SOUND EFFECTS TAXONOMY MANAGEMENT IN PRODUCTION ENVIRONMENTS

Pedro Cano[1], Markus Koppenberger[1], Oscar Celma[1], Perfecto Herrera[1], and Vadim Tarasov[1]

[1] *Music Technology Group, Institut Universitari de l'Audiovisual, Universitat Pompeu Fabra. Ocata 1, 08003 Barcelona, Spain. http://www.iua.upf.es/mtg*

Correspondence should be addressed to Pedro Cano (pcano@iua.upf.es)

ABSTRACT

Categories or classification schemes offer ways of navigating and higher control over the search and retrieval of audio content. The MPEG-7 standard provides description mechanisms and ontology management tools for multimedia documents. We have implemented a classification scheme for sound effects management inspired on the MPEG-7 standard on top of an existing lexical network, WordNet. WordNet is a semantic network that organizes over 100.000 concepts of the real world with links among them. We show how to extend WordNet with the concepts of the specific domain of sound effects. We review some of the taxonomies to describe acoustically sounds. Mining legacy metadata from sound effects libraries further supplies us with terms. The extended semantic network includes the semantic, perceptual and sound effects specific terms in an unambiguous way. We show the usefulness of the approach easing the task for the librarian and providing higher control on the search and retrieval for the user.

1. INTRODUCTION

Sound engineers create the sound that goes along the image in cinema and video productions, as well as spots and documentaries. Some sounds are recorded for the occasion. Many occasions, however, require the engineer to have access to massive audio libraries. Of the three major facets of audio in post-production: music, speech and sound effects, we focus on sound effects (SFX).

Sound effect management systems rely on classical text descriptors to interact with their audio collections. Librarians tag the sounds with textual description and file them under categories. Users can then search for sounds matching keywords as well as navigating through category trees. Audio filing and logging is a labor-intensive error-prone task. Moreover, languages are imprecise, informal and words have several meanings as well as several words for each meaning; sounds are multimodal, multicultural and multifaceted and there is not an agreement in how to describe them.

Despite the difficulties inherent in creating SFX metadata, there is need to catalog assets so as to reuse afterwards. Media assets have value. As Flank *et al.* [8] point out, there are many situations where reusing media content is, not only not economically appealing—think of the cost of sending a team to record Emperor penguins in their natural habitat—but sometimes audio cannot be re-recorded—like natural catastrophes or historical events [8]. Complete digital media management solutions include media archiving and cataloging, digital right management and collaborative creative environments. This article focuses on the knowledge management aspects of sound effect descriptions with the purpose of making metadata easily searchable, less expensive to create and reusable to support possible new users—including computers—and applications.

MPEG-7 offers a framework for the description of multimedia documents [11]. The description tools for describing a single multimedia document consider semantic, structure and content management

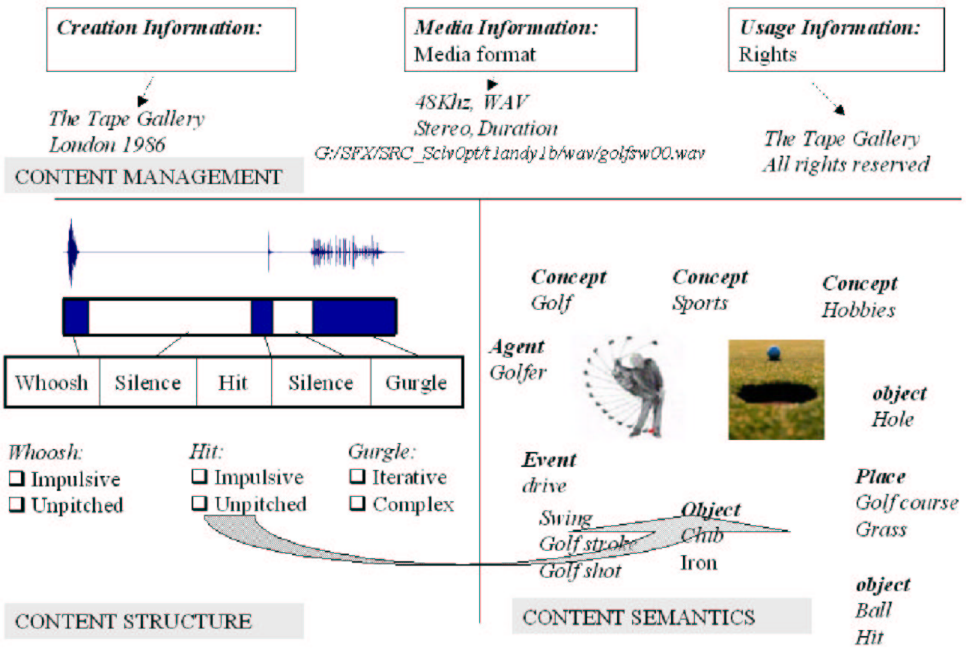

Fig. 1: Example of a SFX description inspired on MPEG-7 Multimedia Description Scheme.

descriptions. MPEG-7 content semantic description tools describe the actions, objects and context of a scene. In sound effects, this correlates to the physical production of the sound in the real world, "1275 cc Mini Cooper Door Closes" , or the context, "Australian Office Atmos Chatter Telephones". MPEG-7 content structure tools concentrate on the spatial, temporal and media source structure of multimedia content. Indeed, important descriptors are those that describe the perceptual qualities independently of the source and how they are structured on a mix. Content management tools are organized in three areas: Media information—which describes storage format, media quality and so on, e.g: "PCM Wav 44100Hz stereo"—, Creation information—which describes the sound generation process, e.g: who and how created the sound—and finally usage information—which describes the copyrights, availability of the content and so on [11]. Figure 1 shows an example on how to describe a SFX inspired on MPEG-7 Multimedia Description Schemes (MDS). The original description is "Golf Swing And Hole" and had been added to the following categories: "Whooshes, Golf, Sports:Golf:Hits:Swings:Swishes".

The use of MPEG-7 description schemes provide a framework suitable for Multimedia description. In order to ensure interoperability and allow the description to be machine readable, the terms within the fields need to be standard. It is important to know whether "bike" refers to "bicycle" or to "motorcycle". MPEG-7 classification schemes allow to define a restrained vocabulary that defines a particular domain as categories with semantic relationships, e.g: Broader term, narrow term, related term and so on. Casey [5] presents an example of using the classification scheme to define a hierarchical sound classification model with 19 leaf nodes. However it is very complicated to devise and maintain taxonomies that account the level of detail needed in a production-size sound effect management system—the categories needed in professional environments exceed the several thousands and they do not follow a hierarchical structure. We have found that it is faster to start developing taxonomies on top on a semantic network such as WordNet rather than starting from scratch. WordNet[1] is an English lexical

[1] http://www.cogsci.princeton.edu/~wn

network designed following psycholinguistic theories of human lexical memory in a long-term collaborative effort [12]. We have developed a WordNet editor to expand it with specific concepts from audio domain, such as "close-up"– which refers to recording conditions—and other specific concepts from the real world, e.g: a Volvo is a type of a car, as well as with the perceptual ontologies. To do so, besides reviewing the existing literature for sound events description, we have mined the concepts associated to sounds by major sound effect library providers and added them to the WordNet. Such a knowledge system is not only useful for retrieval (see Section 3) but it has also been used as ontology backbone for general sounds classification [4].

2. ON SOUND EFFECTS CATALOGING

One of the most time-demanding and error-prone task when building a library of sound effects is the correct labeling and placement of a sound within a category. The information retrieval model commonly used in commercial search engines is based on keyword indexing. Librarians add descriptions for the audio. The systems match the descriptions against the users' query to retrieve the audio. Sounds are difficult to describe with words. Moreover, the librarian must add the text thinking on the different ways a user may eventually look for the sound, e.g: "dinosaur, monster, growl, roar" and at the same time with the maximum detail. We display in Figure 2 some of the fields the librarian could consider when describing a SFX.

The vagueness of the query specification, normally one or two words, together with the ambiguity and informality of natural languages affects the quality of the search: Some relevant sounds are not retrieved and some irrelevant ones are presented to the user. Sound effect management systems also allow browsing for sounds in manually generated categories. It is difficult to manage large category structures. Big corpuses may be labeled by different librarians that follow somewhat different conventions and may not remember under which category sounds should be placed (e.g: Camera:clicks or clicks:camera). Several ways of describing a sound include: source centered description, perceptual, post-production specific and creation description (See Figure 2).

rumbles, roars, explosions, crashes, splashes, booms	Whistles Hisses Puffing
Snorts, Whispers, Murmers, Mumbles, Grumbles, Gurgles	Screeches, Creaks, Rustles, Buzzes, Crackles, Scrapes
Noises make by percussion on: Metal, Wood, Skin, Stone, Pottery, etc.	*Voices of Animals and Men:* Shouts, Screams, Groans, Shrieks, Howls

Table 1: Russolo' Sound-Noise Categories

2.1. Semantic Descriptors

Semantic descriptors usually refer to the source of the sound, that is, what has physically produced the sound, e.g: "car approaching". They also refer to the context, e.g: "Pub atmos". The importance of source-tagging is put in doubt by Mott [10]. Mott explains that the sound engineer should concentrate on the sound independently on what actually produced it because in many occasions the natural sounds do not fulfill the expectations and must be replaced with sounds of distinct origin. There are, however, cases where having the true sound can add quality to a production, e.g: Using the real atmosphere of a Marrakesh market tea house. Besides, describing the source of a sound is sometimes easier than describing the sound itself. It is difficult to describe the "moo of a cow" without mentioning "moo or cow" but just perceptual attributes.

2.2. Perceptual Descriptors

They describe the perceptual qualities independently of the source. Classical research on auditory perception has studied the world of sounds within a multidimensional space with dimensions such as pitch, loudness, duration, timbral brightness, and so on [13]. Since they refer to the properties of sound, sometimes there is a mapping between sound descriptions to perceptual measurable features of the sound.

Another possibility to describe sounds is the use of onomatopoeia, words that imitate sounds and are extensively used in comics (roar, mmm, ring). The futurist painter Russolo [14] proposed in 1913 a categorization of noises in six separate groups: Rumbles, whistles, whispers, screeches, noises obtained

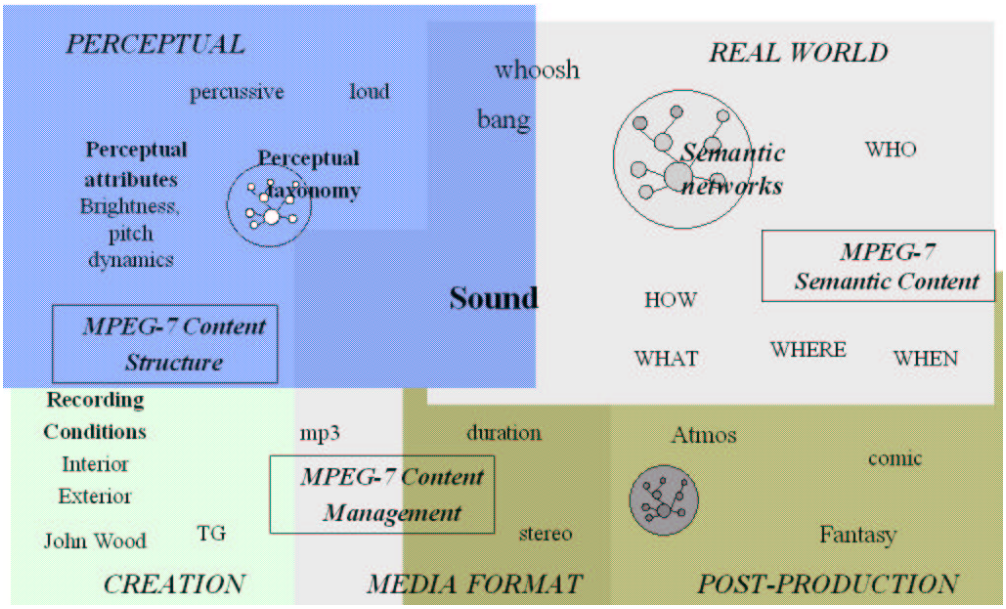

Fig. 2: Block Diagram of the System.

MATTER CRITERIA		
MASS Perception of "noiseness"	HARMONIC TIMBRE Bright/Dull	GRAIN Microstructure of the sound
SHAPE CRITERIA		
DYNAMICS Intensity evolution	ALLURE Amplitude or Frequency Modulation	
VARIATION CRITERIA		
MELODIC PROFILE: pitch variation type	MASS PROFILE Mass variation type	

Table 2: Schaeffer' *Solfege* of sound objects

by percussion and voices of animals and men (see Table 1).

Schaeffer [15], in the search of a lexicon to describe sounds, introduced the reduced listening (*écoute réduite*) which consists in the disposition of the listener to focus on the sound object itself with no reference to the source causing its production. His *solfège* of sound objects (see Table 2) considered attributes such as mass (perception of "pitchiness") or harmonic timbre (bright/dull, round/sharp).

Gaver [9] introduced a taxonomy of environmental sounds on the assertion that sounds are produced by interaction of materials. The hierarchical description of basic sonic events include those produced by vibrating objects (impacts, scraping and others), aerodynamic sounds (explosions, continuous) and liquid sounds (dripping and splashing). The ecological approach to perception distinguishes two types of invariants (i.e.: High-order acoustical properties) in the sound generation: structural and transformational. Structural refer to the objects properties meanwhile transformational refer to the change they undergo [17, 18].

Murray Schafer [16] classifies sounds according to their physical characteristics (acoustics), by the way they are perceived (psychoacoustics), according to their function and meaning (semiotics and semantics); or according to their emotional or affective qualities (aesthetics). Since he is interested in analyzing the sonic environment—soundscape—he adds to Schaeffer sound object description information on the recording settings, e.g: estimated distance from the observer, estimated intensity of the original sound, whether it stands clear out of the background, environmental factors: short reverb, echo.

2.2.1. Post-production Specific Descriptors

Other important searchable metadata are Post-production specific: Natural sounds (actual source sound), characteristic sounds (what a sound should be according to someone), comedy, cartoon, fantasy.

2.2.2. Creation Information

Creation metadata describes relevant information on the creation or recording conditions of the sound. Creation terms that we have found mining SFX descriptions are the library that produced the sound and the engineer that recorded it. Most of the terms we have found refer to the recording conditions of the sound, e.g: to record a "car door closing" one can place the microphone in the interior or in the exterior. Some examples of such descriptors are: interior, exterior, close-up, live recording, programmed sound, studio sound, treated sound. These terms have been added to the taxonomies.

3. ONTOLOGY MANAGEMENT

The use of taxonomies or classification schemes alleviates some of the ambiguity problems inherent to natural languages, yet they pose others. It is very complicated to devise and maintain classification schemes that account for the level of detail needed in a production-size sound effect management system. The MPEG-7 standard provides description mechanisms and ontology management tools for multimedia documents [11]. Celma *et al.* built a flexible search engine for opera works using classification schemes of the MPEG-7 framework [6]. Even though, powerful, the approach would require a huge human effort to extend it for SFX. SFX many times are described referring to the source that produced it. It is not trivial to put terms that describe the world in classification schemes. According to the latest version of WordNet (WordNet 2.0), the number of distinct terms is 152059 and the number of concepts 115424. WordNet is well suited as starting point for ontology-backbone.

Standard dictionaries organize words alphabetically. WordNet organizes concepts in synonym sets, *synsets*, with links between the concepts like: broad sense, narrow sense, part of, made of and so on. It knows for instance that the word piano as a noun has two senses, the musical attribute that refers to "low loudness" and the musical instrument. It also encodes the information that a grand piano is a type of piano, and that it has parts such us a keyboard, a loud pedal and so on. Such a knowledge system is useful for retrieval. It can for instance display the results of a query "car" in types of cars, parts of car, actions of a car (approaching, departing, turning off). The usefulness of using WordNet in Information Retrieval has been proved useful in the case of image retrieval for example in [1] and in general multimedia asset management [7]. Even though WordNet already organizes plenty of terms, it sometimes lacks specific knowledge of the specific domain of sound effects and audio in general, such as "close-up" when referring to the recording technique or that a "747" is an airplane (this property is known as lack of "granularity"). We have developed an editor to expand WordNet with terms and concepts from perceptual taxonomies to describe sounds and mining legacy metadata from sound effects libraries. We have based the editor on the work by Brian [3]. The addition of a new term, for instance Jaguar—referring to the car—would go as follows: We would first search for the term in WordNet to make sure it does not exist. There is on concept associated with jaguar:

```
1. jaguar, panther, Panthera onca, Felis onca --
   a large spotted feline of tropical America
   similar to the leopard;
```

Since this synset does not apply, the second step is finding a node in the taxonomy to which attach the new concept, in our case if ask the synsets for "car":

```
1. car, auto, automobile, machine, motorcar --
   4-wheeled motor vehicle; usually propelled
   by an internal combustion engine;
2. car, railcar, railway car, railroad car --
   a wheeled vehicle adapted to the rails of
   railroad; "three cars had jumped the rails"
3. cable car, car -- conveyance for passengers
   or freight on a cable railway;
4. car, gondola -- car suspended from an airship
   and carrying personnel and cargo and power
   plant)
5. car, elevator car --where passengers ride up
   and down;
```

In this case we would choose connecting Jaguar with the first synset: "car, auto, ...". The type of relationship between Jaguar and car is of hyponim/hypernim (Narrower/Broader). If we ask for the list of hypernims of the "jaguar" we get:

```
1. jaguar, panther, Panthera onca, Felis onca --
   a large spotted feline of tropical America
   similar to the leopard;
   => big cat, cat
     => feline, felid
       => carnivore
         => mammal
           ...
              => animal, animate being, beast
                => organism, being
                  => living thing, animate thing
                    => object, physical object
                      => entity, physical thing

2. Jaguar -- British car ...
 => car, auto, automobile, machine, motorcar
   => motor vehicle, automotive vehicle
     => wheeled vehicle
       => vehicle
         => conveyance, transport
           => instrumentality, instrumentation
             => artifact, artefact
               => object, physical object
                 => entity, physical thing
```

A hyponim acquires all the features of its hypernim. Therefore if we ask the meronyms—relation of being part of—of Jaguar in its second new meaning we get:

```
Sense 2
Jaguar
 HAS PART: accelerator, gas pedal, gas, throttle
 HAS PART: air bag
 HAS PART: automobile engine
 HAS PART: car horn, motor horn, horn, hooter
 HAS PART: bumper
 HAS PART: car door
 ...
```

After the disambiguation of the terms used to label a database of 60857 sounds from over 30 libraries of sound effects, music and music samples, we have 3028 different concepts. The histogram of number of synsets assigned per sound sample is depicted in Figure 3. The higher the number of synsets, the more detailed is the description of the sound. Table 3 shows the most commonly used concepts. The first column indicates the number of sounds that have been labeled with the synset, the second column, the offset (WordNet Synset-ID) and the third the glossary. The distribution of 3028 synsets with respect its syntactic function is as follows: 2381 nouns, 380 verbs, 251 adjectives and 16 adverbs (see Figure 4). The following are examples of disambiguation of captions into synsets:

```
Dalmatian Dog Bark Interior ->
   01778031%n dalmatian,  ...
   01752990%n dog, domestic dog, ...
   00826603%v bark -- make barking sounds
   00915868%a interior -- (situated ...

Cello pizzicato ->
   02605020%n cello, violoncello
         => bowed stringed instrument, string
           => stringed instrument
               => musical instrument, instrument

   00908432%a pizzicato -- ((of instruments in
      the violin family) to be plucked with the
      finger)
   00422634%r pizzicato -- ((music) with a light
      plucking staccato sound)
```

The extended semantic network includes the semantic, perceptual and sound effects specific terms in an unambiguous way, easing the task for the librarian and providing higher control on the search and retrieval for the user. Further work needs to deal with concepts that appear on different parts-of-speech—pizzicato is both an adjective and an adverb—but are equivalent for retrieval purposes.

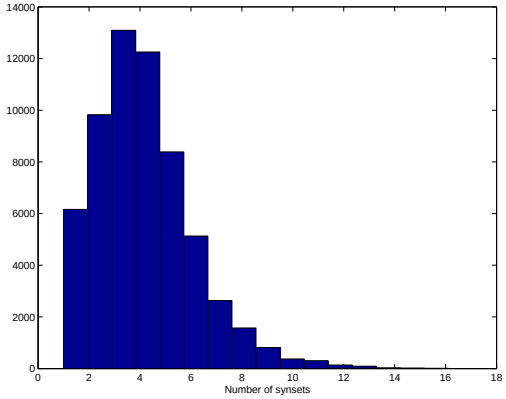

Fig. 3: Histogram of the number of concepts assigned to each SFX. The higher the number of concepts the most detailed the specification of the SFX.

# of Sounds	Synset	Terms and Glossary
5653	03131431%n	drum, membranophone, tympan – (a musical percussion instrument; usually consists of a hollow cylinder with a membrane stretch across each end)
4799	13697183%n	atmosphere, ambiance, ambience – (a particular environment or surrounding influence; "there was an atmosphere of excitement")
4009	06651357%n	rhythm, beat, musical rhythm – (the basic rhythmic unit in a piece of music; "the piece has a fast rhythm"; "the conductor set the beat")
3784	07719788%n	percussion section, percussion, rhythm section – (the section of a band or orchestra that plays percussion instruments)
3619	14421098%n	beats per minute, bpm, metronome marking, M.M.
3168	00006026%n	person, individual, someone, somebody, mortal, human, soul

Table 3: Appearance number of most popular concepts (synsets).

4. BENEFITS OF WORDNET-BASED TAXONOMY MANAGEMENT

The use of a WordNet-based taxonomy management together with Natural Language Processing tools enhances text-search engines used in sound effects retrieval systems by moving from keyword to concept-based search. At the same time it eases the librarian task when describing sounds and it simplifies the management of the categories.

- Higher control on the precision and recall of the results using WordNet concepts. The query "bike" returns both "bicycle" and "motorcycle" sounds and the user is given the option to refine the search.
- Common sense "intelligent" navigation: The concept relations encoded in WordNet can be used to propose related terms. It is generally accepted that recognition is stronger than recall and a user may not know how the librarian tagged a sound.
- Proposal of higher level related term not included in the lexical network. WordNet does not have all possible relations. For instance, "footsteps in mud", "tractor", "cow bells" and "hens" may seem related in our minds when we think of farm sounds but do not have direct links within WordNet. It is possible to recover this type of relations because there are many sounds that have been labeled with the concept "farm". Studying the co-occurrence of synsets allows the system to infer related terms [2].
- There is a lemmatizer, say "bikes" becomes "bike", an inflecter that allows to expand it to "bike, bikes and biking", and a name entity recognition module, that is able to identify "Grand piano" as a specific type of piano both based on WordNet.
- Module for the phonetic matching, e.g: "whoooassh" retrieves "whoosh". Phonetic matching is used in information retrieval to account for the typo errors in a query and thus aims at reducing the frustration of a user. In sound effects retrieval, it is even more important since it is common practice to describe sounds as they sound if one reads them. WordNet has a very complete onomatopoeia ontology.

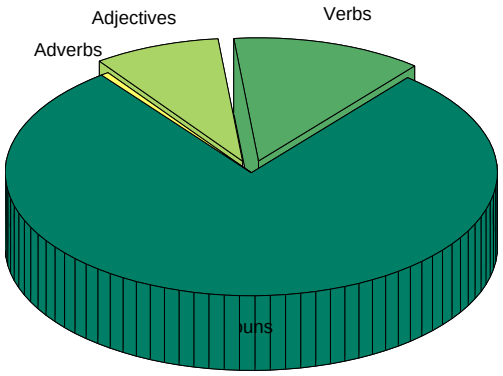

Fig. 4: Distribution of nouns, verbs, adjectives and adverbs after disambiguating a SFX collection.

5. SUMMARY

We have presented some of the problems in cataloging SFX. Specifically, how to generate searchable and machine readable SFX description metadata. We have reviewed some of the literature for audio classification as well as mined legacy SFX metadata. We have implemented a knowledge management system inspired on the MPEG-7 framework for Multimedia and relying on WordNet as taxonomy-backbone. The librarian, does not need to add many terms since many relations are given by the lexicon. Categories can be created dynamically allowing user can search and navigate through taxonomies based on psycolinguistic and cognitive theories. The terms—even though described externally as plain English—are machine readable, unambiguous and can be used for concept-based retrieval. Specific SFX terms as well as external taxonomies can be added to the lexicon.

6. ACKNOWLEDGMENTS

We thank the staff from the Tape Gallery for all the support, discussion and feedback. This work is partially funded by the AUDIOCLAS Project E! 2668 Eureka. We thank the collaboration from Sylvain Le Groux, Julien Ricard and Nicolas Wack. We thank the review and feedback from Alvaro Barbosa, Eloi Batlle, Fabien Gouyon and José Lozano.

7. REFERENCES

[1] Y. Alp Aslandogan, C. Thier, C. T. Yu, and J. Zou nd N. Rishe. Using semantic contents and WordNet in image retrieval. In *Proc. of the SIGIR*, Philadelphia, PA, 1997.

[2] S. Banerjee and T. Pedersen. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2003.

[3] D. Brian. Lingua:WordNet. *The Perl Journal*, 5(2):12–17, Summer 2002.

[4] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, P. Herrera, and N. Wack. Nearest-neighbor generic sound classification with a wordnet-based taxonomy. In *Proc.116th AES Convention*, Berlin, Germany, 2004.

[5] M. Casey. Generalized sound classification and similarity in MPEG-7. *Organized Sound*, 6(2), 2002.

[6] O. Celma and E. Mieza. An opera information system based on MPEG-7. In *Proc. AES 25th Int. Conf.*, London, UK, 2004.

[7] S. Flank. Multimedia technology in context. *IEEE Multimedia*, pages 12–17, July-September 2002.

[8] S. Flank and S. Brinkman. Drinking from the fire hose: How to manage all the metadata. In *Proceedings of the International Broadcasting Convention*, Sep 2002.

[9] W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[10] R. L.Mott. *Sound Effects: Radio, TV, and Film*. Focal Press, 1990.

[11] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7. Multimedia Content Description Interface*. John Wiley & Sons, LTD, 2002.

[12] G. A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, pages 39–45, November 1995.

[13] SOb Project. *The Sounding Object*. Mondo Estremo, 2003.

[14] L. Russolo. *The Art of Noises*. Pendragon Press, 1986.

[15] P. Schaeffer. *Traité des Objets Musicaux*. Editions du Seuil, 1966.

[16] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Alfred Knopf, Inc., 1977.

[17] J.J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin, 1979.

[18] W.H. Warren and R.R. Verbrugge. Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *J. of Exp. Psych.*, 10:704–712, 1984.