two columns of table 2, where a slight improvement over the original raw data set performance is evident (using or not using CFS filtering). Adding the BER derived set did not improved the performance and therefore they were discarded for any other analysis.

Feature subsets	All initial descriptors	CFS filtered
All	83.4 (89)	83.1(38)
All Spectral	80.1(46)	78.7(32)
Temporal	73.4(43)	71.3 (21)
MFCC's	78.9(26)	79.4 (14)
Spectral (see section 2.2.4)	59.7(7)	59.7(7)
BERs	66.6(52)	62.19(31)
BERDerived	65.2(29)	63.3(15)
LogBER	76.4(26)	77.1(31)
LogSpectral	62.4(7)	62.4(7)
Log AllSpectral	**84**(46)	**83.1** (32)

Table 3. Hit rates for different "conceptual" subsets of features. Their cardinality is indicated inside parentheses. See text for explanation

Automatic labeling of 33 different sound classes of unpitched percussion instruments seems to be a difficult task, but fortunately it is not intractable. Previous studies using similar amount of classes of string and wind instruments' sounds, and similar database size have reported performance figures that range from the surprisingly low 35% reported in [43] to the 82% and 92% respectively reported in [8] and [35]. In our case, any strategy based on a "bulk" labeling has not achieved better performance than 83% of correctly classified instances. As confusion matrices reveal, there are different types of errors: confusion between acoustic and electronic sources for the same "class" (e.g. bongo and electronic bongo), which can be sometimes tolerated, and confusion between different classes, being the most prominent those occurring between bongo, conga, tabla and timbale (which, on the other hand could be also expected given the sonic similarities that they show). Electronic bongo and electronic crash seem to be the most difficult classes, whereas bongo (acoustic and electronic), conga, acoustic shaker, and timbale have been the least reliable classes (i.e. other sounds are incorrectly assigned to them).

Hierarchical approaches have been proposed for the classification of string and wind sounds by [44] or, [8], for example, and it has been shown that in some cases they can improve the performance of classification systems (for example by first labeling a sound as "containing or not containing vibrato" or as "pizzicato or continuous sound", and then assigning the proper class specific label). In our case, it is clear that if we could sort out the electronic sounds from the acoustic ones, the first type of confusions, which represent a 22% of the errors, could be reduced (hence improving a 2% the hit rate). Regarding the other source of errors, we do not have yet a clear strategy to alleviate them apart from including better descriptors that are capable of differentiating the nuances of one or of another class.

Another interesting finding is that drum-machine specific sounds are labeled with high precision (even without confusions even between two similar models, the TR-808 and the TR-909 kicks), as they usually are quite idiosyncratic. This type of discrimination, if possible, paves the way for detailed descriptions that go beyond taxonomic labels. We will return on that in the Study 3.

3.2. Study 2: similarities and differences between instrument classes

We have approached the relationships between the different classes of instruments using a strategy that could be termed "synthetic", as it will provide an overall summary about them. we have worked with a technique that has been frequently used for representing the "mental" relationships that listeners have regarding different timbres: Multidimensional Scaling (MDS) [45],[46],[47],[48]. The purpose of MDS is to identify and model the structure and dimensions of a set of objects from the dissimilarity data observed among them. This is accomplished by assigning observations to specific locations in an abstract space (usually two- or three-dimensional) such that the distances between points in the space match the given dissimilarities as closely as possible. A three dimensional perceptual space for percussive instruments (not including bells) has been hypothesized by Lakatos [49]. This percussive perceptual space spans three related physical dimensions: log-attack time, spectral centroid and temporal centroid. Additional evidence supporting them has been gathered during the multimedia content description format standardization process (MPEG-7) and, consequently, they have been included in MPEG-7 as descriptors for timbres [50].

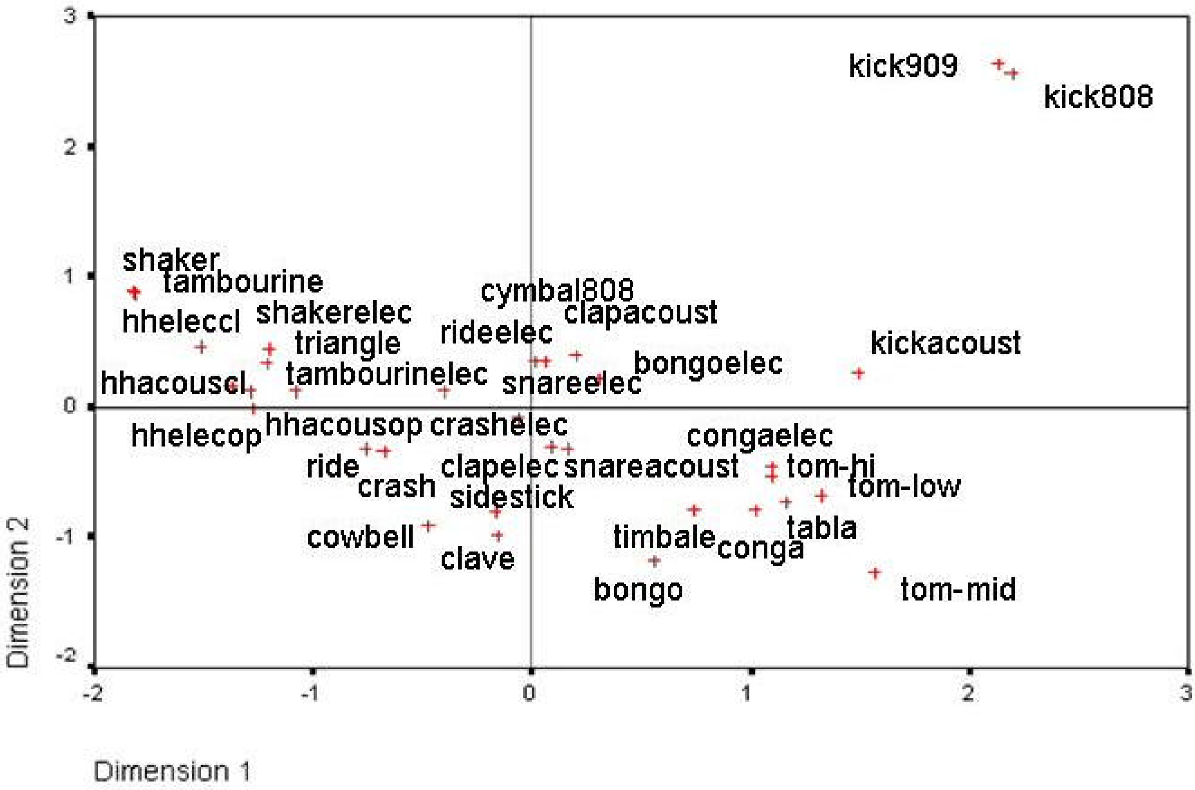

Figure 1. Multidimensional Scaling of the unpitched percussion sound classes

Here we are not concerned with cognitive representations but only with abstract "taxonomic" or "normative" representations that can be derived from the low-level signal descriptors that we have computed. Instead of computing dis-similarities or distances between sounds directly from those data, as it would be computationally very expensive, we have considered as index of similarity between classes the ratio computed between the deviation of each pair of class means from the grand total mean (these deviations are squared, multiplied by the number of sounds in each class and summed). This is also known as the *between-groups F-matrix* in the field of multidimensional data analysis.

Figure 1 shows the 2-dimensional map we have obtained after processing the matrix with an ALSCAL algorithm of MDS. Along Dimension 1 (horizontal axis) we may track the "plates" on the left side and the "skins" on the right side. Dimension 2 (vertical axis) could be considered as a representation of the "acoustic-synthetic" origin of sounds, as in the upper region most of the classes (but not all) correspond to synthetic sounds whereas the lower region is mainly occupied by the acoustic sounds. Another interesting finding is the presence of "clusters" of instrument classes that show similar relatedness that was found after analyzing the errors in the classification models. For example, bongo, conga, timbale and tabla form one of such clusters, hi-hats are all quite close of each other, ride and crash, cymbal808 and electronic ride might form the main plates cluster, and both synthetic kicks cluster in the upper-right corner of the map. Idiophones that are not plates or skins appear in what could be considered as "logical" places: shaker and triangle occupy regions that are close to those of hi-hats and tambourines, claps appear close to snares, and finally, clave, sidestick and cowbell occupy an "outside" region that is between ride and crash –being the cowbell the closest one to them-, but also to snares and even to the conga-tabla-timbale cluster –being the sidestick and clave the closest to them-.

Another interesting approach that we regrettably have left out of this paper is what could be termed the "analytic" strategy, as it would provide specific information about where and how the different instrument classes differ (i.e. which descriptors define a given class). For this kind of analysis, the sets of rules yielded by PART are ideal, though they are not the best strategy for achieving optimal classification.

3.3. Study 3: Identification of manufacturers or models

We have also inquired about the possibility of being able to automatically describe sounds beyond the usual taxonomical labels. One of those possibilities is that of telling if, for example, a given cymbal sound has been produced by a Yamaha or by a Paiste instrument. In the previous studies 1 it has been shown that it is possible to tell if a sound belongs to some peculiar drum machines as the Roland model 808 or model 909. Here we will approach for our first time the automatic labeling of the manufacturer and model of instruments belonging to pop-rock acoustic drum kits.

A database built from the so-called *Absolute* collection[3] was used for this study. As a given instrument (we will use the word "instrument" to denote a combination of manufacturer and model, not a taxonomic class) was recorded using different microphones and miking techniques[4], we can be sure that what our systems learn are some "specific instrument or manufacturer-model signature" and not the "microphone signature". A couple of methodological drawbacks can be mentioned here, though. The first one is the low variability of the available examples, compared to that found in other databases we have used in this report. The second one is that a much more rigorous testing would imply matching, for example, the diameters of the skins or plates for a given class. Anyway, what this study shows is that we can describe instrumental sounds beyond the class label to some more detailed and idiosyncratic level (as could the manufacturer-model or at least, telling if two hits correspond to the same instrument).

In order to model the manufacturer name decision, we have performed separated Canonical Discriminant Analyses for each subset of data corresponding to each instrument class (cymbals were not included because in the used database there was a manufacturer not well enough represented). Six manufacturers (i.e. *Ludwig, Pearl, Rogers, Sonor, Tama, Yamaha*) were used for snares, kicks and toms, whereas only two were used for hihats (i.e *Paiste* and *Zildjian*). Selection of features was done using the F criteria and a forward search, as it is usual with this technique. Retained features differed significantly depending on the instrument class to be considered. Labeling of snare manufacturers required more than two thirds of the available descriptors whereas for the rest of instruments a half of them was enough. Most relevant features usually belonged to the spectral and MFCC's categories, but some instrument-related specific Bark bands were always included (e.g. lowest bands for kick, highest ones for hihat). It is interesting to notice that temporal descriptors (i.e. variances or ZCR) were absent or got very low relevance ratings except for the case of snares manufacturers. Table 4 summarizes the results and shows that yielding this prediction seems to be an easy task for a linear system.

Kick	Snare	Tom	HiHat
96.1	86.3	94.1	99

Table 4. Overall performance of manufacturer labeling for the indicated classes (the learning criteria is the manufacturer name, not the instrument class)

3.4. Study 4: Generalization across different databases

Up until now we have been testing the generalization capabilities of our results by means of the 10-fold cross-validation procedure. Even in the case of including sounds corresponding to different

[3] http://www.drums.sk

[4] Personal communication from the comercial collection producer

recording conditions, as we have done, there exists the possibility of over-fitting or extremely biasing the learning models. In order to evaluate this possibility, we have set up two different databases (S and A), then we have set up some models using database S, and finally we have tested them using database A, and the other way around, in a cross-holdout procedure.

Database S sounds were selected from an internet-based sample-provider (see note 2). This database included 950 acoustic and electronic sounds. Database A, on the other hand, contained 589 acoustic sounds, and was built upon the *Absolute* collection (see section 3.3). Therefore, database A contained the most homogeneous (i.e. less intra-class variance) data we had available. All classification decisions were done using six classes (kick, snare, tom, hihat, ride, and crash), discarding any distinction between electronic or acoustic origins. Table 5 summarizes the most important data. In all cases values were obtained using a kernel density estimator as the learning technique. The diagonal shows the best results when using 10-fold cross-validation (i.e. a database is validated using data from the same database). As it was expected hit rates were very high and the highest one is for the most homogeneous database. The other two cells quantify somehow the generalization power of the tests: using the most heterogeneous database (S) for the learning phase we only lose an 8% when classifying the other one (A), whereas using the least one leads to an 18% loss when trying to classify the former (in both cases we compare data against the 90% that was achieved with 10-fold CV). As an additional control test, a third database, called D, was used to test the models derived from S and from A. D was a database intended to be "comparable" to S in terms of the variability of sounds and number of classes, though it only included 570 examples. Learning from S and testing D yielded similar results than the CV of S on its own (87%), whereas learning from A and then testing D yielded only 80% of hits. On the other hand, learning from D and using S for testing yielded 82% of hits.

The main conclusion to be drawn from this study is that a careful selection of the database examples is needed if we have strong concerns regarding the generalization capabilities of the learnt models. If the database is not good (i.e., and including a large number of instances per class –let's say, more than 100- and therefore incorporating enough variability for each one of the classes), our error rates can be even a 23% over-optimistic (95% versus 72%), whereas if the learning database has been carefully selected, we can expect that our generalization errors are not larger than 8% (90% versus 82%). Unfortunately, most of the research that has been done in the field of classification of instrument sounds has used only one sample-supplier (i.e. very reduced variability in terms of recording conditions, and number of physical instruments recorded), and has performed the validation using sounds coming from the same sample provider[5].

Learning ->	S	A
Testing		
S	90	72
A	82	95

Table 5. Overall best results with using different holdout validation conditions.

4. CONCLUSIONS AND FURTHER WORK

Automatic labelling of percussion sounds is a problem that deserves attention. We have presented here a summary of our current efforts on the subject, and have also discussed some methodological approaches. We have devised a set of descriptors that allows acceptable error rates and cover relevant spectral and temporal information. Anyway, we think that the current descriptor set could be sensibly improved by incorporating more and better temporal information such as the temporal centroid, the attack time, MFCC's deltas, or inter-Bark correlations.

It has been demonstrated that the task is feasible for different types of learning systems though the error rates found, ranging from 10% to 20%, leave some room for improvement. Apart from that of using hierarchical classification systems, as discussed in section 3.1, one promising area for overcoming the current limitations is that of working with ensembles of descriptors and of classifiers. The traditional approach to these types of tasks has been using only one set of features with one specific induction algorithm to learn to partition the space defined by the examples into the relevant classes. Using different classification techniques simultaneously, as if they were a "committee of experts" seems to be a hot topic in the field of Machine Learning (though it's an older one too!). It is clear that an ensemble of learners could improve the class decisions, but this is true if and only if the confusion matrices (i.e. errors) they generate are different. The same reasoning can be

[5] This has been usually the McGill collection of samples.

made for the case of descriptors: if we combine the decisions from two systems which use different subsets of descriptors (even doing the same induction process) we can expect better results than using an only set that subsumes all of them (we need again that the distribution of errors are different for each feature subset).

Though one of our forthcoming steps will be that of using the amassed knowledge for approaching labelling of drum loops, a final stage of our work could deal with some type of percussion description in the context of commercial songs databases (i.e. telling if a song has drums, if they are synthetic or acoustic, if there is Latin or rock percussion, etc, if there are TR-808 sounds or not, etc.). Heittola and Klapuri [51] have approached the problem using sinusoidal plus noise decomposition as a preprocessing stage of analysis. The file obtained after eliminating most of the "harmonic" part is supposed to contain most of the drum sounds (plus other noises such as respirations, pluck noises, etc.). After this pre-processing, two different algorithms have been tested, one dealing with periodicities and the other dealing with timbral descriptors (Mel frequency Cepstral coefficients) and Gaussian mixture modelling. Using a database containing more that 28 hours of audio they have obtained more than 80% of correct decisions regarding if the song has drums or not. Their best results corresponded to the combination of their two classification algorithms. It is clear for us that the descriptors and techniques we have been using, combined with additional signal processing, be they sinusoidal+noise decomposition, ISA or others, deserve a thorough exploration in the light of what we have learnt regarding isolated sounds of unpitched percussion sounds.

5. REFERENCES

[1] T.Zhang and C.-C.Jay Kuo, "Classification and retrieval of sound effects in audiovisual data management," *33rd Asilomar Conference on Signals, Systems, and Computers*, 1999.

[2] M.A.Casey, "MPEG-7 sound recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, No. 6, 2001, pp. 737-747.

[3] G.Tzanetakis, G.Essl, and P.Cook, "Automatic Musical Genre Classification of Audio Signals," *International Symposium on Music Information Retrieval (ISMIR)*, 2001.

[4] B.Whitman, G.Flake, and S.Lawrence, "Artist detection in music with Minnowmatch," *2001 IEEE Workshop on Neural Networks for Signal Processing*, 2001, pp. 559-568.

[5] F.Gouyon and P.Herrera, "Exploration of techniques for automatic labeling of audio drum tracks' instruments," *MOSART: Workshop on Current Directions in Computer Music*, 2001.

[6] P.Herrera, A.Yeterian, and F.Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," *Music and Artificial Intelligence: Second International Conf. Proceedings,* C.Anagnostopoulou, M.Ferrand, and A.Smaill, eds., Berlin, Springer, 2002, pp. 79-91.

[7] P.Herrera, F.Gouyon, and A.Dehamel, *Automatic labeling of drum loops: an integration of bottom-up and top down approaches*, Tech. Report In preparation, 2003.

[8] I.Kaminskyj, "Multi-feature Musical Instrument Sound Classifier," *Australasian Computer Music Association Conference*, 2001, pp. 46-54.

[9] W.A.Schloss, *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High-Level Analysis,* Ph.D. Thesis, Stanford University, 1985.

[10] J.Bilmes, *Timing is the essence: Perceptual and computational techniques for representing, learning and reproducing expressive timing in percussive rhythm,* MSc Thesis, Massachussetts Institute of Technology, Media Laboratory, 1993.

[11] S.McDonald and C.P.Tsang, "Percussive sound identification using spectral centre

trajectories," *1997 Postgraduate Research Conference*, 1997.

[12] J.Sillanpää, *Drum stroke recognition*, Tampere, Finland, 2000. Technical report. Tampere University of Technology

[13] J.Sillanpää, A.Klapuri, J.Seppänen, and T.Virtanen, "Recognition of acoustic noise mixtures by combined bottom-up and top-down approach," *European Signal Processing Conference, EUSIPCO 2000*, 2000.

[14] M.Goto and Y.Muraoka, "A sound source separation system for percussion instruments," *Transactions of the Institute of Electronics, Information and Communication Engineers D-II*, Vol. J77, No. 5, 1994, pp. 901-911.

[15] M.Goto and Y.Muraoka, "A real-time beat tracking system for audio signals," *International Computer Music Conference*, 1995, pp. 171-174.

[16] A.Zils, F.Pachet, O.Delerue, and F.Gouyon, "Automatic Extraction of Drum Tracks from Polyphonic Music Signals," *2nd International Conference on Web Delivering of Music(WedelMusic2002)*, 2002.

[17] F.Gouyon, F.Pachet, and O.Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," *COST G6 Conference on Digital Audio Effects 2000*, 2000.

[18] J.Laroche and J.-L.Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Transactions OnSpeech and Audio Processing*, Vol. 2, No. 2, 1994, pp. 329-344.

[19] B.H.Repp, "The sound of two hands clapping: An exploratory study," *Journal of the Acoustical Society of America*, Vol. 81, 1993, pp. 1100-1109.

[20] A.Freed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive events," *Journal of the Acoustical Society of America*, Vol. 87, No. 1, 1990, pp. 311-322.

[21] R.L.Klatzky, D.K.Pai, and E.P.Krotkov, "Perception of material from contact sounds," *Presence: Teleoperators and Virtual Environments*, Vol. 9, No. 4, 2000, pp. 399-410.

[22] I.Orife, *Riddim: A rythm analysis and decomposition tool based on independent subspace analysis*, Master of Arts Master's dissertation, Dartmouth College, Hanover, NH, 2001.

[23] M.A.Casey and A.Westner, "Separation of mixed audio sources by independent subspace analysis," *Proceedings of the International Computer Music Conference*, 2000.

[24] D.FitzGerald, E.Coyle, and B.Lawlor, "Sub-Band Independent Subspace Analysis for Drum Transcription," *5th International Conference on Digital Audio Effects (DAFX-02)*, 2002, pp. 65-69.

[25] E.Riskedal, *Drum Analysis*, M.Sc. Thesis, Dept. of Informatics, Univ. Bergen, Norway, 2002.

[26] A.Klapuri, "Onset detection by applying psychoacoustic knowledge," *International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[27] A.J.Bell and T.J.Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, Vol. 7, No. 6, 1995, pp. 1129-1159.

[28] M.Jørgensen, "Drumfinder: DSP project on recognition of drum sounds in drum tracks," (http://www.daimi.au.dk/~elmer/dsp)

[29] M.Kragtwijk, *Percussive Music and the Automatic Generation of 3D Animations*, MSc Master of Science thesis, University of Twente, The Netherlands, 2001.

[30] B.Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *International Symposium on Music Information Retrieval, ISMIR 2000*, 2000.

[31] A.Blum and P.Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, Vol. 97, 1997, pp. 245-271.

[32] K.D.Martin and Y.E.Kim, "Musical instrument identification: A pattern-recognition approach," *Proceedings of the 136th meeting of the Acoustical Society of America*, 1998.

[33] I.Fujinaga and K.MacMillan, "Realtime recognition of orchestral instruments," *Proceedings of the 2000 International Computer Music Conference*, 2000, pp. 141-143.

[34] A.Eronen, "Comparison of features for musical instrument recognition," *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)*, 2001.

[35] G.Agostini, M.Longari, and E.Pollastri, "Musical instrument timbres classification with spectral features," *IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 97-102.

[36] J.R.Quinlan, *C4.5: Programs for machine learning*, San Mateo, CA, Morgan Kaufmann, 1993.

[37] K.Jensen and J.Arnspang, "Binary decission tree classification of musical sounds," *Proceedings of the 1999 International Computer Music Conference*, 1999.

[38] A.Wieczorkowska, "Classification of musical instrument sounds using decision trees," *Proceedings of the 8th International Symposium on Sound Engineering and Mastering, ISSEM'99*, 1999, pp. 225-230.

[39] E.Frank and I.H.Witten, "Generating accurate rule sets without global optimization," *Fifteenth International Conference on Machine Learning*, 1998, pp. 144-151.

[40] E.Bauer and R.Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, Vol. 36, No. 1-2, 1999, pp. 105-139.

[41] G.Guo, H.J.Zhang, and F.Li, "Boosting for content-based audio classification and retrieval: An evaluation," *IEEE International Conference on Multimedia and Expo*, 2001.

[42] Y.Freund and R.E.Schapire, "Experiments with a new boosting algorithm," *International Conference on Machine Learning*, 1996, pp. 148-156.

[43] A.Eronen, *Automatic musical instrument recognition*, M.Sc. Thesis, Department of Information Technology, Tampere University of Technology, 2001.

[44] K.D.Martin, *Sound-source recognition: A theory and computational model*, Ph.D. doctoral dissertation, MIT, Cambridge, MA, 1999.

[45] J.R.Miller and E.C.Carterette, "Perceptual space for musical structures," *Journal of the Acoustical Society of America*, No. 58, 1975, pp. 711-720.

[46] J.M.Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, Vol. 61, No. 5, 1977, pp. 1270-1277.

[47] S.McAdams, S.Winsberg, G.de Soete, and J.Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject

classes," *Psychological Research*, Vol. 58, 1995, pp. 177-192.

[48] P.Toiviainen, M.Kaipainen, and J.Louhivuori, "Musical timbre: Similarity ratings correlate with computational feature sapce distances," *JNMR*, No. 24, 1995, pp. 282-298.

[49] S.Lakatos, "A common perceptual space for harmonic and percussive timbres," *Perception and Psychophysics*, Vol. 62, No. 7, 2000, pp. 1426-1439.

[50] G.Peeters, S.McAdams, and P.Herrera, "Instrument sound description in the context of MPEG-7," *Proceedings of the 2000 International Computer Music Conference*, 2000.

[51] T.Heittola and A.Klapuri, *Locating Segments with Drums in Music Signals*, Technical Report, Tampere University of Technology, 2002.

6. ACKNOWLEDGEMENTS

The work reported here has been partially funded by the European Union IST project CUIDADO (http://www.cuidado.mu).