

ALLTHATSOUNDS: ASSOCIATIVE SEMANTIC CATEGORIZATION OF AUDIO DATA

Julian Rubisch, Matthias Husinsky, Hannes Raffaseder

University of Applied Sciences, St. Pölten
Institute for Media Production
3100 St. Pölten, Austria

firstname.lastname@fhstp.ac.at

ABSTRACT

Finding appropriate and high-quality audio files for the creation of a sound track nowadays presents a serious hurdle to many media producers. As most digital sound archives restrict the categorization of audio data to verbal taxonomies, this process of retrieving suitable sounds often becomes a tedious and time-consuming part of their work. The research project AllThatSounds tries to enhance the search procedure by supplying additional, associative and semantic classifications of the audio files. This is achieved by annotating these files with suitable metadata according to a customized systematic categorization scheme. Moreover, additional data is collected by the evaluation of user profiles and by analyzing the sounds with signal processing methods. Using artificial intelligence techniques, similarity distances are calculated between all the audio files in the database, so as to devise a different, highly efficient search algorithm by browsing across similar sounds. The project's result is a tool for structuring sound databases with an efficient search component, which means to guide users to suitable sounds for their sound track of media productions.

1. INTRODUCTION

Supply and demand for digitally stored audio increased rapidly in the recent years. Number and diversity, as well as quality of available sound files reached an unmanageable amount. Efficient processes for the retrieval of audio data from large digital sound libraries play a pivotal role in the process of media production. In many cases, the user is required to know important features of the sound he is seeking, such as its source or excitation, in advance. On the other hand it is hardly possible to search for semantic features of a sound which are closer to human perception rather than technical parameters. Another obstacle towards an adequate verbal description of audio data emanates from the medium's inherent volatility, which makes it difficult to uptake and formally subsume acoustic events. Thus, not sounds themselves, much more the causing events are described.

The research project AllThatSounds aimed at facilitating the process of finding suitable sounds for media productions. For this purpose many different possibilities to categorize and describe sonic events were analyzed, evaluated and linked. Apart from the applicatory use of the tool, the research questions raised by the work on these topics trigger a discussion process about perception, meaning, function and effect of the sound track of a media product.

2. PRESENT SITUATION

As indicated above, sound designers and media producers often face the difficulty of retrieving appropriate sounds from huge digital audio libraries. The indexing of such databases is mostly restricted to verbal descriptions of the items' sources and excitations,

limiting the formulation of adequate search requests to these criteria. In addition, solely verbally tagged sound libraries generally display poor accessibility, since users often have to browse through huge flat text files, which also points at the need for a standardized sound categorization scheme and vocabulary. Hence, even though it might seem natural and sufficient to describe a sound by its source or the event it is caused by, these categorizations usually do not carry any semantic or sound-related information.

Still, the semantic information included in a sonic event, or its signal-related parameters present a useful search criterion for many possible applications. For example, in the field of movie sound design, for the dubbing of objects often sounds are used which in fact have nothing in common with the sonified object, apart from the transported meaning or certain signal characteristics. Even more apparent are use cases that deal with objects that do not even exist in reality, such as spacecraft engines or lightsabers. Since human auditory perception tends to accept differences between what is perceived visually and aurally as long as it occurs simultaneously, and the differences do not exceed a certain tolerable limit – an effect known as *synchresis* [2] – it is most sound designers' primary target to retrieve the sound that expresses a certain meaning best, which may not necessarily be accomplishable by the actual sound of the object in question.

Time is another factor that has to be considered when designing a search interface for sound databases. Since in many cases it is unfeasible to spend a large amount of time for the retrieval of an optimal sound, it was a primary objective of the prototypical development to devise an optimized user interface which is capable of assisting the user in his search process.

3. DESCRIPTION AND CATEGORIZATION OF AUDIO FILES

A central objective of the research project was to design a sound database with enhanced retrieval possibilities. A design issue that had to be tackled results from the fact that sound is a carrier of information in many different dimensions. Aside from technical signal parameters, many acoustic events are tagged with a multitude of meanings and messages that originate e.g. from sociocultural contexts associated with them.

Therefore a primary obstacle in designing a taxonomy for sounds results from the abundance of possible categories that have to be considered. Furthermore, these categories are subject to transformation relating to cultural differences and temporal developments. For example, the encoded meaning of a typewriter has altered from *modern*, 50 years ago, to *nostalgic* or *anachronistic* nowadays.

In order to address these design issues, a multi-dimensional ap-

proach was taken, consisting of four different approaches:

- Descriptive Analysis
- Listeners' Analysis
- Machine Analysis
- Semantic Analysis

3.1. Descriptive Analysis

The descriptive analysis enables for the uniform description of acoustical events from the sound designer's perspective. The aim is that already at the time the upload of a sound into the database takes place, it can be sufficiently and distinctly categorized. Based on relevant literature [5] [6] [7] [8] [9], a general classification scheme of audio events was developed, which allows for a differentiated description in the following categories:

- *Type of Sound*
(music, speech, sound effect, ambience)
- *Source*
(technical/mechanical, musical instruments, objects/materials, ambience, synthetical, human, animal, nature)
- *Stimulation*
(time structure, intensity, type of stimulation)
- *Room*
(reverberation, inside/outside)
- *Timbre*
(harmonicity, brightness, volume)
- *Pitch*
- *Semantic Parameters*
(personal distance, familiarity, emotion/mood)
- *Other Parameters*
(source velocity, complexity)

A discrepancy that has to be addressed here concerns the amount of detail that is necessary for a satisfactory categorization of audio data. While it becomes evident from the considerations mentioned above, that in order to provide an optimal search accuracy, sounds also have to be tagged with high precision and complexity, such a process also requires a lot of time. To keep time short when describing the sound at upload, full description cannot be reached.

3.2. Listeners' Analysis

To overcome the drawback that a single user's description of a sound may lead to an unwanted result in another context, collaborative methods using Web 2.0 technologies are employed. Tags, comments and descriptions of possible usages of a sound provide further assistance when trying to retrieve appropriate audio files.

Moreover, users are given the possibility to contribute alternative categorizations of sounds to the system and thus expressing their own interpretation of the sound in question. A large data pool also represents a basis for further analysis which is explained in section 3.4.

3.3. Machine Analysis

Contemporary off-the-shelf computers offer enough computational power to enable for an automated processing and categorization of audio signals. In particular, such methods seem to be a promising approach regarding the categorization of raw audio data from existing archives or large databases, which are to be integrated in the system. To accomplish a technical analysis of audio files, two different methods were employed.

3.3.1. MPEG-7 Features

The MPEG-7 standard¹ defines a variety of audio descriptors which are usable for the annotation of multimedia files. Some of these features are suitable for the estimation of psychoacoustic measures, such as sharpness, volume, harmonicity and loudness of a sound. In a preliminary listening experiment, the AudioSpectrumCentroid, AudioSpectrumSpread, AudioHarmonicity and AudioPower descriptors respectively were determined to model the mentioned characteristics best.

Based on these descriptors, the users are provided suggestions of these factors, with the intention to accelerate the process of entering data into the system. Through the evaluation of the estimated values by user corrections, it seems that especially the perceived sharpness, harmonicity and loudness can be approximated by the mentioned MPEG-7 features. We are planning to do an exhaustive quantitative analysis of the collected data in the near future.

3.3.2. Similarity Analysis

As described in section 2, a considerable acceleration of the search process can be anticipated by computing similarity measures of all the sounds in the database. To achieve this, a similarity model based on Mel Frequency Cepstral Coefficients (MFCCs) is used to compare all the sounds with one another. This model proved to be very useful in comparing and classifying sounds and music [3].

Similarity measures are obtained by computing the mean values as well as covariances and inverted covariances of the first five MFCC coefficients on a single-frame basis, and calculating a Kullback-Leibler divergence afterwards [4].

For every sound in the database, the 20 most related sounds are stored, which allows the user to browse for viable sounds in a novel way. Early tests have proved this feature to be of great value to the users, because thereby sounds that wouldn't have been found by pure metadata search can be discovered.

3.4. Semantic Analysis

In media production the semantic content of acoustic events plays a crucial role to achieve a certain perception at the consumer. Conclusions regarding semantic content of a sound obtained through signal analysis methods would present a rewarding facilitation of the search process. However, until today it is almost impossible to automatically extract semantic meaning from audio signals, since not only signal parameters, but to a wide extent also cultural and social experience of listeners play a role.

The project aims at combining the collected user metadata and calculated sound features so as to explore possible correlations between signal parameters and semantic denotations. At present, such an approach is still unfeasible due to a lack of a sufficient number of user annotations. We intend to further investigate this field when numerous listeners have contributed to the database for a longer period of time.

4. PROTOTYPE

The first prototype with an improved search tool was released to the public on May 30th 2008 as a web application². As of January 30, 2009, about 3700 sounds primarily recorded by students of

¹<http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>

²<http://www.allthatsounds.net>

the University of Applied Sciences St. Pölten's Media Engineering course. Sounds can be up- and downloaded in uncompressed wave-format and used under the Creative Commons Sampling Plus license³.

4.1. Usage & Architecture

To use the application and contribute to it, a user is required to register. Afterwards he or she is entitled to up- and download sounds to and from the database. Upon upload, the user is asked to provide a short name and description of the sound, as well as keywords that can be used for search purposes afterwards. The mentioned MPEG-7 descriptors are computed and employed to supply recommendation values for the sound's loudness, sharpness, harmonicity and volume. Additionally, the user is required to provide a first categorization according to the mentioned taxonomy. At the same time, the 20 most similar sounds are calculated and stored in the database. Furthermore, similarity distances for all sounds are renewed on a daily basis.

For retrieval purposes, a primitive text-based search can be used as well as the extended feature-based search (see figure 1). The user is furthermore enabled to obtain a sound's most similar counterparts from its detail page.

To improve architectural flexibility, the prototype's core functionality is exposed as a SOAP web service, enabling the implementation of a wide range of clients, including e.g. standalone applications as well as Digital Audio Workstation (DAW) plugins.

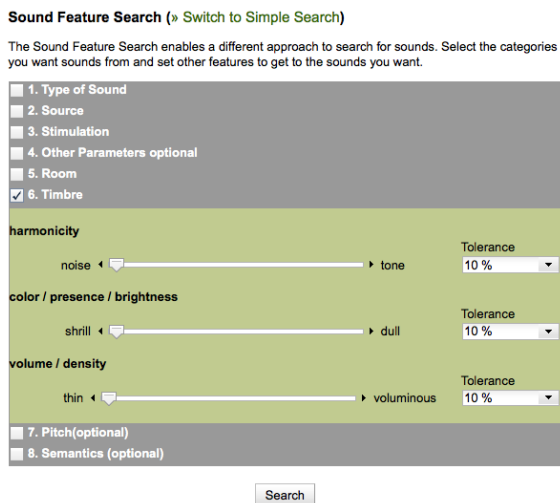


Figure 1: Feature-based search form

4.2. Contents

This first prototype focuses on media sounds (jingles, music beds etc.), environmental or ambient sounds as well as sound effects, or Foley sounds. While it is of course possible to upload music files, the musical features are not separately treated by the implemented algorithms. Durations range from approx. 30 milliseconds to about 9 minutes, with an average of 12.8 seconds.

Most of the sounds currently in the database have been provided by students of the University of Applied Sciences St. Pölten who

were asked to record and upload sounds of a certain timbre, context or envelope as part of a course assessment.

4.3. Evaluation

At the present stage of the prototype, the evaluation process is limited to internal reports by university employees and students, but we also encourage external users to provide feedback and report bugs.

Furthermore, several efforts are underway to form a cooperation with Austrian free radio stations to obtain feedback also from professional media producers.

5. RELATED WORK

Many insights that the concepts of the prototype's machine analysis are based on originate from Music Information Retrieval (MIR) methods. There, researchers deal with pieces or collections of pieces of music, necessitating the consideration of the large area of music theory.

Related studies concerning the automated classification of sound effects have been conducted by Cano et. al. [1]⁴.

6. CONCLUSION AND PERSPECTIVE

The web application prototype released in May 2008 is still in a beta stadium, hence several requirements for use in a production environment are not yet completely met. However, users reported great convenience improvements by the use of the enhanced search process, especially concerning the possibility to browse by similarity.

Even though the prototype has already proved the usefulness of the multi-dimensional approach used in AllThatSounds, it also revealed a lot of new research questions. In future research projects a deeper examination of the semantic content of acoustic events is absolutely necessary. As has already been pointed out, a larger data set is necessary to explore possible correlations between signal characteristics and semantic denotations of a sound.

Moreover, concerning the usability and effectivity of the prototype, further studies have to be conducted. Enhancements of workflow could be reached by the implementation of a Naïve Bayes text classifier for the estimation of probable sound sources and excitation types from the initial set of keywords. Also, pattern recognition methods which can be employed to classify and label sounds solely by taking into account their signal characteristics will have to be evaluated. Additionally, when a larger data pool has been reached, it will be inevitable to revise the current categorization taxonomy, as it sometimes produces redundant information, and it still takes too much time for the user to complete the sound classification.

7. ACKNOWLEDGEMENT

Starting in October 2005 a research team at the *University of Applied Sciences St.Pölten* worked together under project leader Hannes Raffaseder with the Vienna based companies *Audite / Interactive Media Solutions, Team Teichenberg* and, since April 2007 also the

³<http://creativecommons.org/licenses/sampling+/1.0/>

⁴<http://audioclas.iaa.upf.edu/>

University of Applied Sciences Vorarlberg. The works were supported by the *Österreichische Forschungsförderungsgesellschaft (FFG)* in their FHplus funding programme until May 2008.

8. REFERENCES

- [1] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, P. Herrera, N. Wack, "Nearest-neighbor generic sound classification with a wordnet-based taxonomy" in *Proceedings of AES 116th Convention*, Berlin, Germany, 2004
- [2] M. Chion, *Audio-Vision - Sound on Screen*, Columbia University Press, 1994
- [3] D. Feng, W.C. Siu, H.J. Zhang, *Multimedia Information Retrieval and Management*, Springer-Verlag, Berlin, 2003
- [4] Michael M. Mandel, Daniel P.W. Ellis, "Song-Level Features and Support Vector Machines for Music Classification" in *Proceedings of the 6th International Conference on Music Information Retrieval*, London, 2005, pp. 594-599.
- [5] H. Raffaseder, *Audiodesign*, Hanser-Fachbuchverlag, Leipzig, Germany, 2002
- [6] M. R. Schafer, *The Soundscape - Our Sonic Environment and the Tuning of the World*, Destiny Books, Rochester, 1994
- [7] D. Sonnenschein, *Sound Design - The Expressive Power of Music, Voice and Sound Effects in Cinema*, Michael Wiese Productions, Studio City, 2001
- [8] B. Truax, *Acoustic Communication (2nd ed.)*, Ablex Publishing, Westport, 2001
- [9] T. van Leeuwen, *Speech, Music, Sound*, MacMillan Press Ltd., London, 1999