CONTENT-BASED RETRIEVAL FROM UNSTRUCTURED AUDIO DATABASES USING AN ECOLOGICAL ACOUSTICS TAXONOMY

Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera

Universitat Pompeu Fabra
Music Technology Group
Roc Boronat 138, Barcelona
firstname.lastname@upf.edu

ABSTRACT

In this paper we describe a method to search for environmental sounds in unstructured databases with user-submitted material. The goal of the project is to facilitate the design of soundscapes in virtual environments. We analyze the use of a Support Vector Machine (SVM) as a learning algorithm to classify sounds according to a general sound events taxonomy based on ecological acoustics. In our experiments, we obtain accuracies above 80% using cross-validation. Finally, we present a web prototype that integrates the classifier to rank sounds according to their relation to the taxonomy concepts.

1. INTRODUCTION

Virtual environments based on realistic simulations of physical space are becoming a common use of the Internet. Most of them can be divided among multiplayer games and social environments used to meet and chat. In some cases, people have even become interested in purchasing virtual goods and hence virtual economies have emerged. However, the cost of designing such amount of 3D spaces is very high. Virtual environments have followed the trend towards user-centered technologies that dominates the web. Many programs allow users to create and upload their own models and to design their spaces. Sites such as *Google 3D Warehouse* are available as centralized repositories of 3D models that can be placed in different environments.

So far, these environments offer very sophisticated visual simulations but quite basic audio functionality. Still, applications like *Second Life* allow users to upload custom sounds for objects. In this context, open, user-contributed sound repositories such as *freesound.org* [1] can be used to improve the acoustic experience of virtual environments. However, searching for sounds in user-contributed databases is still problematic. Sounds are often insufficiently annotated and with very diverse vocabularies [2]. Some sounds are isolated and segmented, but others consist of very long recordings containing mixtures of environmental sounds. In this situation, content-based tools can help improving the search and retrieval of sounds. For specific domains, such tools can be based on *a priori* knowledge and constraints. For the case of designing the acoustic experience of virtual environments, we do not consider voice and music audio, for which separate channels are typically allocated, i.e. real user voices for avatars and background music streams. With respect to indexing and retrieval of environmental sounds for virtual spaces, we are interested in categorizations that take into account the way we perceive everyday sounds. In this context, the ideas of Gaver have become commonplace. In [3], he emphasized the distinction between musical listening (as defined by Schaeffer [4]) and everyday listening. He also devised a comprehensive taxonomy of everyday sounds based on the workings of ecological acoustics, while pointing at the problems with traditional organization of sound effects libraries. The taxonomy categorizes sounds according to the type of interacting materials (solids, liquids, gases) and the kind of interaction (e.g for solid bodies, sounds are classified as impact, deformation, scraping or rolling). One example of the use of this taxonomy can be found in the *Closed* project [5], where it was used to develop its physically-based sound models [6].

In this paper, we analyze the use of this taxonomy for retrieving audio from unstructured, user-contributed audio repositories. We test different approaches to description and classification of sounds according to this taxonomy using SVM. We then use the learnt models to rank sounds. In addition to the traditional text search interface, we add the option to filter and sort the results according to a category in the taxonomy. This interface makes it easier to retrieve iconic sounds that represent basic auditory events. However, the focus of this paper is the classification part, and we don't formally evaluate the search interface.

2. RELATED WORK

Automatic analysis and categorization of environmental sounds has been traditionally related to management of sound effects libraries. The taxonomies used in these libraries usually do not attempt to provide a comprehensive organization of sounds. 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 representation contexts such as video 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 [7], [8] when using Hidden Markov Models (HMM) on rather specific classes of sound effects. There are two problems with this kind of approach. On the one hand, working with non comprehensive taxonomies omits the fact that real world applications will typically need to deal with much larger vocabularies. Many of these approaches may be difficult to scale to vocabularies and databases orders of magnitude larger. On the other hand, they commonly employ small databases of sounds recorded and edited under controlled conditions. This means that it is unclear how these methods would generalize to noisier environments and databases. In particular, we are concerned with user-contributed media, which involves a wide variety of situations, recording equipment, motivations and skills.

Previous research works have explored the scalability issue by using more efficient classifiers. For example in [9], the problem of extending content-based classification to thousands of labels was approached by using a nearest-neighbor classifier. The system presented in [10] bridges the semantic space to 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 noise and real world recordings, another trend of work has been directed towards the 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 detection of sonic context for mobile phones.

In a way, our problem shares some characteristics of both sound effects and environmental sound classification. This situation comes from the different perceptions and motivations of users at a site like *freesound.org*. Some users will upload sound effects, and many users are interested especially in downloading clean sound effects for using them in music or multimedia productions. But also it is common to upload raw field recordings of different locations and situations as a way to share personal experiences.

The specification of a general taxonomy for environmental sounds remains an elusive problem. Gaver's taxonomy [3] organizes sounds according to how the mechanics of the production of sounds are perceived, from the point of view of ecological acoustics. Further research has given some support to his hypothesis with respect to the perception and categorization of environmental sounds [13]. The taxonomy offers a coherent and general categorization of environmental sounds that is well defined for simple auditory events. However, despite being frequently cited, we don't know of other attempts at automatic classification using this taxonomy.

Gaver proposed a hierarchical classification space, from broad classes to simple sonic events (see figure 1). The root class can be called *Interacting Materials*, since most generally sounds are produced as a result of an interaction of materials. At the next level, the taxonomy divides sounds in three general categories: those involving vibrating solids, gases and liquids. Finally, basic level sonic events are shown at the third level, they are defined by the simple interactions that can cause solids, gases and liquids to produce sound.

3. CLASSIFICATION OF ENVIRONMENTAL SOUNDS

3.1. Overview

We analyze the use of Gaver's taxonomy for general audio segments databases by training and testing a Support Vector Machine (SVM) classifier over a dataset collected from various sources. Our first experiment consists in comparing the performance of different sets of features for the task, in order to assess the importance of describing temporal evolution. A second experiment analyzes two different definitions of the classification problem: as a hierarchical classification or as a direct multiclass problem.

3.2. Datasets creation

For our experiments, we manually selected and labeled sounds according to the taxonomy's categories. We created three datasets: *SoundEvents*, *SoundFx* and *Freesound*.

The *SoundEvents* dataset [14] 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 recordings are repeated multiple times. For example, the sound of metal balls running on plywood is recorded several times in the same session. We discarded the sounds that would correspond to complex events due to the interactions of different materials. A second dataset was collected from a number of sound effect libraries, with different levels of quality. A small number of sounds in this dataset was downloaded from online repositories. Sounds in this dataset generally try to provide a good representation of specific sounds. For instance, for the *explosion* category we select 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. Finally, our third dataset consists of sounds downloaded from *freesound.org*. This set is the noisiest of the three, as sounds are recorded in very different conditions and situations. Most contain background noise and many are not segmented with the purpose of isolating a particular sound event.

The collection of sounds in the dataset presented a number of issues. We now describe the main criteria we used in order to provide a coherent interpretation of the taxonomy.

First of all, in order to allow our classifier to generalize to user-submitted audio, we needed to search sounds with a variable recording setup, recording quality and relative microphone position. Second, we needed to search samples with a less stringent segmentation than the one used in the *SoundEvents* database, where the researchers tried to include just one instance of a basic event in each recorded sample. Thus we considered samples presenting: i) complex temporal pattern repetition of basic events, ii) sounds generated from compound interaction and iii) sounds generated by hybrid interaction. Compound interaction happens when a sound results from the interaction between more than one type of basic event. This is the case of specific door locks, where the sound is generated by a mix of impacts, deformations and scrapings; or the case of missiles, where the sound is generated both by whoosh and explosion. Contrastingly, the sound generated by hybrid interaction happens when a given material interacts with one of a different kind, as in the case of the hybrid event *impact-drip*, when water drips onto a solid surface, or in the case of bubbles, a hybrid between liquid and gas. In order to extend the dataset, we included sound instances that still can be classified at the basic level, but under somewhat less restrictive constraints: repetition patterns of atomic events of the same type, samples containing only a tiny amount of compound or hybrid interactions and samples representing different microphone positions, recording setups and noise conditions.

The taxonomy provides also some parameters related to source attributes that are percieved through sounds. These parameters were useful to qualitatively determine whether samples belonged to a class or not. Some examples: rain was a problematic case, following the original definition, the *Drip* basic event is just water falling into water, with the parameters *viscosity*, *object size*, *object shape* and *force*. In comparison, the parameters of the *Pour* basic event are *viscosity*, *amount* and *height*. Depending on the par-

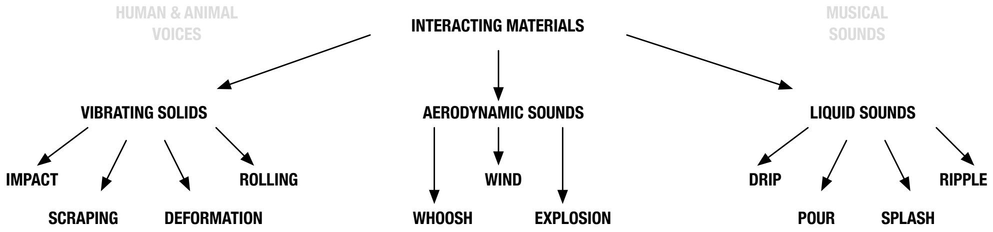


Figure 1: Representation of the Gaver taxonomy.

ticular sample analysed, rain could belong to *Drip*, if individual raindrops were clearly identifyable, or to *Pour*, if the temporal fine structure was undiscernible and the signal closer to noise. Also, if the sound clearly indicated water falling down on a surface, as in the case of rain tapping on a window, the sample was considered to be a hybrid event and discarded, but if the rain contact with the surface was faint, we included it in the *Liquids* category, and it still needed to be categorized into either *Drip* or *Pour*.

3.3. Audio Features

One important question in the discrimination of general auditory events is how much of our ability comes from extracting properties of the instantaneous spectrum, and how much results from 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 (quoted in [3]). In several fields involved with discrimination 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 we think that time varying aspects of the sound are determinant to recognize different classes. This is especially true for impulsive classes such as impacts and explosions or splashes, and to a lower extent by classes that imply some regularity, like rolling.

In this paper we analyze the performance of some descriptors extracted from the time series of frame level descriptors for our classification task. We test two sets of frame-level features:

- *MFCC*: An implementation of Mel Frequency Cepstrum Coefficients using 40 bands and 13 coefficients.
- *Spectral*: A collection of spectral shape descriptors such as spectral centroid, kurtosis, skewness, crest, decrease and rolloff, along with an estimation of pitch and pitch salience.

We use MFCCs as a reference as they are one of the most commonly used representations of the spectrum. Our second set includes descriptors of the spectral shape that were popularized by the MPEG-7 standard [15]. We also include an estimation of pitch and pitch salience, which have been shown to be relevant for the discrimination of environmental sounds [13]. We compute the mean and variance of every frame level descriptor, as well as mean and variance of its first and second derivative. We also test several descriptors computed from the temporal evolution of frame level features, such as the log attack time, a measure of decay [16] and temporal descriptors derived from statistical moments: temporal centroid, kurtosis and skewness (table 1).

Name	Description	# desc.
mv	mean and variance	2
mvd	*mv*, derivatives	6
mvdad	*mvd*, log attack time and decay	8
mvdadt	*mvdad*, temp. centroid, kurtosis, skewness	9

Table 1: Sets of descriptors extracted from the temporal evolution of frame-level features, and the number of descriptors per frame level feature.

Features	*mv*	*mvd*	*mvdad*	*mvdadt*
MFCC	69.35	75.76	74.98	77.80
Spectral	73.17	78.04	80.02	81.29

Table 2: Average classification accuracy (%) for different sets of features.

3.4. Experiments

We use a Support Vector Machine (SVM) classifier [17] in order to assign a given feature vector representing one sound to one of the classes in the taxonomy. Our first experiment consists in an evaluation the performance of the temporal descriptors applied only to MFCC features. We repeatedly evaluate a *one vs one* multiclass SVM classifier using a set of MFCC descriptors where we progressively add temporal evolution descriptors. We then repeat the procedure with the second set of descriptors and compare the results.

The second experiment consists in comparing the *one vs one* classifier to 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). For this experiment we combine both MFCC and spectral shape features with their corresponding temporal descriptors.

Our general procedure starts by resampling the whole database in order to have a balanced number of examples for each class. We then evaluate using ten-fold cross-validation. We run this procedure five 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.

Method	accuracy
Direct	84.10
Hierarchical	80.61

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

4. RESULTS

Table 2 shows the accuracy of the multiclass SVM model for each set of descriptors. While the most important improvement is typically obtained by adding derivatives, the experiment shows that adding the temporal descriptors does help in the discrimination of the different kinds of event. This is true for both MFCC and spectral shape descriptors. On the other hand, it shows that it is possible to obtain reasonably good results with a simple and scalable approach to the description of the temporal evolution of the spectrum. Our best result is obtained when combining both descriptor sets (table 3). As a further step, we plan to compare this results with more complex approaches such as HMM.

Table 3 shows the comparison of the hierarchical approach to the direct classification. While in the hierarchical approach more classification steps are performed, with the corresponding accumulation of error, results are still above 80% on average. This seems to support the underlying hierarchy in Gaver's proposal, in the sense that basic events involving a main class of materials (solid liquid or gas) share some features and can be discriminated from other main classes. This approach has the advantage of providing a model for the top level and consistent models for the lower levels, which may be used for browsing sound databases.

The results of the classification experiments show that a widely available and scalable classifier like SVM may suffice to obtain a reasonable result for such a general set of classes over noisy datasets. Next, we describe the use of the direct approach to rank sounds in the *freesound.org* database. The rank is obtained by training the multiclass model to support probability estimates [17]. The probability estimate is then used as a rank for a query containing one concept of the taxonomy.

5. APPLICATION

A principal objective of the present research is to facilitate the search of environmental sounds in user-contributed audio databases. With that purpose, we integrated the SVM models as a front-end for querying the *freesound.org* database with a combination of textual input and terms from the ecological acoustics taxonomy. First, we review how the taxonomy under study is currently represented as metadata in the *freesound.org* database by social tags.

5.1. Taxonomy concepts in the Freesound folksonomy

Since its inception in 2005, *freesound.org* has become a renowned repository of sounds with non-commercial license, building an active online community at the same time. Currently, it stores $84,222$ sounds, labeled with approximately 18000 unique tags. Sounds are collaboratively labeled with tags, a practice known as folksonomy [18], leading to an unstructured audio database.

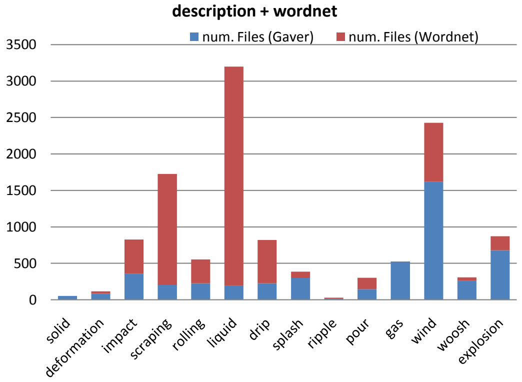


Figure 2: Number of sound files in *freesound.org* containing tags or descriptions with Gaver taxonomy's terms (in red), or their synonyms from Wordnet (in blue).

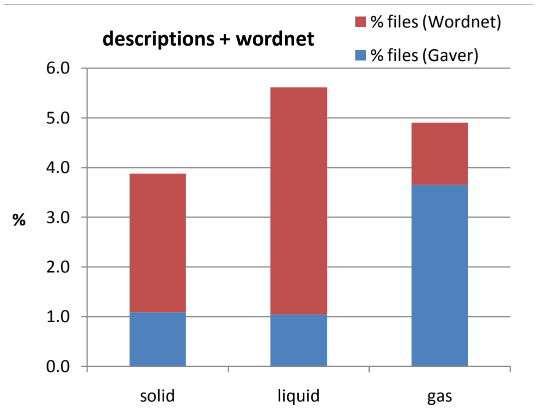
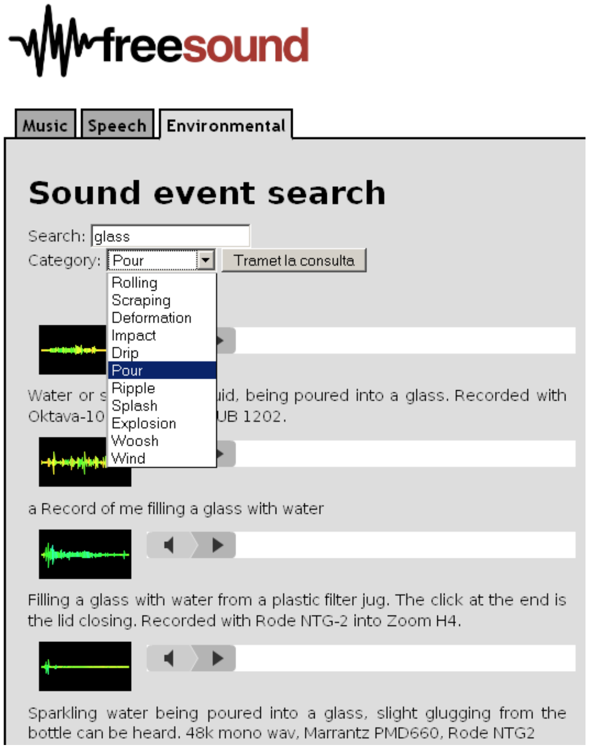


Figure 3: Percentage of sound files in Freesound containing tags or descriptions with Gaver taxonomy's terms (in red), or their synonyms from Wordnet (in blue). Results are grouped by top categories (solid, liquid and gas).

Looking at the database content, one can distinguish three main types of sounds: environmental (e.g. nature recordings), musical (e.g. instrument samples, loops) and speech (e.g. individual, conversational).

Regarding the environmental sounds tagged in *freesound.org*, the presence of the studied ecological acoustics taxonomy is scarce. Figure 2 shows the histogram of the taxonomy's terms (in blue), grouped by the top-level categories (solid, liquid and gas). In order to widen the search, we aggregated to each term various semantically-related tags that appear as a *synset* (synonym set) in the Wordnet database [19]. The number of files retrieved with the *s*ynset are shown in red. Figure 3 shows the histogram of files grouped by the top categories in the taxonomy, i.e. solid, liquid and gas. In this case, values indicate the percentage of files compared to the total files in the database $(84,222)$. Patently, the concepts used in ecological acoustics are infra-represented in the the folksonomy. Next section describes a practical application of how content-based retrieval can assist the search with these concepts.



Sound event search

Search: glass

Category: Pour

Water or s… …uid, being poured into a glass. Recorded with Oktava-10… …UB 1202.

a Record of me filling a glass with water

Filling a glass with water from a plastic filter jug. The click at the end is the lid closing. Recorded with Rode NTG-2 into Zoom H4.

Sparkling water being poured into a glass, slight glugging from the bottle can be heard. 48k mono wav, Marrantz PMD660, Rode NTG2

Figure 4: Screen-shot of the web-based prototype.

5.2. Extending text-queries with a content-based classifier

In this application, we are only concerned with environmental sounds. Hence, the retrieval of musical and speech sounds would have a negative effect on our search results. A pre-process to automatically classify these three sound categories is currently beyond the scope of this paper. Instead, here we opted for filtering the dataset using tags related to environmental sounds in order to reduce the retrieval of musical and speech sounds. The final subset is built by departing from the “field-recording” tag, which has become one of the most general of the database. We consider this tag as one of the main themes of the site along with “voice” and “loop”, which respectively represent speech and music samples. We compute the cosine distance of all tags to the “field-recording” tag and keep all tags within a distance below an empirically determined threshold. We limit the search to files labeled with these tags. Also, we limit the file duration to 10 seconds in order to avoid the retrieval of long soundscape recordings, reducing the whole database to 1934 sounds.

On this restricted dataset, our prototype allows searching by ecological acoustics terms. A free word is input in a text box, and a term from the taxonomy is selected from a list. The idea behind this scheme is that the query is composed of a *subject* represented by the free text word, and a *predicate* represented by a class of event. The query is matched to tags and descriptions, and the results are ranked according to the output of the automatic classifier for the given class, as described in section 3. Figure 4 shows the GUI of the search prototype.

5.3. Discussion

While we didn’t formally evaluate the search interface, we informally analyzed its viability by trying several common queries composed of a word plus a term of the taxonomy. We compared the results to the ones returned from multi-word queries by the regular search engine at *freesound.org*, which matches text queries to tags and descriptions, and ranks the results by popularity (number of downloads). We observed that for some queries (e.g. glass+pour) the content-based approach represents a significant improvement over the traditional text-based search. In many cases, content-based indexing helps reducing the effects of ambiguity and incomplete or noisy text descriptions. As a side effect, we observed that the content-based search is much more restrictive. Depending on the query, it may return an empty list, if none of the matched sounds were classified into the specified category. A development version of the web prototype is publicly available [20].

6. CONCLUSIONS

This research aims to improve the search of environmental sounds in large-scale unstructured audio databases. Specifically, we contribute with a content-based analysis and classification framework built upon the ecological acoustics taxonomy proposed by Gaver [3]. To our knowledge, previous approaches on content-based analysis of environmental sounds, have only addressed very concrete sound categories (e.g bird calls, sirens, car engine), without tackling the usage of a general taxonomy.

We proposed a supervised learning approach, and created a annotated database providing several examples of all categories present in the taxonomy. By means of an automatic classifier, the system ranks the sounds according to the acoustic similarity to each class in the taxonomy. We implemented a search interface to *freesound.org* using this system, with promising results. We plan to experiment with other ways to interact with the database using this framework, such as a more exploratory browsing interface.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Universitat Pompeu Fabra. (2005) Freesound.org. Repository of sounds under the Creative Commons license. [Online]. Available: http://www.freesound.org

[2] E. Martínez, O. Celma, M. Sordo, B. De Jong, and X. Serra, “Extending the folksonomies of freesound.org using content-based audio analysis,” in *Sound and Music Computing Conference*, Porto, Portugal, July 2009.

[3] 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.

[4] P. Schaeffer, *Traité des objets musicaux*, 1st ed. Paris, France: Editions du Seuil, 1966.

[5] http://closed.ircam.fr.

[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. Kuo, "Classification and retrieval of sound effects in audiovisual data management," in *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, vol. 1, 1999, pp. 730–734.

[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, pp. 99–111, 2005.

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

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

[12] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame 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] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception & Psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[14] http://www.psy.cmu.edu/~auditorylab/AuditoryLab.html.

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

[16] F. Gouyon and P. Herrera, "Exploration of techniques for automatic labeling of audio drum tracks instruments," in *Proceedings of the Workshop on Current Directions in Computer Music (MOSART)*, Barcelona, Spain, 2001.

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

[18] A. Mathes, "Folksonomies - cooperative classification and communication through shared metadata," December 2004. [Online]. Available: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[19] C. Fellbaum *et al.*, *WordNet: An electronic lexical database*. MIT Press Cambridge, MA, 1998.

[20] http://dev.mtg.upf.edu/soundscape/freesound-search/.