Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

# Automatic recognition of urban sound sources

Boris Defréville[1, 3], Pierre Roy[2], Christophe Rosin[1] and François Pachet[2]

[1] LASA, 236 bis rue de Tolbiac, 75013 Paris, France,
defreville@lasa.fr, christophe.rosin@adp.fr

[2] SONY CSL PARIS, 6 rue Amyot, 75005 Paris, France
pachet@csl.sony.fr, roy@csl.sony.fr

[3] UNIVERSITY OF CERGY-PONTOISE, Rue d'Eragny, Neuville-sur-Oise, 95031 Cergy-Pontoise, France

## ABSTRACT

The goal of the FDAI project is to create a general system that computes an efficient representation of the acoustic environment. More precisely, FDAI has to compute a noise disturbance indicator based on the identification of six categories of sound sources. This paper describes experiments carried out to identify acoustic features and recognition models that were implemented in FDAI. This framework is based on EDS – Extractor Discovery System – an innovative acoustic feature extraction system for sound feature extraction. The design and development of FDAI raised two critical issues. Completeness: it is very difficult to design descriptors that identify *every* sound source in urban environments, and Consistency: some sound sources are not acoustically consistent. We solved the first issue with a conditional evaluation of a family of acoustic descriptors, rather than the evaluation of a single general-purpose extractor. Indeed, a first hierarchical separation between vehicles (moped, bus, motorcycle and car) and non-vehicles (bird and voice) significantly raised the accuracy of identification of the buses. The second issue turned out to be more complex and is still under study. We give here preliminary results.

## 1. INTRODUCTION

### 1.1. Framework

Automatic recognition of environmental sounds has become an important field of research with the increasing demand coming from applications such as environment monitoring [1-4], media annotation [5] or robotics [6].

In this paper we deal with urban soundscape analysis. The impact of environmental noise can have harmful effects on human health and well-being. The European Directive 2002/49/EC of 25th June 2002 relating to the assessment and management of environmental noise lays down a general framework to produce mappings for city noise management [7]. Available to the public and intended to urban planners, these mappings are strategic tools to prepare and implement action plans i.e. prevent and reduce environmental noise where necessary and preserve environmental noise quality where it is satisfactory. In complement to an indicator based on global energetic properties of the sound, this directive encourages member states to define innovative indicators that have to be introduced in mappings.

We have shown recently that models of environmental noise are substantially more accurate when the nature of sound sources is taken into account explicitly [8]. We have therefore introduced a environmental noise model as a linear combination of basic features of source categories (e.g. time of presence of car, moped, motorcycle, bus, voices and birds) in which coefficients are specific to the type of soundscape considered (e.g. park, market, district road).

FDAI is a general framework to provide an efficient representation of the acoustic environment, conforming to the European directive's recommendations. FDAI provides a visualization of the indicator described above on a city's GIS (Geographic Information System).

Operationally, monitoring systems are located in various urban locations. A monitoring system is composed of a sound acquisition set (microphone and audio acquisition card), a notebook computer and real time analyzer that performs measurements, in particular sound sources identification. Then, raw data are sent to a server via an internet connection. The global indicator is then computed from these measurements and automatically sent to a GIS for visualization.

Technically, the most difficult task of FDAI is to classify automatically the sounds sources occurring in the soundscape. The six categories of sound sources considered are: Car, Moped, Bus, Motorbike, Voice and Bird. This paper describes experiments carried out to identify acoustic features and recognition models of FDAI.

### 1.2. Feature Extraction

A typical audio classification system basically follows a two-phase scheme: (1) a *feature extraction* phase to extract information from the samples to classify; and (2) a *classification* phase, in which classifiers are trained on the results of the features of the first phase on a training dataset. The quality of the resulting classifiers depends on three parameters: (1) the quality of the features, i.e. how well they separate samples of different classes while grouping samples of the same class; (2) the performance of the machine-learning algorithm; and 3) the quality (size and variety) of the training dataset. We use here a new approach to improve the typical classification scheme. The idea is to create automatically *problem-specific* features, i.e. features that perform well on the problem at hand, thus improving the quality the resulting classifiers. From a technical standpoint, our approach uses a genetic algorithm which generates features by combining elementary domain-specific operators (also called low-level descriptors). At each generation, the best features are kept to be the seeds for the next generation; the seeds undergo mutations or are combined together to create new, hopefully better, features.

EDS – Extractor Discovery System – is an implementation of this approach in the domain of acoustic signal classification [9]. Many acoustic features have been devised by audio signal processing researchers, and each of them is known to be adapted to specific classification tasks. For instance, some features are well-suited to musical instrument classification, others perform well on speaker recognition, or musical genre classification. A signal processing expert will typically use features known to perform well on the problem at hand. EDS aims at replacing the signal processing expert. EDS, as opposed to the domain expert, has powerful computational skill but poor domain knowledge. More precisely, the domain knowledge in EDS consists of the following elements (note that $X$ represents the signal): 1) a library of operators (low-level descriptors for audio signal processing) containing approximately 80 operators, e.g. Fft, Db, Arcsin, Normalize, Min, Max, LowPassFilter; 2) heuristics that prevent creating useless features, e.g. Min (Max ($X$)), or LowPassFilter (LowPassFilter ($X$, $freq_1$), $freq_2$); and 3) typing rules that ensure the creation of features that are well-formed, e.g. forbid Normalize (Max ($X$)). The key idea of EDS is to invent features as combinations of operators that are specific to the problem to solve, i.e. features that an expert would never have come up with, as they are not in the literature. This approach has been shown to outperform the standard approach on several classification problems, e.g. musical instrument recognition

## 2. CONSTRUCTION OF DATABASES

### 2.1. Metrology

We carried out recordings in typical urban conditions, such as U-shaped streets, speeds around 50 km/h, irregular streams of vehicles and weak slopes of the road. We recorded on streets of various types of asphalt-coating.

We are interested in recognizing six main categories of sound sources: cars, mopeds, bus, motorbikes, human voices, and birds twittering. Note that we work on real-life sounds, not isolated sounds as found e.g. on CD effect libraries. In particular, the sound sources we try to identify are usually mixed up together and with background noise. Figure 1 shows a typical sound sequence used in this study.

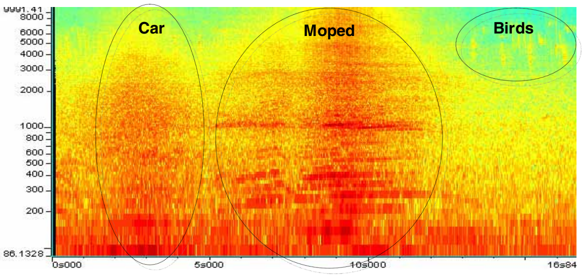

Figure 1. Sonogram of a sequence containing a car, a moped and birds.

Each measurement is based on a protocol and material used by official organizations of measurements, engineering and design acoustic departments. The sound acquisition set (01dB Symphony® system) consists of a transducer linked to a small computer unit (a single channel omni directional Bruel & Kjaer class1 microphone) which transfers data in real time to a notebook computer.

The protocol of measurement is done according to the recommendation of standards [10] i.e. the sites of measurement near buildings must be located at 2 meters in front of the most advanced part of the building, the height of measurement is between 1.2 and 1.5 meters above the ground-level and the microphone is placed at a distance between 2 and 3 meters from the vehicles.

Measurements were carried out in the city of Paris, on one-way streets with isolated vehicles as well as two-way streets. These measurements were made during various moments of the day, on a dry roadway.

We recorded more than 2,000 sequences with the following sound sources, see Table 1:

| | |
|---|---|
| **Cars** | Horns |
| **Mopeds** | Steps |
| **Motorcycles** | Background noise |
| **Buses** | Dogs barking |
| **Birds** | Stroller |
| **Voices** | Police horns |
| Trucks | Slamming door |
| Brake whistling | Rain |

Table 1. Sound sources composing the different sound databases. Names in bold correspond to the six main categories mentioned in Section 1.1

### 2.2. Databases for EDS

In order to avoid features based on intensity, each sample is normalized i.e. its maximum level is brought to the digital maximum level before saturation. Sampling rate is 44,100.

The length of each sample is arbitrary fixed at 500 ms. For each database, 66% of the data was used for training and the other 33% for testing. Typically, we trained classifiers on databases containing between 200 and 800 samples.

## 3. EXPERIMENTAL RESULTS

### 3.1. Parameters of the classifiers

For the experiments reported here, we focused only on two types of classifiers: Gaussian Mixture Models – or GMMs – and Nearest Neighbors – or *k*NN.

The first task is to find the best parameters for the machine-learning algorithms, to optimize the resulting success ratios. Two parameters are considered here: the number of Gaussian Components (M) used in GMMs, and the number of Nearest Neighbor (K) used in *k*NN.

The experience was set on two different sets of features obtained by EDS with respect to two different databases *Moped* and *Bird*. M and K can take the values 20, 30, 40, 50 and 1, 3, 5, 10 respectively. The recognitions rates are counted as follows.

*All*: percentage of correctly recognized samples, *Bird* or *Moped*: percentage of correct recognition of the corresponding class, *Other*: percentage of correct recognition of other sounds. Results are shown in Table 2 and 3.

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Bird | Other | All | Bird | Other |
| K=1 | *100.0* | *100.0* | *100.0* | *69.0* | *46.8* | *92.9* |
| K=3 | *95.0* | *96.7* | *93.3* | *71.8* | *53.2* | *90.5* |
| K=5 | *94.0* | *95.6* | *92.2* | *71.8* | *53.2* | *90.5* |
| K=10 | *93.0* | *95.6* | *91.1* | *70.8* | *53.2* | *90.5* |
| **GMM** | **Train** | | | **Test** | | |
| | All | Bird | Other | All | Bird | Other |
| M=20 | *93.0* | *95.6* | *90.0* | *70.0* | *51.1* | *90.5* |
| M=30 | *94.0* | *96.7* | *91.1* | *71.0* | *51.1* | *92.9* |
| M=40 | *96.0* | *100.0* | *92.2* | *65.0* | *44.7* | *88.1* |
| M=50 | *97.0* | *100.0* | *93.3* | *69.0* | *53.2* | *85.7* |

Table 2. Recognition results (in percentage) for various parameters of the models for the discrimination task between *Bird* and *Other*. The models use 9 features.

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Moped | Other | All | Moped | Other |
| K=1 | *100.0* | *100.0* | *100.0* | *81.0* | *73.2* | *89.4* |
| K=3 | *95.0* | *96.4* | *93.1* | *86.0* | *82.1* | *89.4* |
| K=5 | *94.0* | *96.4* | *92.6* | *84.0* | *77.7* | *89.4* |
| K=10 | *94.0* | *94.6* | *93.5* | *84.0* | *78.6* | *90.3* |
| **GMM** | **Train** | | | **Test** | | |
| | All | Moped | Other | All | Moped | Other |
| M=20 | *94.0* | *96.9* | *91.8* | *83.0* | *80.4* | *85.8* |
| M=30 | *94.0* | *95.1* | *92.2* | *84.0* | *82.1* | *86.7* |
| M=40 | *97.1* | *96.9.* | *97.4* | *86.0* | *85.7* | *85.8* |
| M=50 | *95.0* | *97.8* | *93.1* | *86.0* | *83.9* | *87.6* |

Table 3. Recognition results between *Moped* and *other sounds*. The models use 4 features.

As a conclusion we observe an optimal accuracy for K = 3 and M = 40. Note that all models tend to be more efficient for the recognition of the *other sounds*.

### 3.2. Timbre Features

The following so-called *Timbre* feature is commonly-used for musical instrument recognition [11] or similarity measures between music titles [12] with better performance than basic MPEG-7 features in the general case [13].

$$Mfcc0(Hanning(SplitOverlap(x,2048,0.5)),10) \quad (1)$$

This feature cuts the signal into 2048 points frames (50 ms), and for each frame, computes the short-time spectrum. Then the ten first MFCC (Mel Frequency Cepstrum Coefficients) are computed to estimate the spectral envelope of each frame. A typical 500 ms sample is represented with 10 feature vectors, i.e. 100 coefficients. This feature is tested to complement the features found with EDS.

### 3.3. Mechanical versus Non Mechanical

We ran several experiments on databases containing all seven sound sources (i.e. car, bus, truck, motorcycle, moped, voice, bird). We noticed that we could not manage to create classifiers that have good performance on every sound source (see 3.4). This is what we refer to as the Completeness issue. Instead, we split the recognition process into two phases: a first classifier distinguishes between *mechanical sounds* and *non mechanical sounds* (i.e. voices and birds). The idea is to find in a first step robust features able to separate vehicles from natural, not mechanical sounds (voices, birds, dogs, background noise, etc). In a second step we test others classes of sounds and see if they are grouped together with the extractor. Although this is an arbitrary distinction, this workaround happens to be effective, as shown below.

#### 3.3.1. Database

The repartition of samples of the database is indicated in Table 4.

| Class | Sequences | Quantity |
|---|---|---|
| **Vehicle** | *car* | *60* |
| | *bus* | *60* |
| | *truck* | *60* |
| | *moped* | *60* |
| | *motorcycle* | *60* |
| **Other** | *bird* | *144* |
| | *voice* | *149* |
| | | **623** |

Table 4. Constitution of the database *Mechanical*.

We did not include classes such as Horn or Brake whistling because they always come with strong foreground vehicle sounds.

### 3.3.2. Features

The best features obtained with EDS are shown in Table 5:

| Features |
|---|
| Sqrt(Rangs(SpectralSkewness (Hanning(SplitOverlap(x, 1323, 0.6))))) |
| Sqrt(Sqrt(Range(SpectralSpread (Hanning(SplitOverlap(x, 2205, 0.7))))) |
| Sqrt(Sqrt(Range(SpectralKurtosis (Hann(SplitOverlap(x, 1323, 0.6)))))) |

Table 5. Features obtained by EDS for the discrimination between *Mechanical* and *Non Mechanical* sounds.

### 3.3.3. Models

First we designed optimal models containing the three pre-cited EDS features (Table 6).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Vehi | Other | All | Vehi | Other |
| EDS | *96.0* | *96.6* | *95.3* | *89.2* | *91.2* | *87.2* |
| GMM | Train | | | Test | | |
| | All | Vehi | Other | All | Vehi | Other |
| EDS | *96.5* | *96.6* | *96.3* | *91.2* | *87.2* | *95.1* |

Table 6. Recognition accuracies (%) between *Mechanical* and Non Mechanical sounds using the best features found by EDS.

In a second step, we introduced the timbre feature in addition to the 3 EDS features. The accuracy of the resulting models are shown in Table 7.

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Vehi | Other | All | Vehi | Other |
| EDS + *Timbre* | *100.0* | *100.0* | *100.0* | *95.7* | *97.2* | *94.1* |
| GMM | Train | | | Test | | |
| | All | Vehi | Other | All | Vehi | Other |
| EDS + *Timbre* | *100.0* | *100.0* | *100.0* | *96.6* | *96.3* | *97.1* |

Table 7. Recognition accuracies between *Mechanical* and *Non Mechanical* sounds using the 3 best features found by EDS plus the *Timbre* feature.

A significant increase in recognition rate was brought by the introduction of the *Timbre* feature: the rate raises to 96.6% gaining 3.5%.

## 3.4. Bus

### 3.4.1. Bus vs. all other sound classes

In this experiment we considered the separation between class Bus and all other classes of events. We created a database containing 578 samples. The best features obtained with EDS have a relatively weak fitness (Table 8) predicting poor recognition.

| Features |
|---|
| Sqrt(MaxPos(Derivation(MelBands (x, 8.0)))) |
| Sqrt(Max(Derivation(Sqrt(BarkBands (x, 8.0)))) |
| Iqr(Derivation(Zcr(FilterBank(x, 10.0)) |
| Sqrt(Mean(Power(Chroma(x), -1.0))) |

Table 8. Features obtained by EDS for discrimination between *Bus* and all other classes of sound.

Indeed, the best recognition rate on *Test* is only 67.0% for the model using k-NN (Table 9).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Bus | Other | All | Bus | Other |
| K=3 | *92.0* | *95.3* | *88.7* | *67.0* | *50.0* | *84.2* |
| GMM | Train | | | Test | | |
| | All | Bus | Other | All | Bus | Other |
| M=40 | *96.1* | *99.5* | *92.8* | *64.3* | *47.8* | *80.8* |

Table 9. Recognition accuracies (%) between *Vehicle* and all other sounds classes using the 4 best EDS features.

The recognition is also weak with the feature *Timbre* (Table 10).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Bus | Other | All | Bus | Other |
| K=3 | *99.5* | *100* | *99.0* | *70.6* | *60.4* | *81.1* |
| GMM | Train | | | Test | | |
| | All | Bus | Other | All | Bus | Other |
| M=40 | *98.9* | *100* | *97.9* | *69.1* | *46.9* | *91.6* |

Table 10. Recognition accuracies (%) between *Vehicle* and all other sounds classes using the *Timbre* feature.

### 3.4.2. Bus vs. other vehicles

This experiment addresses the bus versus other vehicles problem. A database was created containing only buses and other vehicles (777 samples). The genetic search with EDS yields the following features (Table 11).

| Features |
|---|
| Log10(Range(SpectralDecrease (SplitOverlap (x, 1323, 0.8)))) |
| Sqrt(MaxPos(Derivation(BarkBands (x, 5.0)))) |
| Power(MaxPos(MelBands(x, 7)), -4.8) |
| Median(Power(Chroma(x), -1.0)) |

Table 11. Features obtained by EDS for discrimination between Bus and other vehicles.

Finally, we have extracted models containing these four features (EDS) plus the feature *Timbre* (Table 12).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Bus | Other | All | Bus | Other |
| EDS | *95.9* | *95.1* | *96.9* | *85.2* | *72.5* | *98.4* |
| EDS + *Timbre* | 100.0 | 100.0 | 100.0 | *85.6* | *75.6* | *96.1* |
| **GMM** | **Train** | | | **Test** | | |
| | All | Bus | Other | All | Bus | Other |
| EDS | *95.3* | *92.8* | *98.0* | *86.8* | *75.6* | *98.4* |
| EDS + *Timbre* | *99.6* | *99.6* | *99.6* | *79.4* | *69.5* | *89.9* |

Table 12 Recognition accuracies (%) between *Bus* and *other vehicles* using 4 best features found by EDS plus feature *Timbre*.

Here, the best recognition rate reaches 86.8% on the *Test* database. It represents a much better result than the one obtained while trying to separate buses from all other classes of events. The *Timbre* feature does not have a significant influence on the results. However, the accuracy of the models is not completely satisfying. It is probably due to the large diversity of sounds within the bus category. It could be fixed by listening the samples in order to separate the sample into sub-groups.

The next experiments address the distinction between vehicles sound and no vehicle sounds.

### 3.5. Moped

A first experiment ran on mopeds gave unsatisfactory results. By listening to the incorrectly classified samples, we noticed that they sounded differently from the other samples. Indeed, these samples were representing a new class that was finally identified as a particular type of mopeds called *Vespa*.

We decided to create a new database but this time, without considering *Vespa* as a part of the class *mopeds*.

The five best features found by EDS are shown in Table 13.

| Features |
|---|
| Sqrt(Mean(SpectralFlatness (SplitOverlap (x, 2205.0, 0.8)))) |
| Log10(Median(MelBands(Normalize (x), 2.0))) |
| Min(SpectralCentroid(Split(x, 256.0))) |
| Power(Mean(Mfcc(x, 9.0) 1.3) |
| Log10(SpectralSkewness(x)) |

Table 13. Features obtained by EDS for discrimination between Moped and other vehicles.

The best model found uses only one feature containing Mfcc as a core operator (Table 14).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Moped | Other | All | Moped | Other |
| K=3 | *83.3* | *85.2* | *81.4* | *88.9* | *96.4* | *81.4* |

Table 14. Recognition accuracy (%) between *Mopeds* and other classes of vehicles. The model uses the only the feature *Power(Mean(Mfcc(x, 9.0) 1.3).*

As a good feature for discrimination between bus and other vehicles, it is noticeable that *Timbre* does not work satisfactorily for the identification of the mopeds (we tried with 10 and 20 Mfcc's).

This experiment shows that the perceptive class *Moped* is acoustically inconsistent. We discuss this issue below (see 4.2.).

### 3.6. Car

The database for Car contains 276 samples of cars and other vehicles. The features obtained by EDS are shown in Table 15.

The best model using the 6 best EDS features yields a recognition rate of 90.0% (Table 16). k-NN is more appropriate than GMMs with a difference on *TestD* equal to 23.3%. The introduction of the *Timbre* features yields a recognition rate of 85.6%. However, the combination between the 6 EDS features and Timbre is not profitable.

| Features |
|---|
| Power(Abs(Sum(SpectralFlatness(Triangle (BpFilter(SplitOverLap(x, 441.0, 0.8), 882.0, 1323.0))))), -0.25) |
| Power(Mean(SpectralRolloff(BpFilter(Split (x, 1764))))) |
| Power(Abs(Median(RHF(Abs(BpFilter (FilterBank(x, 11.0), 882.0, 1323.0)))))), -0.25) |
| Sqrt(Range(SpectralDecrease(Triangle(BpFilter (FilterBank(x, 5.0), 853.0, 970.0))))) |
| Sqrt(Zcr(Percentile(Hann(SplitOverLap (x, 441.0, 0.5)), 58.0))) |
| Sum(Hamming(Derivation(Mfcc(x, 8.0)))) |

Table 15. Features found by EDS to discriminate between *Car* and other classes of vehicles.

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Car | Other | All | Car | Other |
| EDS | 93.0 | 92.5 | 93.5 | 90.0 | 84.4 | 95.6 |
| *Timbre* 10Mfcc | 95.6 | 77.8 | 93.3 | 85.6 | 77.8 | 93.3 |
| EDS + *Timbre* | 100.0 | 100.0 | 100.0 | 85.6 | 77.8 | 93.3 |
| **GMM** | **Train** | | | **Test** | | |
| | All | Car | Other | All | Car | Other |
| EDS | 99.5 | 100.0 | 98.9 | 66.7 | 53.3 | 80.0 |
| *Timbre* 10Mfcc | 98.9 | 100.0 | 97.8 | 84.4 | 73.3 | 95.6 |
| EDS + *Timbre* | 100.0 | 100.0 | 100.0 | 78.9 | 62.2 | 95.6 |

Table 16. Recognition accuracies (%) between *Car* and other classes of vehicles.

Finally, the best model is obtained with the 20 MFCC in a *k*-NN classifier. Its accuracy is slightly improved by 1.1% with the addition of an EDS feature using RHF (Ratio of High Frequencies) exposed in Table 15. The final rate is 94.4% of correct recognitions on *TestD* (Table 17).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Car | Other | All | Car | Other |
| *Timbre* 20Mfcc | 100.0 | 100.0 | 100.0 | 93.3 | 93.3 | 93.3 |
| *Timbre* + EDS | 100.0 | 100.0 | 100.0 | 94.4 | 95.6 | 93.3 |
| **GMM** | **Train** | | | **Test** | | |
| | All | Bus | Other | All | Bus | Other |
| *Timbre* 20Mfcc | 99.5 | 100.0 | 98.9 | 83.9 | 80.0 | 87.8 |
| *Timbre* + EDS | 99.5 | 100.0 | 98.9 | 85.6 | 80.0 | 91.1 |

Table 17. Recognition accuracies (%) between *Car* and other classes of vehicles.

### 3.7. Motorbikes

The database created contains 466 samples of motorbikes and other vehicles. The models created with the best EDS features or *Timbre* features never yield a recognition rate over 82.5 % on *Test*. As in the Moped's case, the samples are not acoustically consistent i.e. there are too many different types of sound in the *Motorbike* category.

### 3.8. Bird

The database created contains 274 samples of birds and other non mechanical sounds. The 4 best features obtained by EDS are shown in Table 18.

| Features |
|---|
| Power(Median(Log10(SpectralRolloff (FilterBank(x,10.0)))), -0.125) |
| Square(Iqr(PitchBands(Sqrt (Abs(x)), 9.0))) |
| Power(RMS(SpectralSpread(Arcsin (Hann(SplitOverlap(x, 220.0, 0.2))))), 3.0) |
| Power(Arcsin(Square(Iqr(Zcr(Integration (Hanning(FilterBank(x, 8.0)))))), 2.6) |

Table 18. Features obtained by EDS to discriminate between *Bird* and other non mechanical sounds.

The better model found yields only 71.8% of recognition rate on *Test*. One explanation of this weak recognition rate could come from the fact that bird twitter is most of the time shorter than the duration of the sample (500 ms) used here. This technical problem could be fixed in further studies by using larger samples (about 100 ms).

### 3.9. Voice

The database created contains 298 samples of voices and other non mechanical sounds. The voices are either male, female or child speaking. After several iterations, EDS did not produce satisfactory accuracy. Thus we decided to use only the *Timbre* features while varying the number of Mfcc. The best result is obtained with 20 coefficients. The global recognition rate on *Test* goes up to 99.0% (Table 19).

| KNN | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | All | Voice | Other | All | Voice | Other |
| 10Mfcc | *100.0* | *100.0* | *100.0* | *89.2* | *100.0* | *79.2* |
| 20Mfcc | *100.0* | *100.0* | *100.0* | *99.0* | *100.0* | *98.1* |
| **GMM** | **Train** | | | **Test** | | |
| | All | Voice | Other | All | Voice | Other |
| 10Mfcc | *100.0* | *100.0* | *100.0* | *87.0* | *100.0* | *75.5* |
| 20Mfcc | *100.0* | *100.0* | *100.0* | *92.2* | *100.0* | *84.9* |

Table 19. Recognition accuracies (%) between *Birds* and other non mechanical sounds using *Timbre* features with 10 or 20 coefficients.

## 4. COMMENTS

During the design and development of FDAI, we faced too critical issues. Completeness: it is very difficult, if not impossible, to design descriptors that identify every sound source in such a complex acoustic environment. Consistency: some sound sources are not acoustically consistent.

### 4.1. Completeness

When dealing with multiple-class classification problems, we face the completeness issue, i.e. it is difficult to design classifiers that perform well on all the classes. In our implementation, we solved this issue by introducing a hierarchy of classifiers instead of a single multi-class classifier. Basically, we considered the distinction a human would spontaneously make between motor-vehicles and "natural" sounds (voices, birds). The classification process we enforce consists in first, using the "motor-vehicles versus natural sounds" classifier, and then, apply either the "car-truck-moped-bus-motorcycle" or the "voice-birds" classifier according to the results yielded by the first classifier.

Obviously, this is an *ad hoc* workaround based on human mental categories. It is not satisfying since there is no guaranty that this hierarchy is optimal, and besides, the idea to distinguish first between motor-vehicles and natural sounds is based on human intuitions, and is therefore difficult to automate.

We are working on a generalization of this idea. First, we want to automate and optimize the composition of the hierarchy of classifiers. Secondly, we will explore more complex organizations than hierarchies, like for instance Bayesian networks of classifiers.

### 4.2. Consistency

When asked to describe their perception of an urban soundscape, people spontaneously mention sound sources corresponding to mental categories that are culturally consistent, e.g. *motorcycle*, *car*, *truck*, or *birds*. However, these categories are not acoustically consistent. For instance, in the motorcycle category, the sound of a Japanese 4-cylinder engine in a sports motorcycle is acoustically very different from the sound of a 2-cylinder Harley-Davidson engine. This phenomenon also occurs with musical instruments: the category called *guitar* corresponds in reality to various instruments that are acoustically very different from one another, e.g. *saturated electric guitar* and *folk guitar*.

The performance of the best classifiers we created on the motorcycle category is substantially worse than the performance obtained on other categories (motorcycle: 82.0% versus bus: 86.8%, moped: 88.9%, car: 90.0%, voice: 99.0%). As explained in this article, the performance depends on the quality of the features, on the performance of the training algorithm, and on the quality of the training dataset.

So far, we do not know whether the poor performance we obtained stems from some shortcoming of our implementation. It could be the case that EDS cannot produce high-quality features for inconsistent categories, or that the machine learning algorithms we used are not well-adapted to this case, although the *GMM* model is designed to handle this situation. Finally, we may improve the performance by using larger training datasets.

However, we do suspect that standard classification systems, even augmented with EDS's automatic feature creation, cannot reach very high success ratios on acoustically inconsistent categories. We are in the process of devising techniques to improve further the classification process by taking into account inconsistent categories explicitly.

## 5. CONCLUSIONS AND PROSPECTS

In this paper, we have presented FDAI, a framework that uses an innovative features extraction system (EDS) to identify sound sources in complex urban soundscape. This system is already used in a real-time environment for noise quality assessment. Our experiments show that the best results are obtained using both EDS and *Timbre* features. As far as we know, no other study gives recognition accuracies for a similar problem. [3] describes an approach with 95.3% of correct recognition of mopeds (vs. 88.9% in this study). But the recognition task was simpler and consisted in recognizing only two classes of sound sources: moped and horns.

The results we have shown are satisfactory with respect to four sound categories: bus (86.8%), moped (88.9%), car (90.0%) and voice (99.0%) and for separation between mechanical and non mechanical sounds (96.6%). As discussed, we need to improve the performance on motorbikes (82.0%) and birds (71.8%). One way to improve the quality of our system (both in terms of performance and extensibility) is to address the two critical issues discussed above; completeness and consistency.

## 6. ACKNOWLEDGEMENTS

This work was supported by the French Ministry of Ecology and Sustainable Development through the research program “Bruit et Nuisances Sonores”.

## 7. REFERENCES

[1] Cowling M., Sitte R. “Comparison techniques for environmental sound recognition.” Pattern Recognition Letters 24 (2003) 2895-2907.

[2] Harma A., Skowronek J., McKinney M.F., “Acoustic monitoring of activity patterns in office, street and garden environment”. Measuring Behavior 2005. 5th International Conference on Methods and Techniques in Behavioral Research, 30 august – 2 September 2005, Wageningen, The Netherlands.

[3] Couvreur L. Laniray M. “Automatic Noise Recognition in urban environments”. Internoise 2004. Prague, Czech Republic, August 22-25.

[4] Dufournet D., Jouenne P., and Rozwadowski A. “Automatic noise source recognition”. The Journal of the Acoustical Society of America, May 1998, Volume 103, Issue 5, p. 2950.

[5] Toyoda Y., Huang J., Ding S., Liu Y. “Environmental sound recognition by multilayered neural networks”. The Fourth International Conference on Computer and Information Technology. Whuan, China, September 14-16, 2004.

[6] Casey, M. “MPGEG-7 Sound Recognition Tolls.” IEEE Trans.Circ.Systems for Video Tech. 11 (6), 737-747.

[7] Directive 2002/49/CE of the European parliament and of the council of 25 june 2002 relating to the assessment and management on environment noise. Brussels, The European Parliament and the Council of the European Union (2002).

[8] Defréville B., Lavandier C. “Unpleasantness of urban sound environment based on identification of sources: a perceptive and an acoustic approach”. Forum Acusticum, Budapest, 2005.

[9] Zils, A. and Pachet, F. “Automatic Extraction of Music Descriptors from Acoustic Signals using EDS”. Proceedings of the 116th AES Convention, May 2004.

[10] ISO 1996-2 ”Acoustics – Description measurement and assessment of environmental noise – Part 2 : Acquisition of data to land use ”. International Organisation for Standardisation, (1996).

[11] Schwarz, D. & Rodet, X. “ Spectral Estimation and Representation for Sound Analysis-Synthesis”. In proc. ICMC 1999

[12] Aucouturier, J.-J. and Pachet F. “Improving Timbre Similarity: How high is the sky?.” Journal of Negative Results in Speech and Audio Sciences, 1(1), 2004.

[13] Kim H., Burred JJ., Sikora T. “How efficient is MPEG-7 for general sounds recognition?” AES, 25th International Conference, London, UK, 204 June 17-19.