SEMANTIC AND PERCEPTUAL MANAGEMENT

OF SOUND EFFECTS IN PRODUCTION SYSTEMS

P. Cano, M. Koppenberger, S. Le Groux, J. Ricard and N. Wack

Universitat Pompeu Fabra, Spain

ABSTRACT

Main professional sound effects (SFX) providers offer their collections using standard text-retrieval technologies. SFX cataloging is an error-prone and labor consuming task. 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. The use of ontologies alleviates some of the ambiguity problems inherent to natural languages, yet they pose others. 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 real world knowledge: e.g.: it relates doors to locks, to wood and to the actions of 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 sounds”, even if unlabeled, or query-by-example. We describe the integration of semantically-enhanced management of metadata using WordNet together with content-based methods in a commercial sound effect management system.

INTRODUCTION

The audio component of audiovisual productions has long been regarded as of minor importance. Nevertheless, in the last years and especially after productions such as Apocalypse Now (1979), its importance has been acknowledged. Sound designers 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). Sound FX providers rely on text descriptions to manage internally and sell their audio collections. Sound engineers search for sounds by matching a query against the descriptive keywords that a librarian has attached to each sound. There are several professional providers that offer SFX using keyword-matching as well as navigating through categories that organize the sounds in classes such as Animal, Cars, Human and so on (e.g.: www.sound-effects-library.com, www.sounddogs.com, www.sonomic.com). Web search engines such as www.altavista.com or www.singingfish.com offer audio search using standard text-based web retrieval indexing the words that appear near audio content in the HTML page.

Limitations of text-based approach

Discussion on the inadequacy of using text descriptors to describe sound is frequent in the literature. They point out that sounds are too difficult to describe with words. Perceptual descriptions are too subjective and may vary for different people. Source descriptions convey sometimes more descriptive power and are objective. However, sound may have been synthesized and have no clear origin. Other cons on current text-based approaches include:

- Library construction, that is, tagging of sounds with textual description, is a labour-consuming, error-prone task and yet the number of sound samples is constantly increasing.
- It is difficult for a librarian to add keywords that would match the ways users may eventually query a sound, e.g.: see Fig 1 for possible keywords to label a "golf drive".
- The sounds without caption are invisible to the users.
- Big corpuses may be labelled by different librarians that follow somewhat different conventions.
- 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 allow browsing for sounds in manually generated categories. The design and maintenance of category trees is complicated. It is very time consuming for a librarian to place a sound in the corresponding categories. Finally, It is difficult for users to navigate through somebody else's hierarchy.

In order to overcome the above shortcomings, solutions have been proposed to manage media assets from a content-based audio perspective, both from the academia and the industry. However, even though text-search has some shortcomings, content-based functionality should complement and not substitute the text search approach for several reasons: first, because the production systems work, second, because there is a great deal of legacy meta-data and new sound effects are released by the major vendors with captions, third, because text-retrieval is generally faster than content-based, and finally because users are familiar with using words to search for media assets.

Contributions

In this context, we present how to construct a SFX management system that incorporates content-based audio techniques as well as knowledge based tools built on top of one of the biggest sound effects providers database .

Overview on Sound FX representation problem: We review some taxonomic proposals for audio description found in the literature, which types of descriptors are actually found in SFX commercial systems and how it is dealt with within a multimedia standardization process such as MPEG7 [1].

Content-based tools to browse and search audio: We describe the integration of state-of-the-art content-based audio technologies in a commercial SFX management system.

From textual to concept-based: We describe how the use of a general knowledge network, such as WordNet [2], augmented with audio and post-production specific terms, can significantly:

- Ease the task of the librarian.
- Improve the control over the search process for the users.

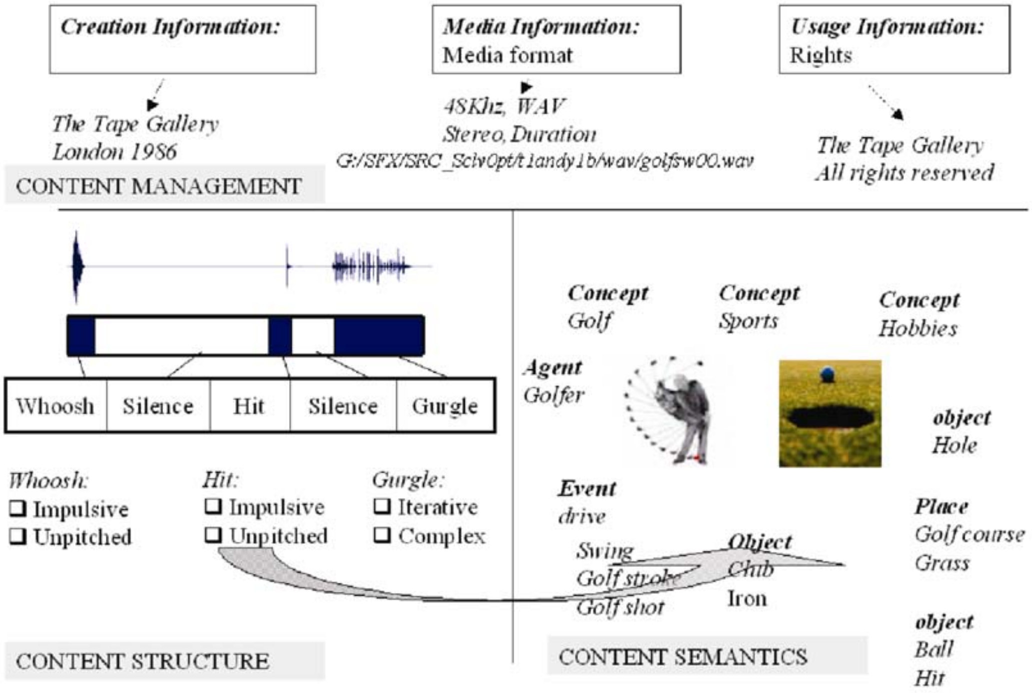


Figure 1 - Example of possible searchable metadata to describe a golf swing.

SOUND EFFECTS CATALOGING

MPEG7 offers a framework for the description of multimedia documents [1]. MPEG7 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, "Dog bark shepherd single bark indoor", or the context, "Airport crowds announcements".

Several ways of describing a sound include: semantic or source-centered description, perceptual, post-production specific and creation description (See Figure 1).

Semantic Descriptors

Semantic descriptors 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". 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.

The importance of source-tagging is put in doubt by Mott [3]. Mott explains that the sound engineer should concentrate on the sound independently of what actually produced it because in many occasions the natural sounds do not fulfil the expectations and must be replaced with sounds of distinct origin. Mimi Arsham, who worked on Ben-Hur, explains that the whip cracking sound was a hefty steak being slapped on a thigh. In any case, even if one thinks of the sound of the steak against the thigh, it is much easier, and cheaper, to download a steak sound than getting the steak and do the recording.

Perceptual Descriptors

They describe the perceptual qualities independently of the source. Since they refer to the properties of sound, e.g.: loudness, brightness, sometimes there is a "direct" 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: "roar, mmm, ring". Onomatopoeia are commonly used by librarians. Schaeffer did try to find a lexicon to describe sounds. He introduced the reduced listening 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 solfége of sound objects considered attributes such as mass (perception of "pitchiess") or harmonic timbre (bright/dull, round/sharp).

Post-production Specific Descriptors

Other important searchable metadata are post-production specific. According to Mott [3] the categories of sound effects are: natural sounds (actual source sound), characteristic sounds (what a sound should be according to someone), comedy, cartoon, fantasy.

Creation Metadata

Creation metadata describe relevant information on the creation or 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.

SYSTEM OVERVIEW

Text-based and content-based methods alone do not seem to suffice for specifying a sound effect. In the implemented system we aim to combine the best of both worlds to offer tools for the users to refine and explore a huge collection of audio.

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.

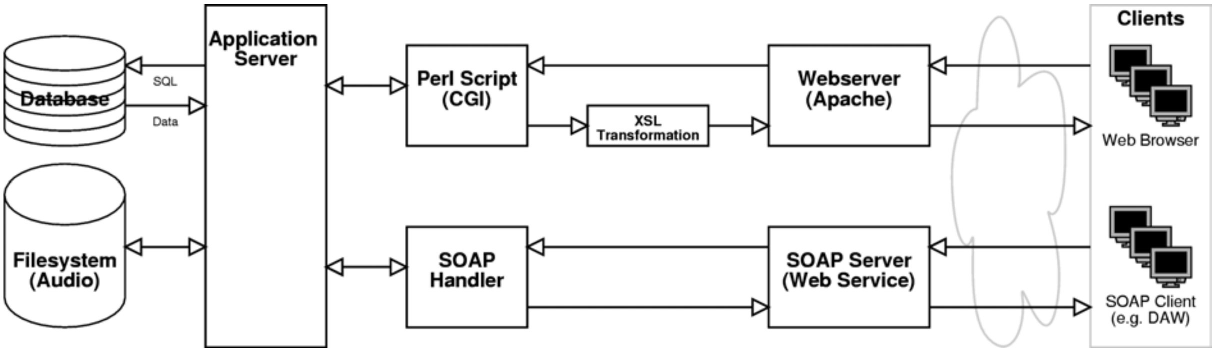


Figure 2 – Architecture of the system

System Architecture

The system architecture is depicted in Figure 2. The audio content is stored in the file system. Information about the audio files is stored in a SQL database. The main core functionalities reside on the Application server. Web-based technologies, including XML and Web Services allow seamless interoperatibility with external applications. The XML interface together with XSLT is used for the web interface. There are two possibilities of accessing the system: a web interface and via a SOAP interface. The SOAP interface (http://www.w3.org/TR/soap) makes possible the interaction between different platforms and digital audio workstation (DAW) environments.

Content-Based Audio Tools

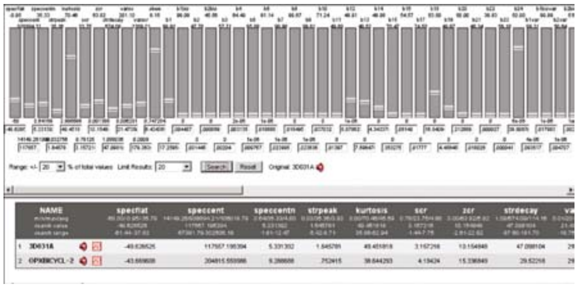
Content-based audio tools ease the work of the librarian and enhance the possibilities of search for the user. It simplifies the work of the librarian when labelling new sounds because many keywords are automatically proposed. To achieve it, the new sound is compared to the collection with Nearest Neighbour search and the text associated with the similar matches is presented in an ranked list.

Our experimental database consists of 80.000 sounds from the Sound-Effects-Library (www.sound-effects-library.com). These sounds have been unambiguously tagged with concepts of an enhanced WordNet (see [7] for details). Thus a violin sound with the following caption: "violin pizzicato D#" may have the following concepts:

- violin, fiddle - (bowed stringed instrument that is the highest member of the violin family; this instrument has four strings and a hollow body and an unfretted fingerboard and is played with a bow)
- pizzicato - ((of instruments in the violin family) to be plucked with the finger)
- re, ray - (the syllable naming the second (supertonic) note of any major scale in solmization)
- sharp - ((music) raised in pitch by one chromatic semitone; "C sharp")

Content-based tools offer the user functionalities such as:

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. Given the subjectiveness associated with distances, the user can adjust the weight of each dimension of the sound with the help of sliders (see Fig 3).



Range: +/- 20 % of total values Limit Results: 20

Original: 3D031A

	NAME	specflat	speccent	speccentn	strpeak	kurtosis	scr	zcr	strdecay
1	3D031A	-49.626525	117557.195394	5.331302	1.645785	49.451818	3.167218	13.154848	47.098104
2	OPXBCYCL-2	-43.669608	204815.555986	6.288608	.752415	38.644293	4.19424	15.336849	29.52258

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.

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

Filter by sound category: Another possibility is having trained models of certain classes of sounds, such as animal sounds, and answer queries like: give me baby cries that sound like a cat.

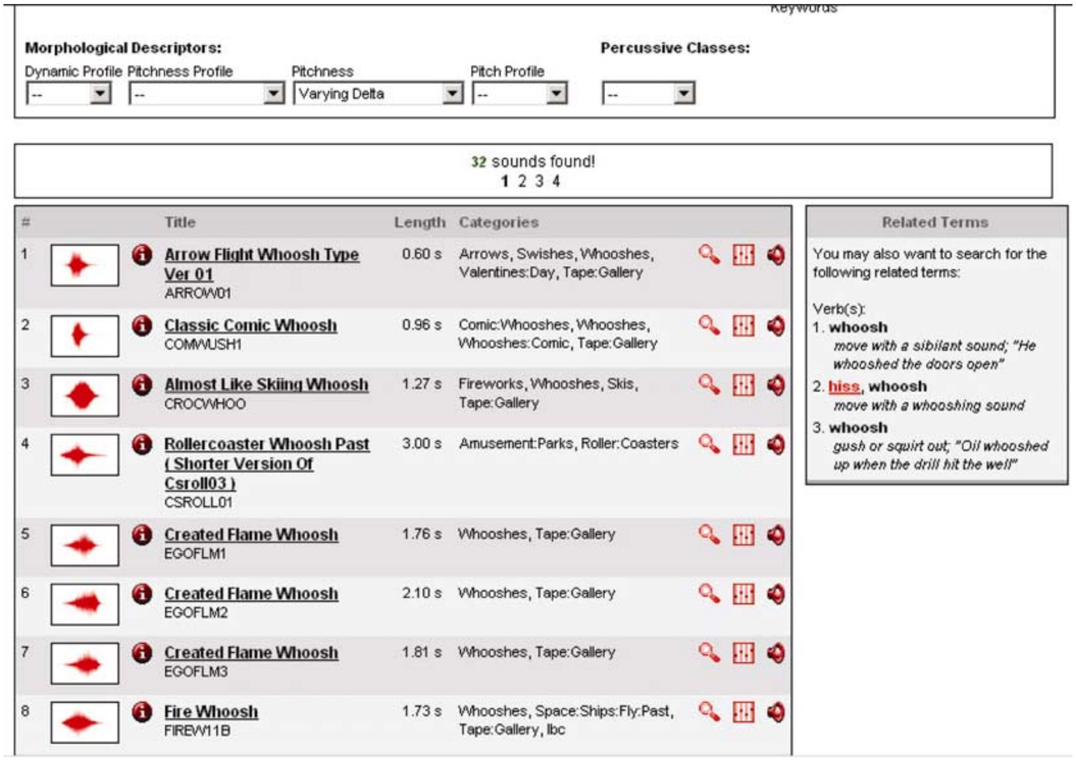
The content-based engine is described in [6]. The similarity measure is used for metadata generation: a sound sample will be labelled with the descriptions from the similar sounding examples of the annotated database. This type of classification is known as one-nearest neighbour decision rule (1-NN)[4]. The terms borrowed from the closest match are unambiguous due to the use of WordNet [2] as the taxonomy back-end. With unambiguous tagging, we refer to assigning concepts and not just terms to sounds. For instance, the sound of a "bar" is ambiguous, it could be "bar" as "rigid piece of metal or wood" or as "establishment where alcoholic drinks are served". The similarity measure is also used for the query-by-example and to browse through "perceptually" generated hyperlinks.

The evaluation of perceptual similarity distances which are the bases of some of the above functionalities is a tricky subject. Perceptual listening tests are expensive. Another possibility is to evaluate the goodness of the similarity measure examining the performance in the automatic metadata generation task. 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. Finally, another possibility is the consistency on the ranking and robustness to distortions such as resampling, transcoding (converting to MP3 format at different compression rates and back), equalization (low-pass, band-pass and high-pass filtering), background noise.

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.

- 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 navigation: The concept relations encoded in the lexical resource is used to propose 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 .
- 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.
- Module for the phonetic matching, e.g.: "whoooassh" retrieves "whoosh". Phonetic matching is used in information retrieval to account for 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.
- 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 labelled with the concept "farm". Studying the co-occurrence of concepts allows the system to infer related terms..



Keywords

Morphological Descriptors:

Dynamic Profile: --
Pitchness Profile: --
Pitchness: Varying Delta
Pitch Profile: --

Percussive Classes: --

32 sounds found!
1 2 3 4

#	Title	Length	Categories
1	**Arrow Flight Whoosh Type Ver 01** ARROW01	0.60 s	Arrows, Swishes, Whooshes, Valentines:Day, Tape:Gallery
2	**Classic Comic Whoosh** COMWUSH1	0.96 s	Comic:Whooshes, Whooshes, Whooshes:Comic, Tape:Gallery
3	**Almost Like Skiing Whoosh** CROCWHOO	1.27 s	Fireworks, Whooshes, Skis, Tape:Gallery
4	**Rollercoaster Whoosh Past (Shorter Version Of Csroll03)** CSROLL01	3.00 s	Amusement:Parks, Roller:Coasters
5	**Created Flame Whoosh** EGOFLM1	1.76 s	Whooshes, Tape:Gallery
6	**Created Flame Whoosh** EGOFLM2	2.10 s	Whooshes, Tape:Gallery
7	**Created Flame Whoosh** EGOFLM3	1.81 s	Whooshes, Tape:Gallery
8	**Fire Whoosh** FIREW11B	1.73 s	Whooshes, Space:Ships:Fly:Past, Tape:Gallery, lbc

Related Terms

You may also want to search for the following related terms:

Verb(s):
1. **whoosh**
   move with a sibilant sound; "He whooshed the doors open"
2. hiss, **whoosh**
   move with a whooshing sound
3. **whoosh**
   gush or squirt out; "Oil whooshed up when the drill hit the well"

Figure 3 – Snapshot of the HTML search front-end.

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 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 aforementioned similarity distance [4]. Another possibility of interaction with the sounds is using visualization techniques, specifically Multidimensional scaling (MDS), self-organizing maps (SOM) or FastMap [5], one can map the audio samples into points of an Euclidean space. MDS, for example, is used to discover the underlying (spatial) structure of a set of data from the similarity, or dissimilarity, information among them. It has been used for some years in e.g. social sciences, psychology, market research, physics. Basically the algorithm projects each object to a point in a k-dimensional space trying to minimize the stress function.

SUMMARY

We have introduced the difficulties inherent in describing a sound effect, both for the librarian and as a sound entry that can be accessed afterward by a user. 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 been described. WordNet, commonly used in other multimedia retrieval systems has been extended for sound effects retrieval.

All the functionality of the system can be accessed and evaluated at http://www.audioclas.org.

REFERENCES

[Banerjee and Pedersen, 2003] 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.

[5] 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.

[6] Cano, P., Koppenberger, M., Groux, S. L., Ricard, J., Herrera, P., and Wack, N., 2004. Nearest-neighbor generic sound classification with a WordNet-based taxonomy. In Proceedings of the 116th AES Convention.

[7] Cano, P., Koppenberger, M., Herrera, P., and Celma, O., 2004. Sound Effect Taxonomy Management in Production Environments. In Proceedings of the AES 25th International Conference.

[4] 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.

[3] L.Mott, R. (1990). Sound Effects: Radio, TV, and Film. Focal Press.

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

[2] Miller, G. A., 1995. WordNet: A lexical database for english. Communications of the ACM, pp 39-45.

ACKNOWLEDGMENTS

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.