KNOWLEDGE AND CONTENT-BASED AUDIO RETRIEVAL USING WORDNET

Pedro Cano and Markus Koppenberger and Sylvain Le Groux
and Perfecto Herrera and Julien Ricard and Nicolas Wack
Institut de l'Audiovisual, Universitat Pompeu Fabra
c/Ocata 3, 08003 Barcelona, Spain

Keywords: MPEG7, content-based audio, sound effects, ontology management, WordNet, audio classification, audio asset management

Abstract: Sound producers 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 libraries of music and sound effects. Of the three major facets of audio in post-production: music, speech and sound effects, this document focuses on sound effects (Sound FX or SFX). Main professional on-line sound-fx providers offer their collections using standard text-retrieval technologies. Library construction is an error-prone and labor consuming task. Moreover, the ambiguity and informality of natural languages affects the quality of the search. The use of ontologies alleviates some of the ambiguity problems inherent to natural languages, yet it is very complicated to devise and maintain an ontology that account for the level of detail needed in a production-size sound effect management system. To address this problem we use WordNet, an ontology that organizes over 100.000 concepts of real world knowledge: e.g: it relates doors to locks, to wood and to the actions of opening, closing or knocking. However a fundamental issue remains: sounds without caption are invisible to the users. Content-based audio tools offer perceptual ways of navigating the audio collections, like "find similar sound", even if unlabeled, or query-by-example, possibly restricting the search to a semantic subspace, such as "vehicles". The proposed content-based technologies also allow semi-automatic sound annotation. We describe the integration of semantically-enhanced management of metadata using WordNet together with content-based methods in a commercial sound effect management system.

1 INTRODUCTION

The audio component is a fundamental aspect in an audiovisual production. Around 75% of the sound effects (SFX) of a movie are added at post-production. Many sounds are useless due to the noise in the recording session and some are simply not picked up by the production microphones. Sometimes sounds are replaced in order to improve the dramatic impact, e.g.: arrow sounds of the "Lord of the Rings" are replaced by "whooshes". There are also artistic reasons, for example, in the movie "All the President's Men", in order to strengthen the message that the pen is mightier that the sword, the typewriter keys sounds were mixed with the sound of gunfire(Weis, 1995).

Many occasions, not only movies but also computer games, audio-visual presentations, web-sites require sounds. These sounds can be recorded as well as recreated using foley techniques—for the sound of the knife entering the body in Psycho' shower scene, Hitchcock used a melon (Weis, 1995). Another possibility is the use of already compiled SFX libraries. Accessing library sounds can be an interesting alternative to sending a team to record sounds or to recreate them in a studio because one needs then a foley pit and the rest of the recording equipment.

A number of SFX providers, for example: www.sounddogs.com, www.sonomic.com or www.sound-effects-library.com, offer SFX on-line. The technology behind these services is standard text-search. Librarians tag sounds with descriptive keywords that the users may search. Some companies also keep categories—such as "automobiles", "horror" or "crashes"—to ease the interaction with the collections. This approach presents several limitations. The work of the librarian is error-prone and a very time-consuming task. Solutions have been proposed to manage media assets from a audio

Cano P., Koppenberger M., Le Groux S., Herrera P., Ricard J. and Wack N. (2004).
KNOWLEDGE AND CONTENT-BASED AUDIO RETRIEVAL USING WORDNET.
In *Proceedings of the First International Conference on E-Business and Telecommunication Networks*, pages 301-308
DOI: 10.5220/0001397503010308

content-based audio perspective, both from the academia and the industry (see Section 2). However none seems to have impacted in professional sound effects management systems. Another source of problems is due to the imprecision and ambiguity of natural languages. Natural languages present polysemy—“bike” can mean both “bicycle” and “motorcycle”—and synonymy—both “elevator” and “lift” refer to the same concept. This, together with the difficulty associated to describing sounds with words, affects the quality of the search. The user has to guess how the librarian has labeled the sounds and either too many or too few results are returned.

In this context we present a SFX retrieval system that incorporates content-based audio techniques and semantic knowledge tools implemented on top of one of the biggest sound effects providers database. The rest of the paper is organized as follows: in Section 2 we review what what existing literature proposes to improve sound effect management. From Section 3 to 5 we describe the implemented enhancements of the system.

2 RELATED WORK

Related work to sound effect management falls into three categories: Content-based audio technologies, approaches to describe sound events and taxonomy management.

2.1 Content-based audio classification and retrieval

Content-based functionalities aim at finding new ways of querying and browsing audio documents as well as automatic generating of metadata, mainly via classification. Query-by-example and similarity measures that allow perceptual browsing of an audio collection is addressed in the literature and exist in commercial products, see for instance: www.findsounds.com, www.soundfisher.com.

Existing classification methods normally concentrate on small domains, such as musical instrument classification or very simplified sound effects taxonomies. Classification methods cannot currently offer the detail needed in commercial sound effects management, e.g: “female steps on wood, fast”. In audio classification, researchers normally assume the existence or define a well defined hierarchical classification scheme of a few categories (less than a hundred at the leaves of the tree). On-line sound effects and music sample providers have several thousand categories. For further discussion on classification of general sound, we refer to (Cano et al., 2004a).

2.2 Description of Audio

Sounds are multifaceted, multirepresentional and usually difficult to describe in words. MPEG-7 offers a framework for the description of multimedia documents, see (Manjunath et al., 2002). 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, “moo cow solo”, or the context, “Airport atmos announcer”.

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. Since they refer to the properties of sound, e.g: Loudness, brightness. Other important searchable metadata are post-production specific descriptions, e.g.: horror, comic or science-fiction. Creation metadata describe how the sound was recorded. For example, 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. For a more complete review on SFX description, we refer to (Cano et al., 2004b).

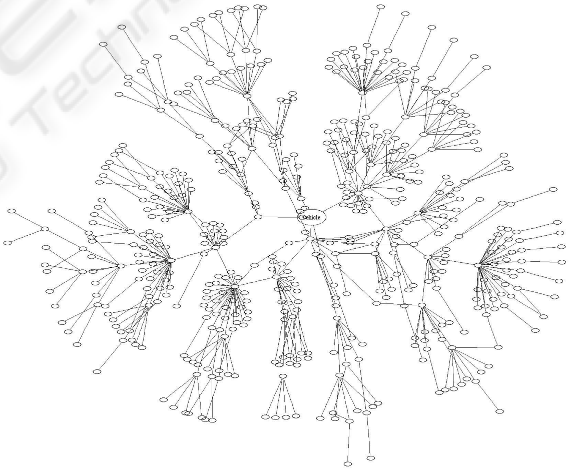


Figure 1: Snapshot of the vehicle taxonomy in WordNet 2.0. Only the hypernym type of relation is displayed.

2.3 Taxonomy 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 (Manjunath et al., 2002). (Celma and Mieza, 2004) show an opera information system that exploits taxonomies built with classification schemes using MPEG-7 framework. However, it is very complicated to extend them to the level of detail needed in a production-size sound effect management system. We have found that it is much faster to start developing ontologies on top on a semantic network such as WordNet rather than starting from scratch.

WordNet (Miller, 1995)/(http://www.cogsci.princeton.edu/ wn/) is a lexical network designed following psycholinguistic theories of human lexical memory. Standard dictionaries organize words alphabetically. WordNet organizes concepts in synonym sets, *synsets*, with links between the concepts like: broader sense, narrower sense, part of, made of and so on. It knows for instance that the word "piano" 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). Figure 1 displays a subset of the "vehicle" graph from WordNet 2.0. The usefulness of such knowledge systems has been justified for image retrieval in (Aslandogan et al., 1997) and in general multimedia asset management (Flank, 2002).

3 SYSTEM OVERVIEW

Text-based and content-based methods alone do not seem to suffice for a complete interaction with vast sound effects repositories. In the implemented system we aim to combine the best of two worlds to offer tools for the users to refine and explore a huge collection of audio. Similar work on integrating perceptual and semantic information in a more general multimedia framework is MediaNet (Benitez et al., 2000). The system we present is specialized for SFX. The current prototype uses 80.000 sounds from a major on-line sound effects provider: www.sound-effects-library.com. Sounds come with a textual description which has been disambiguated with the augmented WordNet ontology.

3.1 Functional blocks

The system has been designed to ease the use of different tools to interact with the audio collection and with speed as a major design issue. On top of these premises we have implemented the blocks of Fig. 2. The sound analysis, audio retrieval and metadata generation blocks are described in Section 4. The text retrieval, text processor and knowledge manager blocks are described in Section 5.

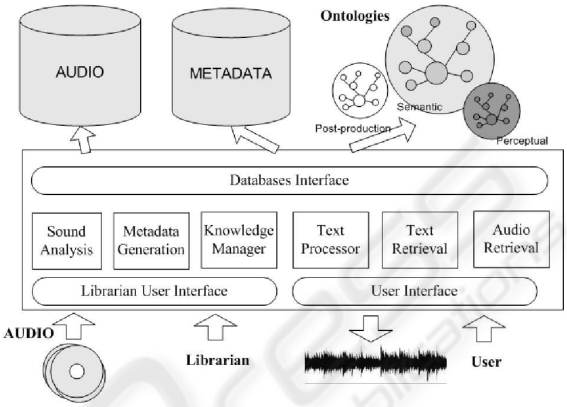


Figure 2: Functional Block Diagram of the System.

3.2 System architecture

The audio processing engines use the CLAM Framework, a C++ Library for Audio and Music, developed at the MTG and distributed under GNU/GPL License (http://www.iua.upf.es/mtg/clam). The ontology management and integration of different parts is done with Perl and a standard relational database management system. The functionality is available via a web interface and exported via SOAP (http://www.w3.org/TR/soap). The SOAP interface provides some exclusive functionality—such as interaction with special applications, e.g.: Sound editors and annotators— which is not available via the web interface. See Figure 3 for a diagram of the architecture.

4 CONTENT-BASED AUDIO TOOLS

Content-based audio tools ease the work of the librarian and enhance the search possibilities for the user. It simplifies the labeling of new sounds because many keywords are automatically presented to the librarian. To achieve it, the new sound is compared to the collection with Nearest Neighbor search and the text associated with the similar matches is presented to the librarian. The sound analysis module (see Figure 2), besides extracting sound descriptors used for the similarity search, generates searchable descriptors

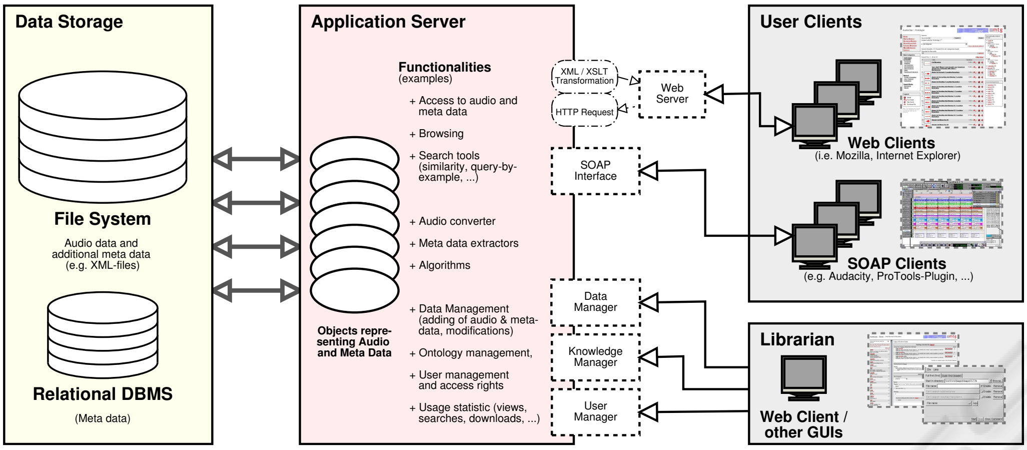

Data Storage
File System
Audio data and additional meta data (e.g. XML-files)
Relational DBMS
(Meta data)
Application Server
Functionalities (examples)
+ Access to audio and meta data
+ Browsing
+ Search tools (similarity, query-by-example, ...)
+ Audio converter
+ Meta data extractors
+ Algorithms
+ Data Management (adding of audio & meta-data, modifications)
+ Ontology management,
+ User management and access rights
+ Usage statistic (views, searches, downloads, ...)
Objects representing Audio and Meta Data
XML / XSLT Transformation
HTTP Request
Web Server
SOAP Interface
Data Manager
Knowledge Manager
User Manager
User Clients
Web Clients
(i.e. Mozilla, Internet Explorer)
SOAP Clients
(e.g. Audacity, ProTools-Plugin, ...)
Librarian
Web Client / other GUIs

Figure 3: System Architecture

as those detailed in Subsection 4.2 (crescendo, noisy, etc.). Content-based tools offer the user functionalities such as:

Virtual Foley Mode: Find perceptually similar sounds. A user may be interested in a glass crash sound. If none of the retrieved sounds suits him, he can still browse the collection for similar sounds even if produced by different sources, even if unlabeled.

Clustering of sounds: Typically a query like “whoosh” may retrieve several hundred results. These results are clustered and only one representative of each class is displayed to the user. The user can then refine the search more easily.

Morphological Descriptors: Another option when the list of results is too large to listen to is filtering the results using morphological descriptors (see Section 4.2).

Query by example: The user can provide an example sound or utter himself one as a query to the system, possibly restricting the search to a semantic subspace, such as “mammals”.

4.1 Similarity Distance

The similarity measure is a normalized Manhattan distance of features belonging to three different groups: a first group gathering spectral as well as temporal descriptors included in the MPEG-7 standard; a second one built on Bark Bands perceptual division of the acoustic spectrum and which outputs the mean and variance of relative energies for each band; and, finally a third one, composed of Mel-Frequency Cepstral Coefficients and their corresponding variances (see (Cano et al., 2004a) for details):

$$d\left(x,y\right) = \sum_{k=1}^{N} \frac{|x_k - y_k|}{(max_k - min_k)}$$

where x and y are the vectors of features, N the dimensionality of the feature space, and max_k and min_k the maximum and minimum values of the kth feature.

The similarity measure is used for metadata generation: a sound sample will be labeled with the descriptions from the similar sounding examples of the annotated database. This type of classification is known as one-nearest neighbor decision rule (1-NN)(Jain et al., 2000). The choice of a memory-based nearest neighbor classifier avoids the design and training of every possible class of sound which is of the order of several thousands. Besides, it does not need redesign or training whenever a new class of sounds is added to the system. The NN classifier needs a database of labeled instances and a similarity distance to compare them. An unknown sample will borrow the metadata associated with the most similar registered sample.

The similarity measure is also used for the query-by-example and to browse through “perceptually” generated hyperlinks.

4.2 Morphological Sound Description

The morphological sounds descriptor module extracts a set of descriptors that focused on intrinsic perceptual qualities of sound based on Schaeffer’s research on *sound objects* (Schaeffer, 1966). The extractor of morphological descriptors (Ricard and Herrera, 2004) currently generates the following metadata:

Pitchness: (Organization within the spectral dimensions) Pitch, Complex and Noisy.

Dynamic Profile: (Intensity description) Unvarying, Crescendo, Decrescendo, Delta, Impulsive, Iterative, Other.

Pitchness Profile: (Temporal evolution of the internal spectral components) Varying and Unvarying

Pitch Profile: (Temporal evolution of the global spectrum) Undefined, Unvarying, Varying Continuous, Varying Stepped.

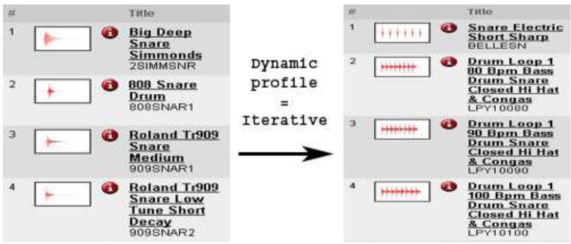

#
Title
1
Big Deep Snare Simmonds
2SIMMSNR
2
808 Snare Drum
808SNAR1
3
Roland Tr909 Snare Medium
909SNAR1
4
Roland Tr909 Snare Low Tune Short Decay
909SNAR2
Dynamic profile = Iterative
#
Title
1
Snare Electric Short Sharp
BELLESN
2
Drum Loop 1 80 Bpm Bass Drum Snare Closed Hi Hat & Congas
LPY10080
3
Drum Loop 1 90 Bpm Bass Drum Snare Closed Hi Hat & Congas
LPY10090
4
Drum Loop 1 100 Bpm Bass Drum Snare Closed Hi Hat & Congas
LPY10100

Figure 4: Morphological descriptor filtering. The iterative dynamic profile allows to discriminate between snare samples and loops

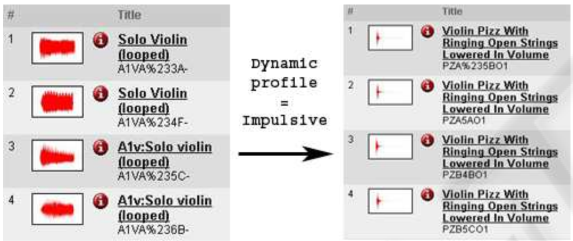

#
Title
1
Solo Violin (looped)
A1VA%233A-
2
Solo Violin (looped)
A1VA%234F-
3
A1v:Solo violin (looped)
A1VA%235C-
4
A1v:Solo violin (looped)
A1VA%236B-
Dynamic profile = Impulsive
#
Title
1
Violin Pizz With Ringing Open Strings Lowered In Volume
PZA%235BO1
2
Violin Pizz With Ringing Open Strings Lowered In Volume
PZA5AO1
3
Violin Pizz With Ringing Open Strings Lowered In Volume
PZB4BO1
4
Violin Pizz With Ringing Open Strings Lowered In Volume
PZB5CO1

Figure 5: Morphological description filtering. The impulsive dynamic profile allows to discriminate violin pizzicati.

These descriptors can be used to retrieve abstract sounds as well as refine other types of searches. Besides applying to all types of sounds, the use of an automatic extractor avoids expensive human labeling while it assures consistency. For details on the construction and usability evaluation of the morphological sound description we refer to (Ricard and Herrera, 2004).

4.3 Clustering and Visualization Tools

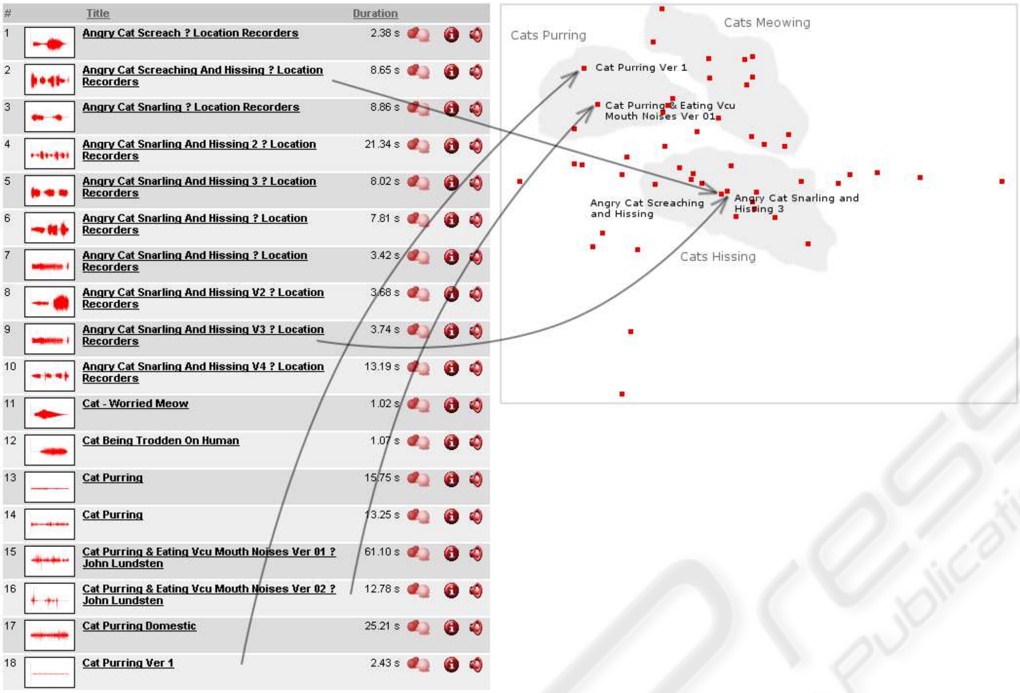
Usually, systems for content-based retrieval of similar sounds output a list of similar sounds ordered by increasing similarity distance. The list of retrieved sounds can rapidly grow and the search of the appropriate sound becomes tedious. There is a need for a user-friendly type of interface for browsing through similar sounds. One possibility for avoiding having to go over, say 400 gunshots, is via clustering sounds into perceptually meaningful subsets, so that the user can choose what perceptual category of sound he or she wishes to explore. We used a hierarchical tree clustering with average linkage algorithm and the above mentioned similarity distance (Jain et al., 2000). Another possibility of interaction with the sounds is using visualization techniques, specifically Multidimensional scaling (MDS), self-organizing maps (SOM) or FastMap (Cano et al., 2002), to map the audio samples into points of an Euclidean space. Figure 6 displays a mapping of the audio samples to a 2D space. In the example it is possible to distinguish different classes of cat sounds, e.g.: “purring”, “hissing” and “miaow” sounds.

5 NATURAL LANGUAGE PROCESSING AND KNOWLEDGE MANAGER

This module enhances existing text-search engines used in sound effects retrieval systems. It eases the librarian work and it simplifies the management of the categories.

- There is a lemmatizer, say “bikes” becomes “bike”, an inflecter that allows to expand it to “bike, bikes and biking”, and a named entity recognition module, that is able to identify “Grand piano” as a specific type of piano.
- 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.
- 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.
- Proposal of related terms. It is generally accepted that recognition is stronger than recall. A user may not know how the librarian tagged a sound. WordNet can be used to propose alternative search terms.
- 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 (Banerjee and Pedersen, 2003).

For further details on the implementation and evaluation of WordNet as backbone for sound effect ontology management, we refer to (Cano et al., 2004b).



#		Title	Duration
1		Angry Cat Screach ? Location Recorders	2.38 s
2		Angry Cat Screaching And Hissing ? Location Recorders	8.65 s
3		Angry Cat Snarling ? Location Recorders	8.86 s
4		Angry Cat Snarling And Hissing 2 ? Location Recorders	21.34 s
5		Angry Cat Snarling And Hissing 3 ? Location Recorders	8.02 s
6		Angry Cat Snarling And Hissing ? Location Recorders	7.81 s
7		Angry Cat Snarling And Hissing ? Location Recorders	3.42 s
8		Angry Cat Snarling And Hissing V2 ? Location Recorders	3.68 s
9		Angry Cat Snarling And Hissing V3 ? Location Recorders	3.74 s
10		Angry Cat Snarling And Hissing V4 ? Location Recorders	13.19 s
11		Cat - Worried Meow	1.02 s
12		Cat Being Trodden On Human	1.07 s
13		Cat Purring	15.75 s
14		Cat Purring	13.25 s
15		Cat Purring & Eating Vcu Mouth Noises Ver 01 ? John Lundsten	61.10 s
16		Cat Purring & Eating Vcu Mouth Noises Ver 02 ? John Lundsten	12.78 s
17		Cat Purring Domestic	25.21 s
18		Cat Purring Ver 1	2.43 s

Cats Purring
Cats Meowing
Cat Purring Ver 1
Cat Purring & Eating Vcu Mouth Noises Ver 01
Angry Cat Screaching and Hissing
Angry Cat Snarling and Hissing 3
Cats Hissing

Figure 6: FastMap visualization screenshot. The points of the 2D map refer to different audio samples. The distances on the euclidean space try to preserve distances in the hyper-dimensional perceptual space defined by the similarity distance of subsection 4.1

6 EXPERIMENTS

We have used 40.000 sounds from the Sound-Effects-Library (http://www.sound-effects-library.com) for the experiments. These sounds have been unambiguously tagged with concepts of an enhanced WordNet. Thus a violin sound with the following caption:"Concert Grand Piano - piano" may have the following synsets (the numbers on the left are the unique WordNet synset identifiers):

- 02974665%n concert grand, concert piano – (a grand piano suitable for concert performances)
- 04729552%n piano, pianissimo – ((music) low loudness)

6.1 Experimental setup

The evaluation of similarity distances is a tricky subject. Perceptual listening tests are expensive. Another possibility is to evaluate the goodness of the similarity measure examining the performance in a Nearest Neighbor (NN) classification task. As we will see in 6.3, the overlap between semantic and perceptual taxonomies complicates the evaluation. In musical instruments, the semantic taxonomy more or less follows an acoustic classification scheme, basically due to the physical construction, and so instruments are wind (wood and brass), string (plucked or bowed) and so on (Herrera et al., 2003).

We have tried three ways of assessing the perceptual similarity between sounds:

- Perceptual listening experiments
- Classification or metadata generation performance
- Consistency on the ranking and robustness to distortions such as resampling, transcoding (converting to MP3 format at different compression rates and back). The harmonic instruments have been transcoded and resampled into WAV PCM format and Ogg format (www.vorbis.com). The classification accuracy dropped from 92% to 71.5% in the worst case.

6.2 Perceptual listening tests

In order to test the relevance of our similarity measures, we asked users of our system to give a personal perceptual evaluation on the retrieval of sounds by similarity. This experiment was accomplished on 20 users who chose 41 different queries, and produced 568 evaluations on the relevance of the similar sound retrieved. During the evaluation, the users were presented with a grading scale from 1—not similar at all—to 5—closely similar. The average grade

Table 1: The classifier assigns the metadata of the sounds of the second column to the sounds of the first.

Query Sound Caption	Nearest-neighbor Caption
1275cc Mini Cooper Door Closes Interior Perspective	Trabant Car Door Close
Waterfall Medium Constant	Extremely Heavy Rain Storm Short Loop
M-domestic Cat- Harsh Meow	A1v:Solo violin (looped)
Auto Pull Up Shut Off Oldsmobile Cutlass	Ferrari - Hard Take Off Away - Fast

was 2.6. We have at our disposal the semantic concepts associated with the 40.000 sounds used in the experiment. It turned out that the semantic class of a sound is crucial in the user's sensation of similarity. The conclusion of our experiment is that the users gave better grades to retrieved sounds that are from the same semantic class as the query sound (40% of the best graded sounds belonged to the same semantic class). In the prototype, in addition to the purely content-based retrieval tools, the use of the knowledge-based tools allows searches for similar sounds inside a specific semantic family.

6.3 Metadata annotation performance

The first experiment on metadata annotation performance consisted in finding a best-match for all the sounds in the database. Table 1 shows some examples: on the left column the original caption of the query sound and on the right the caption of the nearest neighbor. The caption on the right would be assigned to the query sound in an automatic annotation system.

As can be inferred from table 1 it is difficult to quantitatively evaluate the performance of the system. An intersection on the terms of the captions would not yield a reasonable evaluation metric. The WordNet based ontology can inform us that both “Trabant” and “Mini Cooper” are narrow terms for the concept“car, automobile”. Thus, the comparison of the number of common synsets on both query and nearest-neighbor could be used as a better evaluation. The number of concepts (synsets) that the sound in the database and their best match have in common was bigger than one—at least one synset—half of the time. Yet there are many cases when this metric is not appropriate for similarity evaluation. The intersection of source descriptions can be zero for very similar sounding sounds. The closest-match for a “paper bag” turns out to be a “eating toast”. These sounds are semantically different but perceptually equivalent. The ambiguity is a disadvantage when designing and assessing perceptual similarity distances.

In a second experiment we have tested the general approach in reduced domain classification regime mode: percussive instruments, harmonic instruments and we achieve acceptable performances. The assumption is that there is a parallelism between semantic and perceptual taxonomies in musical instruments. The psychoacoustic studies of (Lakatos, 2000) revealed groupings based on the similarities in the physical structure of instruments. We have therefore evaluated the similarity with classification on the musical instruments space, a subspace of the universe of sounds.

In the 6 class percussive instrument classification we achieve a 85% recognition (955 audio files) using 10 fold validation. The results for a 8 class classification of harmonic instruments is a 77.3% (261 audio files). We refer to (Cano et al., 2004a) for further discussion on general sound similarity and classification.

7 SUMMARY

We have introduced the difficulties inherent in interacting with sound effect repositories, both for the librarian who designs such content repositories and for potential users who access this content. We have presented several technologies that enhance and fit smoothly into professional sound effects providers working processes. Several content-based audio tools have been integrated providing possibilities of accessing sounds which are unrelated from the text caption but sound the same—even if they are unlabeled. Several natural language processing tools have also been described. WordNet, previously proposed for multimedia retrieval has been extended for sound effects retrieval.

The system can be accessed and evaluated at http://www.audioclas.org.

ACKNOWLEDGEMENTS

We thank the staff from the Tape Gallery for all the support, discussion and feedback. This work is developed under a EU E! 2668 Eureka AudioClas. We thank Oscar Celma and Fabien Gouyon for the review.

REFERENCES

Aslandogan, Y. A., Thier, C., Yu, C. T., and nd N. Rishe, J. Z. (1997). Using semantic contents and WordNet in image retrieval. In *Proc. of the SIGIR*, Philadelphia, PA.

Banerjee, S. and Pedersen, T. (2003). 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.

Benitez, A. B., Smith, J. R., and Chang, S.-F. (2000). Medianet: A multimedia information network for knowledge representation. In *Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems*, volume 4210.

Cano, P., Kaltenbrunner, M., Gouyon, F., and Batlle, E. (2002). On the use of FastMap for audio information retrieval. In *Proceedings of the International Symposium on Music Information Retrieval*, Paris, France.

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

Cano, P., Koppenberger, M., Herrera, P., and Celma, O. (2004b). Sound effects taxonomy management in production environments. In *Proc. AES 25th Int. Conf.*, London, UK.

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

Flank, S. (July-September 2002). Multimedia technology in context. *IEEE Multimedia*, pages 12–17.

Herrera, P., Peeters, G., and Dubnov, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1).

Jain, A. K., Duin, R. P., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychoacoustics*, (62):1426–1439.

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

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

Ricard, J. and Herrera, P. (2004). Morphological sound description: Computational model and usability evaluation. In *Proc.116th AES Convention*, Berlin, Germany.

Schaeffer, P. (1966). *Trait des Objets Musicaux*. Editions du Seuil.

Weis, E. (1995). Sync tanks: The art and technique of post-production sound. *Cineaste*, 21(1):56.