INTELIGENT INFRASTRUCTURE FOR ACCESSING SOUND AND RELATED MULTIMEDIA OBJECTS

Ivan Damnjanovic, Chris Landone, Panos Kudumakis and Josh Reiss

Queen Mary, University of London
Mile End Road, E1 4NS, London
United Kingdom
{ivan.damnjanovic, chris.landone, panos.kudumakis, josh.reiss}@elec.qmul.ac.uk

ABSTRACT

In recent years, digital sound material becomes more and more available. However, there is still lack of qualitative solutions for access to digital sound archives. Not only that there is no consistency in formats of archived materials with related media often in separate collections, but related metadata are given in non-standard specialist format, incomplete or even erroneous. Hence, the full value of the archived material is hidden from the end user. EASAIER addresses these issues with the development of an innovative remote access system which extends beyond standard content management and retrieval systems. The EASAIER system focuses on sound archives, libraries, museums, broadcast archives, and music schools. However, the tools may be used by anyone interested in accessing archived material; amateur or professional, regardless of the material involved. Furthermore, it enriches the access experience enabling the user to experiment with the materials in exciting new ways. The system features; enhanced cross media retrieval functionality, multi-media synchronisation, audio and video processing, analysis and visualisation tools, all combined within in a single user configurable interface.

Index Terms— Sound Archives, Multimedia Retrieval, Music Ontology, marking, looping, time-scaling.

1. INTRODUCTION

Recent works [1-3] has provided a systematic study of what end users want from Music Information Retrieval systems. User needs studies [4] and extensive research by the JISC [5-7] has identified a number of key features that are required in order to enrich sound and music archives. These findings stress the need for web-based access, integration of other media, time-stretching functionality, and alignment of scores with audio, among many others. They confirm and build on previous work on user needs for digitized audio collections such as the ground breaking exploratory studies carried out more than 10 years ago by the Library of Congress [8] or the Jukebox and Harmonica projects[9, 10]. The development of Indiana University's world-leading Variations project [11] was founded on close analysis of users' needs – particularly music students' need for annotation and visualisation tools to help them learn with digital music content. However, a recent Scottish study into training needs analysis in e-learning [12] reported that audio is still an under-used technology. The UK Arts and Humanities Research Council's ICT Programme has recently funded work surveying the needs of the research community in searching and analysis tools for audio streams [13].

EASAIER address several areas that still lack a deep, systematic, and focused approach: multi and cross- media retrieval, interactivity tools, integration of speech and music processing methods, and systemic archive analysis. In order to cope with these kinds of problems, innovative audio processing, data mining, and visualization techniques, alongside proper user needs and evaluation studies, are being developed and integrated into prototypes. These will be deployed in several sound archives in order to demonstrate a qualitative jump in usability, effectiveness and accessibility.

During designing stage of the EASAIER system a number of key challenges were identified:

- Integrating speech and music technologies to enable user to have common access point for archives and web resources
- Access to related materials (image, video, text) that is stored in the archive or aggregated from the web.
- Establishing a *common set of metadata* and provide a mapping for various existing archive ontologies.
- Establishing a common timeline to enable easier access to the documents and its segments.
- Integrating low-level similarity and metadata retrieval
- Synchronisation of media components for enriched access and visualisation

In following section, system architectures and needed components to address these challenges are described.

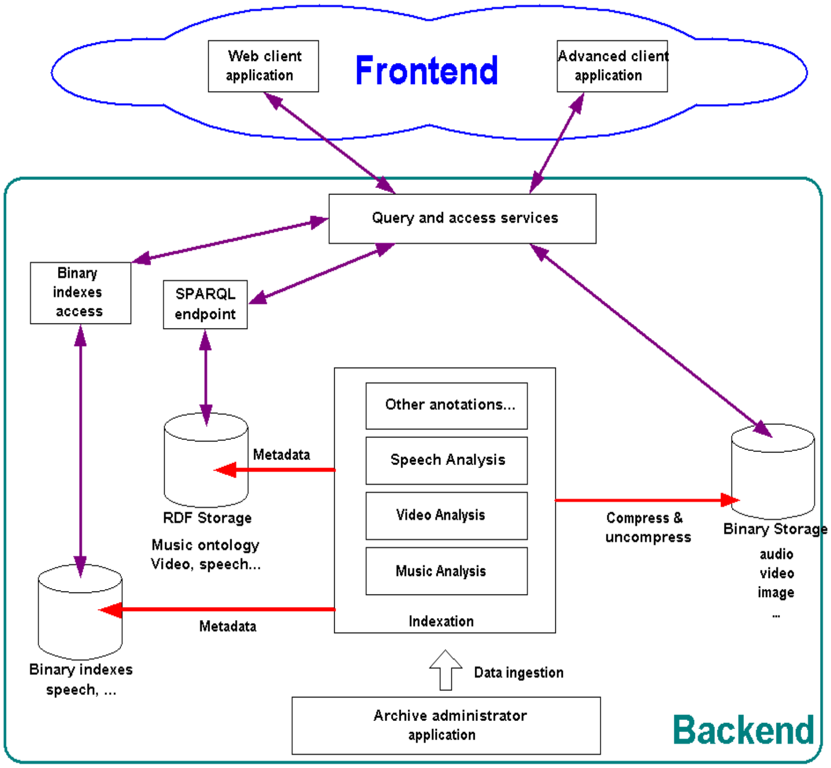


Figure 1 EASAIER system architecture

2. EASAIER SYSTEM ARCHITECTURE

The EASAEIR system is realised in typical client/server architecture (Figure 1). It is designed with special consideration of following requirements:

- Possibility to express relations between every data assets. Data and metadata are treated at the same level, linking one to each other.
- Interlinking distributed archives.
- Automatic and semi-automatic multimedia assets feature extraction.
- Enriched access to sound archives.

In order to address needs of professionals as well amateurs the front-end consists of two types of clients: web client and advanced user application. The web client application (Figure 2) allows the user to browse and query the archive according to the retrieval system functionalities. Its functionalities are restricted due to the web application framework and subsequent limitations in speed, memory and processing power. Simple retrieval, playback and visualization are provided but expensive processing like real-time audio filters are not supported. The advanced user application, beside functionalities of web client, allows enriched access, visualization of an audio file and its related metadata and media. The user is able to interact with audio resources through separating and identifying sources, processing and modifying the audio, and aligning various sources at playback.

On the back-end side, there are components to support advanced retrieval functionalities:

- The database storage composed of an RDF storage for storing objects identifiers, metadata and related media in a semantic form and a media storage for binary data (audio, video, images…)
- The SPARQL end-point to query the RDF storage and return results to the front-end applications (by the way of a web server for the web client).
- The database administration application which is in charge of database administration (add, modify or remove data from the database, manage user and rights). The administration tool is assisted by a feature extraction module to extract automatically metadata from a multimedia file.

The metadata addressed, ranges from low-level features to higher level editorial metadata on individual pieces of audio material. To address all these types of metadata EASAIER retrieval system is build around the Music Ontology [14] together with its extension the Speech Ontology. In addition, for purpose of integrating existing archives EASAIER provides automatic mapping to a metadata representation standards, such as Dublin Core [15]. However, to fulfill custom metadata representation mapping needs to be done manually. Such a case was the HOTBED [16] archive that we mapped to the Music Ontology, so EASAIER retrieval systems and tools can be used for querying and processing HOTBED material.

The Music Ontology is built on the Timeline ontology [17] to cover temporal information. The Timeline ontology defines a TimeLine concept, which represents a coherent backbone for addressing temporal information. An instance of this class covers the physical time line: the one on which we can address a date. Another instance may back an audio signal, and can be used to address a particular sample.

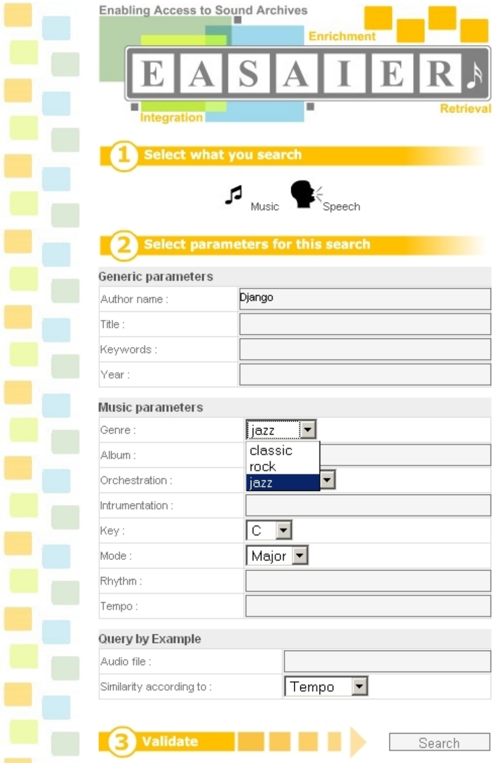
The Music Ontology also uses the Event ontology [18], which defines an Event concept, having a number of factors (such as a musical instrument, for example), agents (such as a particular performer), products (such as the physical sound that a performance produces) and a location in space and time (according to the Timeline ontology). This definition is broad enough to include performances, compositions, recordings, but also ‘artificial’ classifications: structural segmentation, chord extraction, onset detection, etc.

Two more ontologies are incorporated to the Music Ontology: the Friend-of-a-friend (FOAF [19]) ontology and the Functional Requirements for Bibliographic Records ontology (FRBR [20]). First enables modeling of musical groups, artists, labels, and other music-related agents. Latter is used for its concepts of Work, Manifestation and Item.

Furthermore, to retrieve related metadata and material from the web, the EASAIER Semantic Music Retrieval [21] incorporates content from Google, YouTube, LyricWiki, Yahoo Images & Music, Amazon, eBay, and last.fm. The EASAIER approach strongly relies on the structure of the MusicOntology enabling for example, in the case of the MusicBrainz dataset, to browse artists in a structured manner, e.g. according to their unique MusicBrainz ID, while also providing unstructured or multimedia-based content that is aggregated by keywords.

3. FEATURE EXTRACTION

Retrieval of media assets from a repository is traditionally accomplished by forming a query based on textual annotations that are entered by content managers during the archival process. Typically these annotations are limited to information concerning the origination of the assets, with very little data describing the actual content, especially as far as audio and video assets are concerned. The EASAIER client enhances the interaction capabilities of an end-user with sound archives through the exploitation of the content within. The client application enables different representations of an audio stream, such as waveforms, spectrograms and several other visualizations useful for education and analysis. Available metadata is increased by the extraction of various low, mid and high level features which enables greater search functionality.



Generic parameters

Author name : Django
Title :
Keywords :
Year :

Music parameters

Genre : jazz
classic
rock
jazz
Album :
Orchestration :
Intrumentation :
Key : C
Mode : Major
Rhythm :
Tempo :

Query by Example

Audio file :
Similarity according to : Tempo

Figure 2: EASAIER Query Interface (web client)

Automatically extracted metadata is broadly divided into four main categories.

- *Description of the logical container* - This data describes how the actual content of the asset is encoded. In the instance of an audio media file this information would indicate the type of audio coding strategy, sampling rate and number of audio channels.
- *Embedded information* - An increasing number of media file formats have embedded metadata containers that allow information such as the title, artist name, genre, copyright and other “tags” to be stored within the file itself. This information is extracted and can be used to complement or substitute the textual data that is manually entered by the content manager.
- *Machine-readable content descriptions* - In order to enable the system to perform computationally efficient similarity-based searches, low-level features that describe a models of the asset's content are extracted. For instance, in the instance of the EASAIER music retrieval prototype, a model based on MFCC is created during the ingestion process which is later used to retrieve assets that exhibit similar timbral profiles.
- *Human readable content descriptions* - This type of automatically extracted metadata describes the content of the asset using a terminology that is familiar to the end user and that can therefore be used directly as parameters in queries. In EASAIER this type of

metadata is used to describe the musical and/or speech qualities of an audio asset, for instance the tempo, meter, key, instrumentation, gender of a speaker and so on.

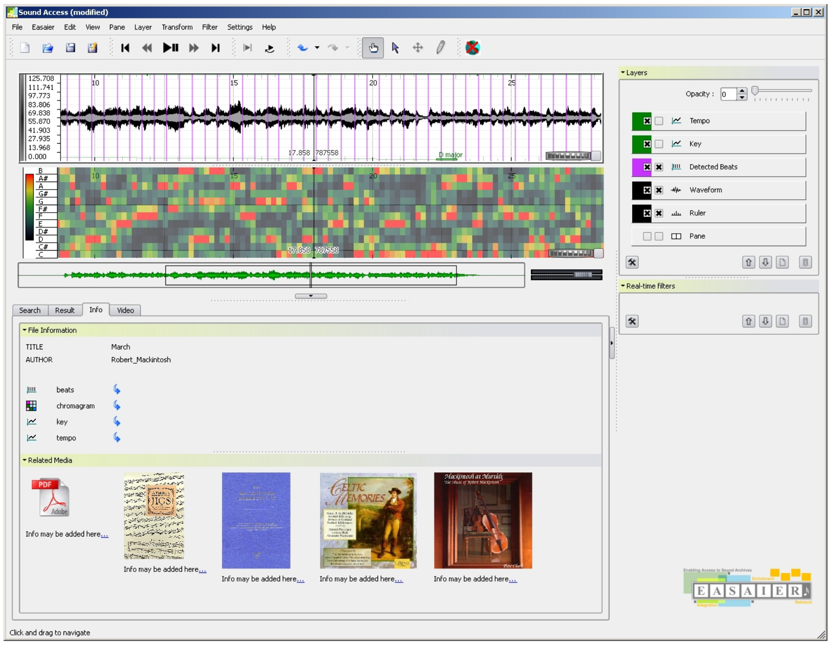
The automatic extraction of metadata in is carried out by a stand-alone service that communicates with the binary storage, RDF storage and Archiver components of the EASAIER system.

4. ENHANCED CONTENT RETRIEVAL

The EASAIER system provides multiple online retrieval systems, allowing for searching of content and metadata using multiple techniques and modalities**:**

-
- *Music retrieval* - Music retrieval involves searching and organising audio collections according to their relevance to music-related queries. This process consists of the generation of compact representations for both the query and the collection and the search for similarities between these representations. Most music retrieval systems use low-level features which allow fingerprinting of audio files, but are limited to only exact match retrieval. By using appropriate higher level features, ranked lists of audio files are obtained related to the query through melodic and harmonic similarity.
- *Speech retrieval* - The speech retrieval features complement the music retrieval features in order to support the interrogation of archives with mixed content. Furthermore the speech-specific archives (legal recordings, lectures, audio books, broadcast recordings) will use solely the speech retrieval features. In the case of mixed archives, the speech and music parts of the sound materials are managed separately, using the adequate algorithms. The speech/non-speech/music segmentation ensures the separate preparation, sound object identification and indexing.
- *Cross-media retrieval* – This allows the user to search media in various formats (audio recordings, video recordings, notated scores, images etc…) and find related material across different media. For instance, a search for similar media to a piece of music could result in musically similar pieces as well as relevant text and video linked to the song or performer. The establishment of linked metadata will enrich the content by allowing for association of separate media.

One of the key challenges in designing retrieval systems in EASAIER was integration of metadata and low-level similarity queries. In the case of the music retrieval module, there is both audio similarity search and search on features. The audio similarity metric used is in Soundbite [22] (audio similarity engine used in EASAIER), is fundamentally different from the metric used for searching across metadata. Furthermore, it is proprietary and undisclosed, whereas the similarity metric for searching on metadata is part of the open EASAIER system architecture.



Opacity : 0

TITLE March
AUTHOR Robert_Mackintosh

Click and drag to navigate

Figure 3: EASAIER Client Interface

Suppose the search is over some number N of features, f_1, f_2, … f_N. For the query, each feature has an assigned confidence, q_1, q_2, … q_N , and for a track in the collection each feature has an assigned confidence, t_1, t_2, … t_N. As an example, consider a query song with bpm known to be 120 with a confidence of 1. On the other hand, a track in the collection can have automatically estimated bpm of 120 with a confidence of 0.7. If we refer to the audio similarity measure as AS and its weighting as w_{AS}, then a combined similarity measure may be given as:

$$\frac{\sqrt{w_1 q_1 (1-t_1)^2 + w_2 q_2 (1-t_2)^2 + ...w_N q_N (1-t_N)^2 + w_{AS} AS^2}}{\sqrt{w_1 q_1 + w_2 q_2 + ...w_N q_N + w_{AS}}} \qquad (1)$$

Weightings w_1, w_2,… w_N are assigned to each feature as well, giving a fully flexible similarity measure that takes into account both the importance of each feature and the confidence in the estimation of that feature, for both query and retrieved track.

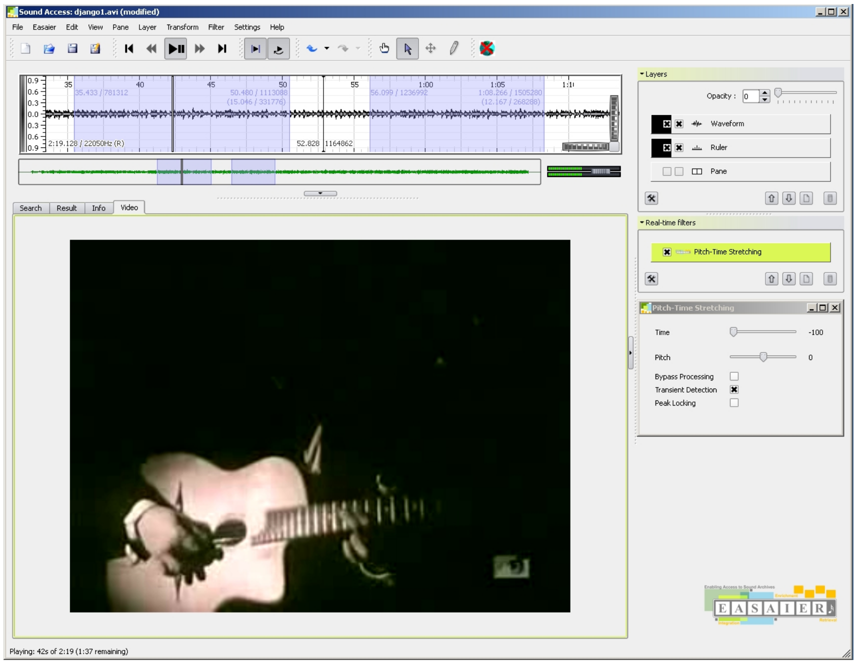
5. ENRICHED INTERACTIVITY

Tools to allow time aligned textual markup are provided to the user. Sections of audio can be selected manually or automatically and looped seamlessly for learning or analysis purposes. Advanced audio signal processing tools allow the user to specify how they listen to and interact with the media content. A source separation tool allows a user to listen to individual instruments within the piece of music while a noise reduction tool can be used to eliminate unwanted artefacts. Figure 3 shows a screen shot of the client interface with audio content in both the time and frequency domain along with related media such as pictures and text.

Time-stretching allows a user to slow down (or speed up) recordings, without modifying the pitch in real-time. This enables a music student or musicologist for example, to easily learn or analyze a piece of music. The ability to speed up the audio content also gives the user the ability to browse long segments of audio rapidly. The same technology also allows for pitch-shifting of the audio without affecting the time scale. A key innovation also allows the video stream to be synchronised with the audio during time and pitch scaling. It is also possible to zoom the video content, allowing closer inspection of an instrumentalist's particular technique. This suite of enriched access tools is presented in real-time with all functionalities accessible simultaneously.

5. CONCLUSIONS

A major driving force for the work presented is that sound archives still lack of a solution for qualitative access to materials. Our main aim is to develop a state-of-the-art access system for sound archives, incorporating multiple, integrated retrieval systems, and enriched access tools



Opacity : 0

☒ ☒ Waveform
☒ ☒ Ruler
☐ ☐ Pane

☒ Pitch-Time Stretching

Pitch-Time Stretching

Time -100
Pitch 0
Bypass Processing ☐
Transient Detection ☒
Peak Locking ☐

Playing: 42s of 2:19 (1:37 remaining)

Figure 4: EASAIER Client Interface with video time scaling

which allow manipulation of the resources. The user requirements for such a system are carefully studied and considered during design process. Having this in mind, the key challenges to integration and development of needed technologies were identified. System architecture and integrated technologies, such as Music ontology and metrics for integrated audio similarity and metadata retrieval, addressing mentioned challenges are described. At present, the EASAIER project has integrated most required functionalities, including the music retrieval system and demonstrators of integrated components, such as the metadata extractor and client prototype. The potential user community for sound archives using EASAIER is large and wide ranging. It is thus hoped that we will be able to deploy the system on a large scale, benefiting sound archives in cultural heritage institutions which until now, have been unable to provide advanced access systems to their users.

ACKNOWLEDGMENTS

This work has been partially supported by the European Community 6th Framework Programme project EASAIER (IST-033902): www.easaier.org.

REFERENCES

[1] S. Downie and S. J. Cunningham, "Toward a theory of music information retrieval queries: System design implications," Proceedings of the Third International Conference on Music Information Retrieval (ISMIR), Paris, France, 2002.

[2] D. Bainbridge, S. J. Cunningham, and S. Downie, "How people describe their music information needs: A grounded theory analysis of music queries.," Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR), Baltimore, Maryland, 2003.

[3] S. J. Cunningham, N. Reeves, and M. Britland, "An ethnographic study of music information seeking: implications for the design of a music digital library," Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries (JCDL), Houston, Texas, pp. 5 - 16, 2003.

[4] S. Barrett, C. Duffy, and K. Marshalsay, "HOTBED (Handing On Tradition By Electronic Dissemination)," Royal Scottish Academy of Music and Drama, Glasgow, Report March 2004.www.hotbed.ac.uk

[5] M. Asensio, "JISC User Requirement Study for a Moving Pictures and Sound Portal," The Joint Information Systems Committee, Final Report November 2003.
www.jisc.ac.uk/index.cfm?name=project_study_picsounds

[6] "British Library/JISC Online Audio Usability Evaluation Workshop," Joint Information Systems Committee (JISC), London, UK 11 October 2004.
www.jisc.ac.uk/index.cfm?name=workshop_html

[7] S. Dempster, "Report on the British Library and Joint Information Systems Committee Usability Evaluation Workshop, 20th October 2004," JISC Moving Pictures and Sound Working Group, London, UK 20 October 2004

[8] "Final Report of the American Memory User Evaluation, 1991-1993," American Memory Evaluation Team, Library of Congress, Washington, DC 1993.
memory.loc.gov/ammem/usereval.html

[9] E. Fønss-Jørgensen, "Applying Telematics Technology to Improve Public Access to Audio Archives (JUKEBOX)," Århus, Denmark 1997.
www.statsbiblioteket.dk/Jukebox/finalrep.html

[10] R. Tucker, "Harmonised Access & Retrieval for Music-Oriented Networked Information (HARMONICA)," 1997-2000. /projects.fnb.nl/harmonica/

[11] J. W. Dunn, M. W. Davidson, J. R. Holloway, and G. Bernbom, "The Variations and Variations2 Digital Music Library Projects at Indiana University," in Digital Libraries: Policy, Planning and Practice, J. Andrews and D. Law, Eds.: Ashgate Publishing, 2004, pp. 189-211.

[12] "Higher Education Training Needs Analysis (HETNA)," Scottish Higher Education Funding Council (SHEFC), Sheffield, UK November 2004.
www.shefc.ac.uk/about_us/departments/learning_teaching/hetna/hetna.html

[13] A. Marsden, "ICT Tools for Searching, Annotation and Analysis of Audio-Visual Media," UK Arts and Humanities Research Council, Lancaster, UK September 2005.
www.ahrbict.rdg.ac.uk/activities/marsden.htm

[14] Y. Raimond, S. Abdallah, Mark Sandler and Frederick Giasson, “The Music Ontology”, Proceedings of the International Conference on Music Information Retrieval, 2007

[15] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, RFC 2413 - Dublin Core Metadata for Resource Discovery, 1998

[16] S. Barrett, C. Duffy, and K. Marshalsay, “HOTBED (Handing On Tradition By Electronic Dissemination),” Royal Scottish Academy of Music and Drama, Glasgow, Report March 2004. www.hotbed.ac.uk

[17] Y. Raimond and S. A. Abdallah, “The Timeline Ontology”, 2006

[18] Y. Raimond and S. A. Abdallah, “The Event Ontology”, 2006

[19] Dan Brickley and Libby Miller, FOAF Vocubulary Specification, 2005

[20] I. Davis and R. Newman, Expression of Core FRBR Concepts in RDF, 2005

[21] M. Luger, Y. Ding, F. Scharffe, R. Duan, Z. Yan, “EASAIER Semantic Music Retrieval Portal”, The 2nd international conference on Semantics And digital Media Technologies (SAMT), Genova, Italy, 5-7 December 2007.

[22] M. Levy, M. Sandler and C. Sutton, “Soundbite”, http://www.isophonics.net/SoundBite