Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 960863, 11 pages
doi:10.1155/2010/960863

Research Article

Ecological Acoustics Perspective for Content-Based Retrieval of Environmental Sounds

Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera, and Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain

Correspondence should be addressed to Gerard Roma, gerard.roma@upf.edu

Received 1 March 2010; Accepted 22 November 2010

Academic Editor: Andrea Valle

In this paper we present a method to search for environmental sounds in large unstructured databases of user-submitted audio, using a general sound events taxonomy from ecological acoustics. We discuss the use of Support Vector Machines to classify sound recordings according to the taxonomy and describe two use cases for the obtained classification models: a content-based web search interface for a large audio database and a method for segmenting field recordings to assist sound design.

1. Introduction

Sound designers have traditionally made extensive use of recordings for creating the auditory content of audiovisual productions. Many of these sound effects come from commercial sound libraries, either in the form of CD/DVD collections or more recently as online databases. These repositories are organized according to editorial criteria and contain a wide range of sounds recorded in controlled environments. With the rapid growth of social media, large amounts of sound material are becoming available through the web every day. In contrast with traditional audiovisual media, networked multimedia environments can exploit such a rich source of data to provide content that evolves over time. As an example, virtual environments based on simulation of physical spaces have become common for socializing and game play. Many of these environments have followed the trend towards user-centered technologies and user-generated content that has emerged on the web. Some programs allow users to create and upload their own 3D models of objects and spaces and sites such as *Google 3D Warehouse* can be used to find suitable models for these environments.

In general, the auditory aspect of these worlds is significantly less developed than the visual counterpart. Virtual worlds like *Second Life* (http://secondlife.com/) allow users to upload custom sounds for object interactions, but there is no infrastructure that aids the user in searching and selecting sounds. In this context, open, user-contributed sound repositories such as *Freesound* [1] could be used as a rich source of material for improving the acoustic experience of virtual environments [2]. Since its inception in 2005, *Freesound* has become a renowned repository of sounds with a noncommercial license. Sounds are contributed by a very active online community, that has been a crucial factor for the rapid increase in the number of sounds available. Currently, the database stores about 84000 sounds, labeled with approximately 18000 unique tags. However, searching for sounds in user-contributed databases is still problematic. Sounds are often insufficiently annotated and the tags come from very diverse vocabularies [3]. Some sounds are isolated and segmented, but very often long recordings containing mixtures of environmental sounds are uploaded. In this situation, content-based retrieval methods could be a valuable tool for sound search and selection.

With respect to indexing and retrieval of sounds for virtual spaces, we are interested in categorizations that take into account the perception of environmental sounds. In this context, the ideas of Gaver have become commonplace. In [4], he emphasized the distinction between musical listening—as defined by Schaeffer [5]—and everyday listening. He also devised a comprehensive taxonomy of everyday sounds based

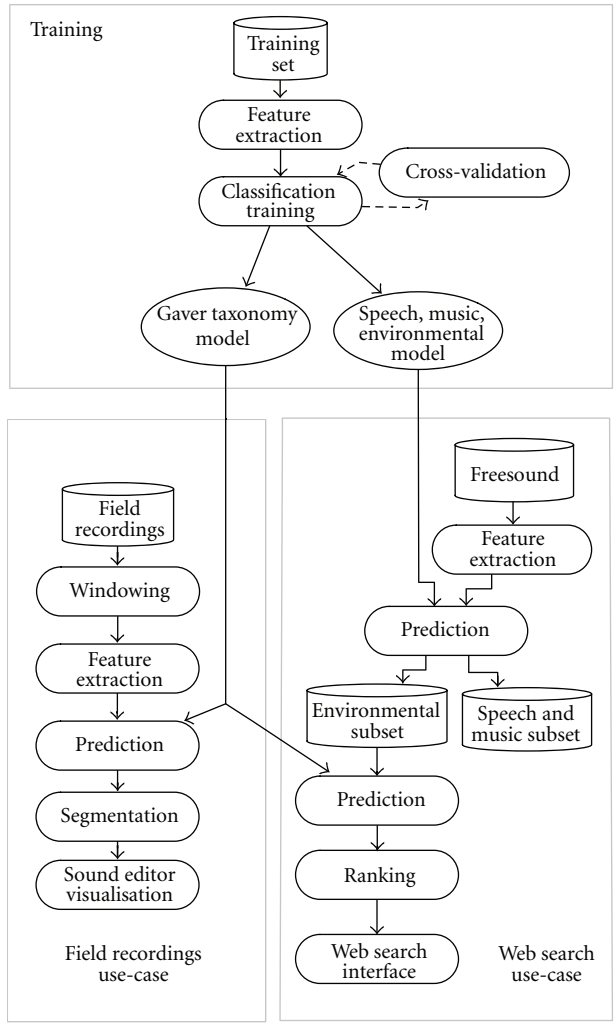

Figure 1: Block diagram of the general system. the models generated in the training stage are employed in the two proposed use-cases.

on the principles of ecological acoustics while pointing out the problems with traditional organization of sound effects libraries. The CLOSED project (http://closed.ircam.fr/), for example, uses this taxonomy in order to develop physically based sound models [6]. Nevertheless, most of the previous work on automatic analysis of environmental sounds deals with experiment-specific sets of sounds and does not make use of an established taxonomy.

The problem of using content-based methods with unstructured audio databases is that the relevant descriptors to be used depend on the kind of sounds and applications. For example using musical descriptors on field recordings can produce confusing results. Our proposal in this paper is to use an application-specific *perspective* to search the database. In this case, this means filtering out music and speech sounds and using the mentioned taxonomy to search specifically for environmental sounds.

1.1. Outline. In this paper, we analyze the use of Gaver's taxonomy for retrieving sounds from user-contributed audio repositories. Figure 1 shows an overview of this supervised learning approach. Given a collection of training examples, the system extracts signal descriptors. The descriptors are used to train models that can classify sounds as speech, music, or environmental sound, and in the last case, as one of the classes defined in the taxonomy. From the trained models, we devise two use cases. The first consists in using the models to search for sound clips using a web interface. In the second, the models are used to facilitate the annotation of field recordings by finding audio segments that are relevant to the taxonomy.

In the following section, we review related work on automatic description of environmental sound. Next, we justify the taxonomical categorization of sounds used in this project. We then describe the approach to classification and segmentation of audio files and report several classification experiments. Finally, we describe the two use cases to illustrate the viability of the proposed approach.

2. Related Work

Analysis and categorization of environmental sounds has traditionally been related to the management of sound effects libraries. The taxonomies used in these libraries typically

do not attempt to provide a comprehensive organization of sounds, but it is common to find semantic concepts that are well identified as categories, such as animal sounds or vehicles. This ability for sounds to represent or evoke certain concepts determines their usefulness in contexts such as video production or multimedia content creation.

Content-based techniques have been applied to limited vocabularies and taxonomies from sound effects libraries. For example, good results have been reported when using Hidden Markov Models (HMM) on rather specific classes of sound effects [7, 8]. There are two problems with this kind of approach. On one hand, dealing with noncomprehensive taxonomies ignores the fact that real world applications will typically have to deal with much larger vocabularies. Many of these works may be difficult to scale to vocabularies and databases orders of magnitude larger. On the other hand, most of the time they work with small databases of sounds recorded and edited under controlled conditions. This means that it is not clear how this methods would generalize to noisier environments and databases. In particular, we deal with user-contributed media, typically involving a wide variety of situations, recording, equipment, motivations, and skills.

Some works have explored the vocabulary scalability issue by using more efficient classifiers. For example in [9], the problem of extending content-based classification to thousands of labels was approached using a nearest neighbor classifier. The system presented in [10] bridges the semantic space and the acoustic space by deriving independent hierarchical representations of both. In [11], scalability of several classification methods is analyzed for large-scale audio retrieval.

With respect to real world conditions, another trend of work has been directed to classification of environmental sound using only statistical features, that is, without attempting to identify or isolate sound events [12]. Applications of these techniques range from analysis and reduction of urban noise, to the detection of acoustic background for mobile phones (e.g., office, restaurant, train, etc.). For instance, the classification experiment in [13] employs a fixed set of 15 background soundscapes (e.g., restaurant, nature-daytime, etc.).

Most of the mentioned works bypass the question of the generality of concepts. Generality is sometimes achieved by increasing the size of the vocabulary in order to include any possible concept. This approach retains some of the problems related to semantic interaction with sound, such as the ambiguity of many concepts, the lack of annotations, and the difficulty to account for fake but convincing sound representations used by foley artists. We propose the use of a taxonomy motivated by ecological acoustics which attempts to provide a general account of environmental sounds [4]. This allows us to approach audio found in user-contributed media and field recordings using content-based methods. In this sense, our aim is to provide a more general way to interact with audio databases both in the sense of the kind of sounds that can be found and in the sense of the diversity of conditions.

3. Taxonomical Organization of Environmental Sound

3.1. General Categorization. A general definition of environmental sound is attributed to Vanderveer: "any potentially audible acoustic event which is caused by motions in the ordinary human environment" [14]. Interest in categorization of environmental sounds has appeared in many disciplines and with different goals. Two important trends have traditionally been the approach inherited from *musique concrète*, which focuses on the properties of sounds independently of their source, and the *representational* approach, concentrating on the physical source of the sound. While the second view is generally used for searching sounds to match visual representations, the tradition of foley artists shows that taking into account the acoustic properties is also useful, especially because of the difficulty in finding sounds that exactly match a particular representation. It is often found that sounds coming from a different source than the described object or situation offer a much more convincing effect. Gaver's ecological acoustics hypothesis states that in everyday listening (different from musical listening) we use the acoustic properties of sounds to identify the sources. Thus, his taxonomy provides a generalization that can be useful for searching sounds from the *representational* point of view.

One important aspect of this taxonomy is that music and animal voices are missing. As suggested in [15], the perception of animal vocalizations seems to be the result of a specialization of the auditory system. The distinction of musical sounds can be justified from a cultural point of view. While musical instrument sounds could be classified as environmental sounds, the perception of musical structures is mediated by different goals than the perception of environmental sounds. A similar case could be made for artificial acoustic signals such as alarms or sirens, in the sense that when we hear those sounds the message associated to them by convention is more important than the mechanism that produces the sound.

Another distinction from the point of view of ecological acoustics can be drawn between "sound events" and "ambient noise". Sound is always the result of an interaction of entities of the environment, and therefore it always conveys information about the physical event. However, this identification is obviously influenced by many factors such as the mixture of sounds from different events, or the location of the source. Guastavino [16] and Maffiolo [17] have supported through psychological experiments the assumptions posed by Schafer [18] that sound perception in humans highlights a distinction between sound events, attributed to clearly identified sources, and ambient noise, in which sounds blur together into a generic and unanalyzable background noise.

Such salient events that are not produced by animal voices or musical instruments can be classified, as suggested by Gaver, using the general acoustic properties related with different kinds of materials and the interactions between them (Figure 2). In his classification of everyday sounds, three fundamental sources are considered: *Vibrating Solids*,

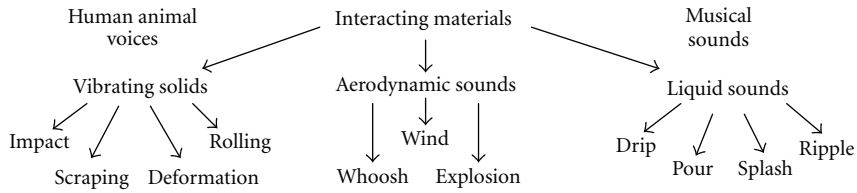

Figure 2: Representation of the Gaver taxonomy.

Aerodynamic sounds(gasses), and *Liquid sounds*. For each of these sources, he proposes several basic auditory events: deformation, impact, scraping, and rolling (for solids); explosion, whoosh and wind (for gas); drip, pour, splash, and ripple (for liquids). We adopt this taxonomy in the present research, and discuss the criteria followed for the manual sound annotation process in Section 6.

3.2. Taxonomy Presence in Online Sound Databases Metadata. Traditionally, sound effects libraries contain recordings that cover a fixed structure of sound categories defined by the publisher. In user-contributed databases, the most common practice is to use free tags that build complex metadata structures usually known as *folksonomies*. In this paper, we address the limitations of searching for environmental sounds in unstructured user-contributed databases, taking *Freesound* as a case study. During several years, users of this site have described uploaded sounds using free tags in a similar way to other social media sites.

We study the presence of the studied ecological acoustics taxonomy terms in *Freesound* (91443 sounds), comparing it to two online-sound-structured databases by different publishers, *SoundIdeas* (http://www.soundideas.com/) (150191 sounds), and *Soundsnap* (http://www.soundsnap.com/) (112593 sounds). Figure 3 shows three histograms depicting the presence of the taxonomy's terms in the different databases. In order to widen the search, we extend each term of the taxonomy with various synonyms extracted from the Wordnet database [19]. For example, for the taxonomy term "scraping", the query is extended with the terms "scrap", "scratch", and "scratching". The histograms are computed by dividing the number of files found for a concept by the total number of files in each database.

Comparing the three histograms, we observe a more similar distribution for the two structured databases (middle and bottom) than for *Freesound*. Also, the taxonomy is notably less represented in the *Freesound*'s folksonomy than in SoundSnap or SoundIdeas databases, with a percentage of retrieved results of 14.39%, 27.48%, and 22.37%, respectively. Thus, a content-based approach should facilitate the retrieval of sounds in unstructured databases using these concepts.

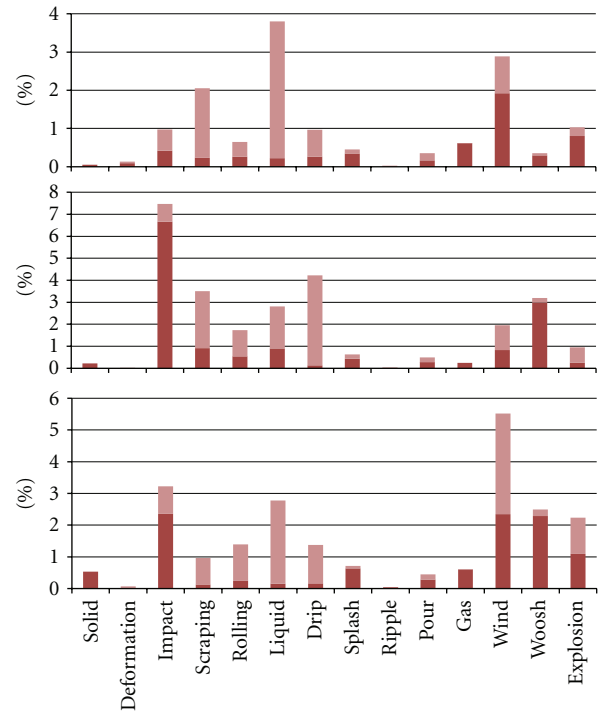

Figure 3: Percentage of sound files in different sound databases, containing taxonomy's terms (dark) and hyponyms from Wordnet (light). *Freesound* (top), Soundsnap (middle), and Soundideas (bottom).

4. Automatic Classification of Environmental Sounds

4.1. Overview. We consider automatic categorization of environmental sounds as a multiclass classification problem. Our assumption is that salient events in environmental sound recordings can be generally classified using the mentioned taxonomy with different levels of confidence. In the end, we aim at finding sounds that provide clear representations of physical events. Such sounds can be found, on the one hand, in already cut audio clips where either a user or a sound designer has found a specific concept to be well represented, or, on the other hand, in longer field recordings without any segmentation. We use sound files from the first type to create automatic classification models, which can later be used to detect events examples both in sound snippets or in longer recordings.

4.2. Sound Collections. We collected sound clips from several sources in order to create ground truth databases for our classification and detection experiments. Our main classification problems are first to tell apart music, voice, and environmental sounds, and then find good representations

of basic auditory events in the broad class of environmental sounds.

4.2.1. Music and Voice Samples. For the classification of music, voice, and environmental sounds, we downloaded large databases of voice and music recordings, and used our sound events database (described below) as the ground truth for environmental sounds. We randomly sampled 1000 instances for each collection. As our ground truth for voice clips, we downloaded several speech corpuses from *voxforge* (http://www.voxforge.org/), containing sentences from different speakers. For our music ground truth, we downloaded musical loops from *indaba* (http://www.indabamusic.com/), where more than 8 GB of musical loops are available. The collection of examples for these datasets was straightforward, as they provide a good sample of the kind of music and voice audio clips that can be found in *Freesound* and generally around the internet.

4.2.2. Environmental Sound Samples. Finding samples that provide a good representation of sound events as defined in the taxonomy was more demanding. We collected samples from three main sources: the *Sound Events* database (http://www.psy.cmu.edu/auditorylab/AuditoryLab.html), a collection of sound effects CDs, and *Freesound*.

The *Sound Events* collection provides examples of many classes of the taxonomy, although it does not match it completely. Sounds from this database are planned and recorded in a controlled setting, and multiple recordings are made for each setup. A second set was collected from a number of sound effect libraries, with different levels of quality. Sounds in this collection generally try to provide good representations of specific categories. For instance, for the *explosion* category we selected sounds from gunshots, for the *ripple* category we typically selected sounds from streams and rivers. Some of these sounds contain background noise or unrelated sounds. Our third collection consists of sounds downloaded from *Freesound* for each of the categories. This set is the most heterogeneous of the three, as sounds are recorded in very different conditions and situations. Many contain background noise and some are not segmented with the purpose of isolating a particular sound event.

In the collection of sounds, we faced some issues, mainly related to the tradeoff between the pureness of events as described in the theory and our practical need to allow the indexing of large databases with a wide variety of sounds. Thus, we included sounds dominated by basic events but that could include some patterned, compound, or hybrid events [4].

(i) *Temporal patterns* of events are complex events formed by repetitions of basic events. These were avoided especially for events with a well-defined energy envelope (e.g., impacts).

(ii) *Compound events* are the superposition of more than one type of basic event, for example, specific door locks, where the sound is generated by a mix of impacts, deformations, and scrapings. This is very common for most types of events in real world situations.

(iii) *Hybrid events* result of the interaction between different materials, such as when water drips onto a solid surface. Hybrid events were generally avoided. Still, we included some rain samples as a *drip* event when it was possible to identify individual raindrops.

The description of the different aspects conveyed by basic events in [4] was also useful to qualitatively determine whether samples belonged to a class or not. For example, in many liquid sounds it can be difficult to decide between *splash* (which conveys *viscosity*, *object size* and *force*) or *ripple* (*viscosity*, *turbulence*). Thus the inability to perceive *object size*, and *force* can determine the choice of the category.

4.3. Audio Features. In order to represent the sounds for the automatic classification process, we extract a number of frame-level features using a window of 23 ms and a hop size of 11.5 ms. One important question in the discrimination of general auditory events is how much of our ability comes from discriminating properties of the spectrum, and how much is focused on following the temporal evolution of the sound. A traditional hypothesis in the field of ecological acoustics was formulated by Vanderveer, stating that interactions are perceived in the temporal domain, while objects determine the frequency domain [14]. However, in order to obtain a compact description of each sound that can be used in the classification, we need to integrate the frame-level features in a vector that describes the whole sound. In several fields involved with classification of audio data, it has been common to use the *bag of frames* approach, meaning that the order of frames in a sound is ignored, and only the statistics of the frame descriptors are taken into account. This approach has been shown to be sufficient for discriminating different sound environments [12]. However, for the case of sound events it is clear that time-varying aspects of the sound are necessary to recognize different classes. This is especially true for impulsive classes such as impacts, explosions, splashes, and to a lower extent by classes that imply some regularity, like rolling. We computed several descriptors of the time series of each frame-level feature. We analyze the performance of these descriptors through the experiment in Section 5.

We used an implementation of Mel Frequency Cepstrum Coefficients (MFCCs) as a baseline for our experiments, as they are widely used as a representation of timbre in speech and general audio. Our implementation uses 40 bands and 13 coefficients. On the other hand, we selected a number of descriptors from a large set of features mostly related with the MPEG-7 standard [20]. We used a feature selection algorithm that wraps the same SVM used for the classification to obtain a reduced set of descriptors that are discriminative for this problem [21]. For the feature selection, we used only mean and variance of each frame-level descriptor. Table 1 shows the features that were selected in this process. Many of them have been found to be related to the identification of environmental sounds in psychoacoustic studies [22, 23]. Also, it is noticeable that

Table 1: Frame-level descriptors chosen by the feature-selection process on our dataset.

High frequency content
Instantaneous confidence of pitch detector (yinFFT)
Spectral contrast coefficients
Silence rate (−20 dB, −30 dB and −60 dB)
Spectral centroid
Spectral complexity
Spectral crest
Spectral spread
Shape-based spectral contrast
Ratio of energy per band (20–150 Hz, 150–800 Hz, 800–4 k Hz, 4 k–20 kHz)
Zero crossing rate
Inharmonicity
Tristimulus of harmonic peaks

Table 2: Sets of descriptors extracted from the temporal evolution of frame-level features and the number of descriptors per frame level feature.

Name	Description	No. desc.
mv	mean and variance	2
mvd	*mv*, derivatives	6
mvdad	*mvd*, log attack time and decay	8
mvdadt	*mvdad*, temp. centroid, kurtosis, skewness, flatness	12

several time-domain descriptors (such as the zero-crossing rate or the rate of frames below different thresholds) were selected.

In order to describe the temporal evolution of the frame level features, we computed several measures of the time series of each feature, such as the log attack time, a measure of decay [24], and several descriptors derived from the statistical moments (Table 2). One drawback of this approach is to deal with the broad variety of possible temporal positions of auditory events inside the clip. In order to partially overcome this limitation, we crop all clips to remove the audio that has a signal energy below −60 dB FSD at the beginning and end of the file.

4.4. Classification. Support Vector Machines (SVMs) are currently acknowledged as the leading general discriminative approach for machine learning problems in a number of domains. In SVM classification, a training example is represented using a vector of features x_i and a label $y_i \in \{1, -1\}$. The algorithm tries to find the optimal separating hyperplane that predicts the labels from the training examples.

Since data is typically not linearly separable, it is mapped to a higher dimensional space by a kernel function. We use a Radial Basis Function (RBF) kernel with parameter γ:

$$K\left(x_i, x_j\right) = e^{\left(-\gamma |x_i - x_j|^2\right)}, \quad \gamma > 0. \tag{1}$$

Using the kernel function, the C-SVC SVM algorithm finds the optimal hyperplane by solving the dual optimization problem:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{2}$$

subject to

$$0 \le \alpha_i \le C, \quad i = 1, \ldots, N,$$
$$y^T\alpha = 0. \tag{3}$$

Q is a $N \times N$ matrix defined as $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ and e is the vector of all ones. C is a cost parameter that controls the penalty of misclassified instances given linearly nonseparable data.

This binary classification problem can be extended to multiclass using either the *one versus one* or the *one versus all* approach. The first trains a classifier for each pair of classes, while the second trains a classifier for each class using examples from all the other classes as negative examples. The *one versus one* method has been found to perform generally better for many problems [25]. Our initial experiments with the *one versus all* approach further confirmed this for our problem, and thus we use the *one versus one* approach in our experiments. We use the *libsvm* [26] implementation of C-SVC. Suitable values for C and γ are found through grid search with a portion of training examples for each experiment.

4.5. Detection of Salient Events in Longer Recordings. In order to aid sound design by quickly identifying regions of basic events in a large audio file, we apply the SVM classifier to fixed-size windows taken from the input sound and grouping consecutive windows of the same class into segments. One tradeoff in fixed window segmentation schemes is the window size, which basically trades confidence in classification accuracy for temporal accuracy of the segment boundaries and noise in the segmentation. Based on a similar segmentation problem presented in [27], we first segment the audio into two second windows with one second of overlap and assign a class to each window by classifying it with the SVM model. The windows are multiplied with a Hamming window function:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right). \tag{4}$$

The SVM multiclass model we employ returns both the class label and an associated probability, which we compare with a threshold in order to filter out segmentation frames that have a low-class probability and are thus susceptible to being misclassified.

In extension to the prewindowing into fixed-sized chunks as described above, we consider a second segmentation scheme, where windows are first centered on onsets found in a separate detection step and then fitted between the onsets

with a fixed hop size. The intention is to heuristically improve localization of impacts and other acoustic events with transient behavior. The onset detection function is computed from differences in high-frequency content and then passed through a threshold function to obtain the onset times.

5. Classification Experiments

5.1. Overview. We now describe several experiments performed using the classification approach and sound collections described in the previous section. Our first experiment consists in the classification of music, speech, and environmental sounds. We then focus on the last group to classify it using the terms of the taxonomy.

We first evaluate the performance of different sets of features, by adding temporal descriptors of frame level features to both MFCC and the custom set obtained using feature selection. Then we compare two possible approaches to the classification problem: a *one versus one* multiclass classifier and a hierarchical classification scheme, where we train separate models for the top level classes (solids, liquids, and gases) and for each of the top level categories (i.e., for solids we train a model to discriminate impacts, scraping, rolling, and deformation sounds).

Our general procedure starts by resampling the whole database in order to have a balanced number of examples for each class. We then evaluate the class models using ten-fold cross-validation. We run this procedure ten times and average the results in order to account for the random resampling of the classes with more examples. We estimate the parameters of the model using grid search only in the first iteration in order to avoid overfitting each particular sample of the data.

5.2. Music, Speech, and Voice Classification. We trained a multiclass SVM model for discriminating music, voice, and speech, using the collections mentioned in Section 4. While this classification is not the main focus of this paper, this step was necessary in order to focus our prototypes on environmental sounds. Using the full stacked set of descriptors (thus without the need of any specific musical descriptor) we achieved 96.19% of accuracy in cross-validation. Preliminary tests indicate that this model is also very good for discriminating the sounds at *Freesound*.

5.3. Classification of Sound Events. For the comparison of features, we generated several sets of features by progressively adding derivatives, attack and decay, and temporal descriptors to the two base sets. Figure 4 shows the average f-measure for each class using MFCC as frame-level descriptors, while Figure 5 shows the same results using the descriptors chosen by feature selection. In general, the latter set performs better than MFCC. With respect to temporal descriptors, they generally lead to better results for both sets of features. Impulsive sounds (*impact*, *explosion*, and *woosh*) tend to benefit from temporal descriptors of the second set of features. However, in general adding these descriptors does not seem to change the balance between the better detected classes and the more difficult ones.

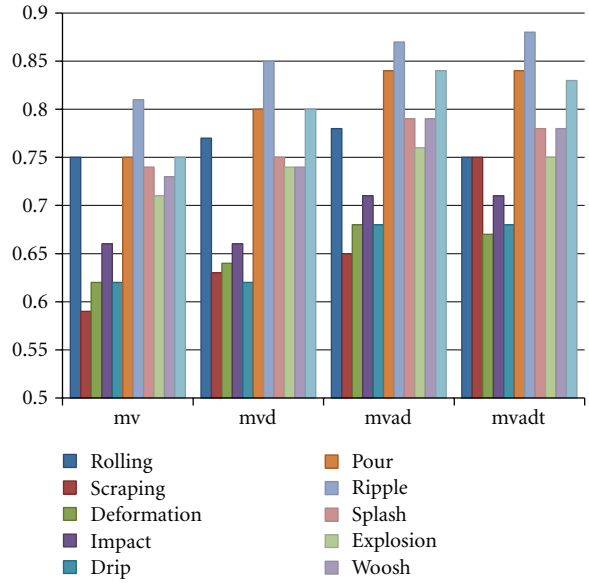

Figure 4: Average f-measure using MFCC as base features.

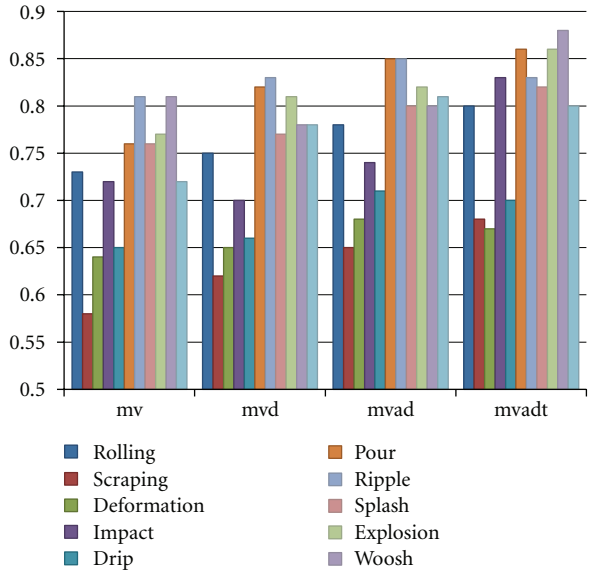

Figure 5: Average f-measure using our custom set of features.

Table 3: Average classification accuracy (%) for direct versus hierarchical approaches.

Method	accuracy
Direct	84.56
Hierarchical	81.41

5.4. Direct versus Hierarchical Classification. For the comparison of the hierarchical and direct approach, we stack both sets of descriptors described previously to obtain the best accuracy (Table 3). While in the hierarchical approach more

Table 4: Confusion matrix of one cross-validation run of the direct clasifier.

	rolling	scraping	deformation	impact	drip	pour	ripple	splash	explosion	woosh	wind
rolling	88	3	7	0	2	0	0	0	0	0	0
scraping	6	71	11	1	1	0	1	2	1	4	2
deformation	3	10	73	3	2	1	1	5	1	0	1
impact	1	1	5	89	1	0	0	0	1	2	0
drip	3	3	3	2	70	7	4	2	1	3	2
pour	1	0	2	1	5	87	0	4	0	0	0
ripple	2	1	0	0	6	0	87	1	0	0	3
splash	1	2	5	0	0	2	2	87	1	0	0
explosion	2	1	1	3	1	0	0	0	89	1	2
woosh	1	1	1	1	2	0	0	0	4	90	0
wind	3	1	0	0	3	0	5	0	2	0	86

Table 5: Confusion matrix of one cross-validation run of the hierarchical clasifier.

	rolling	scraping	deformation	impact	drip	pour	ripple	splash	explosion	woosh	wind
rolling	82	11	2	1	1	0	2	0	0	1	0
scraping	6	73	8	4	4	0	0	1	0	3	1
deformation	4	11	73	3	4	0	2	0	1	1	1
impact	1	3	2	84	0	0	0	1	5	4	0
drip	2	3	3	2	70	6	4	3	0	6	1
pour	1	0	1	0	2	91	0	4	0	1	0
ripple	0	0	1	0	6	0	85	2	0	1	5
splash	2	3	4	1	0	2	4	81	1	0	2
explosion	2	3	1	5	0	0	0	0	84	3	2
woosh	1	3	1	4	0	0	0	0	2	87	2
wind	5	2	1	0	4	0	4	0	2	1	81

classification steps are performed, with the corresponding accumulation of errors, results are quite similar to the direct classification approach. Tables 4 and 5 show confusion matrices for one cross-validation run of the hierarchical and direct approach respectively. The first level of classification in the hierarchical approach does not seem to help in the kind of errors that occur with the direct approach, both accumulate most errors for *scraping*, *deformation*, and *drip* classes. Most confusions happen between the first two and between *drip* and *pour*, that is, mostly in the same kind of material. This seems to imply that some common features allow for a good classification of the top level. In this sense, this classifier could be interesting for some applications. However, for the use cases presented in this work, we use the direct classification approach as it is simpler and produces less errors.

The results of the classification experiments show that a widely available and scalable classifier like SVMs, general purpose descriptors, and a simple approach to describing their temporal evolution may suffice to obtain a reasonable result for such a general set of classes over noisy datasets. We now describe two use cases were these classifiers can be used. We use the direct classification approach to rank sounds according to their probability to belong to one of the classes. The rank is obtained by training the multiclass model to support probability estimates [26].

6. Use Cases

The research described in this paper was motivated by the requirements of virtual world sonification. Online interactive environments, such as virtual worlds or games have specific demands with respect to traditional media. One would expect content to be refreshed often in order to avoid repetition. This can be achieved, on the one hand, by using dynamic models instead of preset recordings. On the other hand, sound samples used in these models can be retrieved from online databases and field recordings. As an example, our current system uses a graph structure to create complex patterns of sound objects that vary through time [28]. We build a model to represent a particular location, and each event is represented by a list of sounds. This list of sounds can be extended and modified without modifying the soundscape generation model.

Content-based search on user-contributed databases and field recordings can help to reduce the cost of obtaining new sounds for such environments. Since the popularization of digital recorders, it has become easy and convenient to record environmental sounds and share this recordings. However, cutting and labeling field recordings can be a tedious task, and thus often only the raw recording is uploaded. Automatic segmentation of such recordings can help to maximize the amount of available sounds.

In this section, we present two use cases where the presented system can be used in the context of soundscape design. The first prototype is a content-based web search system that integrates the model classifiers as a front-end of the *Freesound* database. The second prototype aims to automatically identify the relevant sound events in field recordings.

6.1. Sound Event Search with Content-Based Ranking. Current limitations of searching in large unstructured audio databases using general sound event concepts have been already discussed in Section 3. We implemented a basic prototype to explore the use of the Gaver taxonomy to search sounds in the *Freesound* database. We compare here the use of the classifier described in Section 4 to rank the sounds to the search method currently used by the site.

The prototype allows to articulate a two-word query. The basic assumption is that two words can be used to describe a sound event, one describing the main object or material perceived in the sound, and the other describing the type of interaction. The current search engine available at the site is based on the classic Boolean model. An audio clip is represented by the list of words present in the text description and tags. Given a multiword query, by default, documents containing all the words in the query are considered relevant. Results are ranked according to the number of downloads, so that the most popular files appear first.

In the content-based method, sounds are first classified as voice, music, or environmental sound using the classifier described in Section 5.2. Boolean search is reduced to the first word of the query, and relevant files are filtered by the content-based classifier, which assigns both a class label from the taxonomy and a probability estimate to each sound. Thus, only sounds where the label corresponds to the second term of the query are returned, and the probability estimate is used to rank sounds. For example for the query *bell* + *impact*, sounds that contain the word *bell* in the description and that have been classified as *impact* are returned, sorted by the probability that the sound is actually an impact.

For both methods, we limit the search to sounds shorter than 20 seconds in order to filter out longer field recordings. Figure 6 shows the GUI of the search prototype.

We validated the prototype by means of a user experiment. We selected a number of queries by looking at the most popular searches in *Freesound*. These were all single word queries, to which we appended a relevant term from the taxonomy. We removed all the queries that had to do with music and animal voices, as well as the ones that would return no results in some of the methods. We also removed all queries that mapped directly to terms of the taxonomy, except for *wind*, which is the most popular search of the site. Also we repeated the word *water* in order to test two different liquid interactions. We asked twelve users to listen to the results of each query and subjectively rate the relevance of the 10 top-ranked results obtained by the two retrieval methods described before. The instructions they received contained no clue about the rationale of the two methods used to generate the lists of sounds, just that they

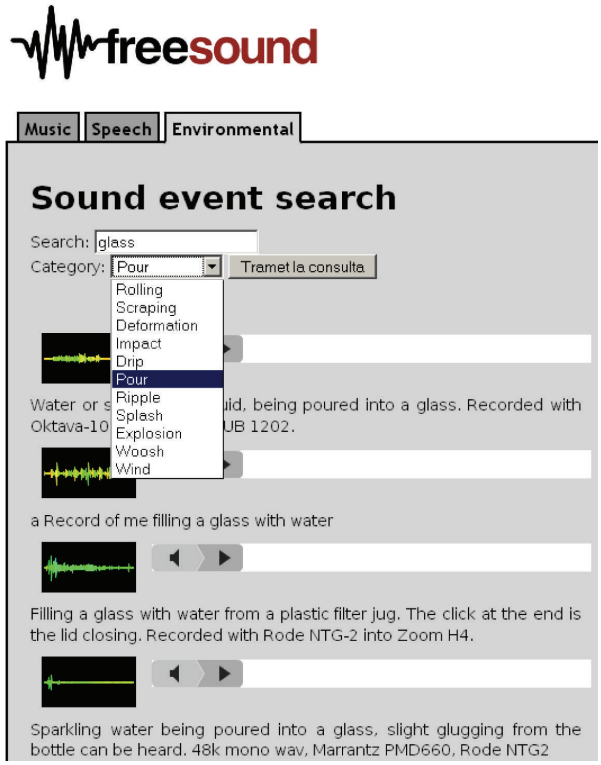

Figure 6: Screenshot of the web-based prototype.

were obtained using different methods. Table 6 contains the experiment results, showing the average number of relevant sounds retrieved by both methods. Computing the precision (number of relevant files divided by the number of retrieved files), we observe that the content-based method has a precision of 0.638, against the 0.489 obtained by the text-based method. As mentioned in Section 3.2, some categories are scarcely represented in *Freesound*. Hence, for some queries (e.g., *bell* + *impact*), the content-based approach returns more results than using the text query. The level of agreement among subjects was computed as the Pearson correlation coefficient of each subject's results against the mean of all judgments, giving an average of $r = 0.92$. The web prototype is publicly available for evaluation purposes (http://dev.mtg.upf.edu/soundscape/freesound-search).

6.2. Identification of Iconic Events in Field Recordings. The process of identifying and annotating event instances in field recordings implies listening to all of the recording, choosing regions pertaining to a single event, and finally assigning them to a sound concept based on subjective criteria. While the segmentation and organization of the soundscape into relevant sound concepts refers to the cognitive and semantic level, the process of finding audio segments that fit the abstract classes mainly refers to the signal's acoustic properties. Apart from the correct labeling, what is interesting for the designer is the possibility to quickly locate regions that are contiguously labeled with the same class, allowing him/her to focus on just relevant segments rather than on

Table 6: Results of the user experiment, indicating the average number of relevant results for all users. We indicate in brackets the number of retrieved results for each query.

word + term	Content-based	Text-based
wind + wind	6.91 (10)	0.91 (10)
glass + scraping	4.00 (10)	4.00 (5)
thunder + explosion	5.36 (10)	5.36 (10)
gun + explosion	9.09 (10)	4.45 (10)
bell + impact	7.18 (10)	1.55 (3)
water + pour	8.73 (10)	6.64 (10)
water + splash	8.82 (10)	6.91 (10)
car + impact	2.73 (10)	1.27 (4)
door + impact	8.73 (10)	0.73 (4)
train + rolling	2.27 (10)	1.00 (1)

Table 7: Normalized segment overlap between segmentation and ground truth for the onset-based and the fixed-window segmentation schemes.

	Onset-based	Fixed-window
Normalized segment overlap	20.08	6.42

the entire recording. We try to help automating this process by implementing a segmentation algorithm based on the previously trained classification models. Given a field recording, the algorithm generates high-class probability region labels. The resulting segmentation and the proposed class labels can then be visualized in a sound editor application (http://www.sonicvisualiser.org/).

In order to compare the fixed window and the onset-based segmentation algorithms, we split our training collection described in Section 4 into training and test sets. We used the former to train an SVM model and the later to generate an evaluation database of artificial concatenations of basic events. Each artificial soundscape was generated form a ground truth score that described the original segment boundaries. The evaluation measure we employed is the overlap in seconds of the found segmentation with the ground truth segmentation for the corresponding correctly labeled segment, normalized by the ground truth segment length. With this measure, our onset-based segmentation algorithms performs considerably better than the fixed-size window scheme (Table 7). In all our experiments we used a window size of two seconds and an overlap of one second.

Figure 7 shows the segmentation result when applied to an artificial sequential concatenation of basic interaction events like scraping, rolling and impacts. The example clearly shows that most of the basic events are being identified and classified correctly. Problems in determining the correct segment boundaries and segment misclassifications are mostly due to the shift variance of the windowing performed before segmentation, even if this effect is somewhat mitigated by the onset-based windowing.

Since in real soundscapes basic events are often not identifiable clearly—not even by human listeners—and recordings usually contain a substantial amount of background noise, the segmentation and annotation of real recordings

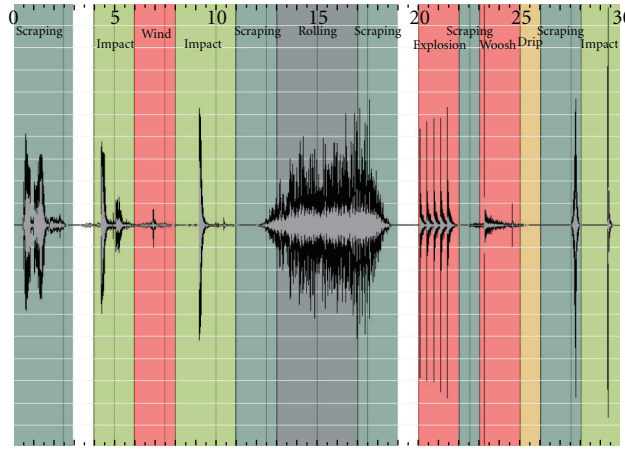

Figure 7: Segmentation of an artificial concatenation of basic events with a window length of two seconds with one second overlap and a class probability threshold of 0.6.

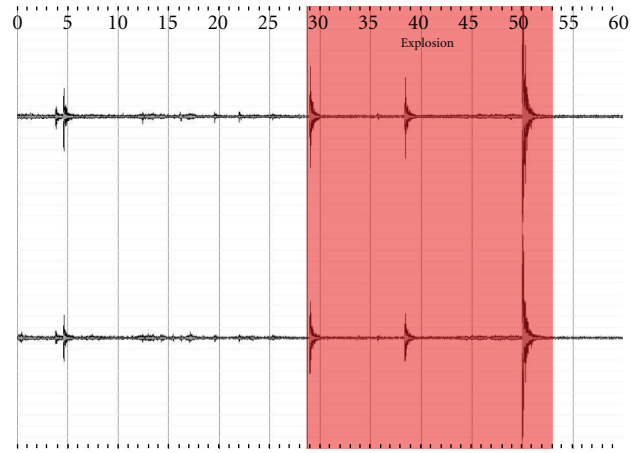

Figure 8: Identification of basic events in a field recording of firecracker explosions with a window length of two seconds with one second overlap using the onset-based segmentation algorithm and a class probability threshold of 0.6.

is a more challenging problem. Figure 8 shows the analysis of a one-minute field recording of firecracker explosions. Three of the prominent explosions are located and identified correctly, while the first one is left undetected.

Although the output of our segmentation algorithm is far from perfect, this system has proven to work well in practice for certain applications, for example, for quickly locating relevant audio material in real audio recordings for further manual segmentation.

7. Conclusions and Future Work

In this paper we evaluated the application of Gaver's taxonomy to unstructured audio databases. We obtained surprisingly good results in the classification experiments, taking into account for the amount of noisy data we included. While our initial experiments were focused on very specific

recordings such as the ones in the *Sound Events* dataset, adding more examples allowed us to generalize to a wide variety of recordings. Our initial prototype shows the potential of using high level concepts from ecological acoustics for interfacing with online repositories. Still, we consider this topic an open issue. For example the use of the taxonomy in more explorative interfaces should be further analyzed, for example, by further clustering the sounds in each class, or by relating the taxonomy to existing concepts in folksonomies. The use of content-based methods using these concepts should also be evaluated in the context of traditional structured audio repositories. With respect to segmentation of longer field recordings, the presented method showed potential to aid the identification of interesting segments for the synthesis of artificial soundscapes. However, it could use further improvements in order to make it more robust to background noise. It also should be further adapted to use different temporal resolutions for each class.

Acknowledgment

This paper was partially supported by the ITEA2 Metaverse1 (http://www.metaverse1.org/) project.

References

[1] Universitat Pompeu Fabra, "Repository of sound under the Creative Commons license," Freesound.org, 2005, http://www.freesound.org/.

[2] J. Janer, N. Finney, G. Roma, S. Kersten, and X. Serra, "Supporting soundscape design in virtual environments with content-based audio retrieval," *Journal of Virtual Worlds Research*, vol. 2, October 2009, https://journals.tdl.org/jvwr/article/view/635/523.

[3] E. Martínez, Ò. Celma, B. De Jong, and X. Serra, *Extending the Folksonomies of Freesound.org Using Content-Based Audio Analysis*, Porto, Portugal, 2009.

[4] W. W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, pp. 1–29, 1993.

[5] P. Schaeffer, *Trait'e des Objets Musicaux*, Editions du Seuil, Paris, France, 1st edition, 1966.

[6] D. Rocchesso and F. Fontana, Eds., *The Sounding Object*, Edizioni di Mondo Estremo, 2003.

[7] M. Casey, "General sound classification and similarity in mpeg-7," *Organised Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[8] T. Zhang and C. C. J. Kuo, "Classification and retrieval of sound effects in audiovisual data management," in *Proceedings of the Conference Record of the 33rd Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 730–734, 1999.

[9] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, N. Wack, and P. Herrera, "Nearest-neighbor automatic sound annotation with a WordNet taxonomy," *Journal of Intelligent Information Systems*, vol. 24, no. 2-3, pp. 99–111, 2005.

[10] M. Slaney, "Semantic-audio retrieval," in *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. IV/4108–IV/4111, May 2002.

[11] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st International ACM Conference on Multimedia Information Retrieval (MIR '08)*, pp. 105–112, ACM, New York, NY, USA, August 2008.

[12] J. J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[13] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with timeFrequency audio features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, Article ID 5109766, pp. 1142–1158, 2009.

[14] N. J. Vanderveer, *Ecological acoustics: human perception of environmental sounds*, Ph.D. dissertation, Cornell University, 1979.

[15] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[16] C. Guastavino, "Categorization of environmental sounds," *Canadian Journal of Experimental Psychology*, vol. 61, no. 1, pp. 54–63, 2007.

[17] V. Maffiolo, *De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain*, Ph.D. dissertation, Universite du Mans, Le Mans, France, 1999.

[18] R. Murray Schafer, *The Tuning of the World*, Knopf, New York, NY, USA, 1977.

[19] C. Fellbaum et al., *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Mass, USA, 1998.

[20] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*, John Wiley & Sons, 2005.

[21] R. Kohavi and G. H. John, *Wrappers for Feature Subset Selection*, 1996.

[22] A. Minard, N. Misdariis, G. Lemaitre et al., "Environmental sound description: comparison and generalization of 4 timbre studies," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '02)*, p. 6, New York, NY, USA, February 2008.

[23] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception and Psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

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

[25] C. wei and C.-J. Lin, "A comparison of methods for multi-class support vector machines," 2001.

[26] C. C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[27] K. Lee, *Analysis of environmental sounds*, Ph.D. dissertation, Columbia University, Department of Electrical Engineering, 2009.

[28] A. Valle, V. Lombardo, and M. Schirosa, "A framework for soundscape analysis and re-synthesis," in *Proceedings of the 6th Sound and Music Computing Conference (SMC '09)*, F. Gouyon, A. Barbosa, and X. Serra, Eds., pp. 13–18, Porto, Portugal, July 2009.