From Multimedia to the Semantic Web using MPEG-7 and Computational Intelligence

G. Tummarello, C. Morbidoni, P. Puliti, A. F. Dragoni, F. Piazza
Dipartimento di Elettronica Intelligenza Artificiale e Telecomunicazioni
Universita' Politecnica delle Marche (Ancona, ITALY)
info@semanticweb.deit.univpm.it

Abstract

In this paper we present an architecture that provides Semantic Web annotations of sound clips described by MPEG-7 audio descriptions. The great flexibility of the MPEG-7 standard makes it especially difficult to compare descriptions coming from heterogeneous sources. To cope with this, the architecture will first obtain "normalized" versions of the audio descriptions using different adaptation techniques. Once in a "normalized" format, descriptions can be then projected into uniform and semantically relevant vector spaces, ready to be fed to a variety of well known computational intelligence techniques. As higher semantic results are then available, these can be exported as interoperable (RDF) annotations about the resource that was originally fed into the system.
As novel aspect, through the use and interchange of MPEG-7 descriptions, the framework allows building applications (e.g. classificators) which can provide annotations on distributed audio resource sets.

1. Introduction

Even since the first version, published in 2001 [1], MPEG-7 featured a very large quantity of descriptors allowing annotation of audio video sequences with information ranging from low level statistical features to purely semantic concepts. Although the standard was developed along with an experimental codebase (the so called XM, experimental model [2]), there is currently a notable lack of actual applications using it. This sort of initial "implementation flop" can be explained by looking at the complexity of the standard under the interoperability aspect. In fact, while easy to create MPEG-7 compliant descriptions, the freedom in terms of structures and parameters is such that generically understanding MPEG-7 produced by others is difficult at least.

While computational intelligence techniques are directly mappable to the applications evised for the standard [11], the actual usage of these is not direct. As MPEG-7 descriptions of identical object could in fact be very different from each other when coming from different sources with the sole requirement of syntactical "MPEG-7 compliance", special care is needed so that these can be projected in a uniform vector space. While it would always be possible to "filter out" sources which do not comply directly to a simplified, pre-specified fixed mpeg-7 subset, it is clear that having a database as large as possible leads to more interesting results.

Recognizing the intrinsic difficulty of a full interoperability, works are currently undergoing [3] to standardize subsets of the base features as "profiles" for different specific applications, generically trading off generality and expressivity in favor of the ease and lightness of the implementation. Necessarily, this also means to give up on interesting scenarios especially as multimedia distribution and clients are more and more network applications (as opposed to standalone "cd player" like appliances), and expected to comply to a tolerant, decentralized (Semantic) Web philosophy and architecture [13].

2. Comparable works

While there are many works related to specific areas of audio classification and intelligent processing (among the most recent, [14][15]), just a few use exclusively MPEG-7 as base features and generically all address very specific use cases.

Semantic information from audio files is used to allow browsing of libraries in [16]. However the system appears to be fully centralized, requiring all the audio files to be available to the server as a prerequisite, and only partially relying on standard MPEG-7 features.

Also in [12] as well as in [17], MPEG-7 low level descriptors (LLD) are used but often associated with others that are not currently in the standard. The feature used as well as the database structure crafted to solve the specific problem (identification of music instrument by timbre and audio fingerprinting respectively) appear to perform well, but still requires the complete set of audio

files to extract metadata uniformly. No indication is also given that the optimized DB structure there used could address other use cases.

In [4] interoperability and manageability of MPEG-7 structures by means of a large number of existing XML databases is evaluated. The results show that the hierarchical structures spawned by the standard have so specific semantics and requirements that even the best systems appear of very limited use (if any, when LLD are involced).

Higher level interoperability is studied in [8] by the means of a mapping ontology expecially between space time segments and OWL (the semantic web ontology language) union/disjunction operator. While this in fact succeeds in respect to making the higher level mpeg-7 descritors interoperate on the Semantic Web, no mention is done on how mapping of features that are machine extractable.

Instead of concentrate of the specific details of mpeg-7 performance when applied to specific use cases (e.g. genre classifications, matching etc..) in this paper we focus on MPEG-7 low level interoperability and merging with the Semantic Web. Steps in this direction include:

- Usage of web standard Uniform Resource Identifiers (URIs) both to indentify and to annotate resources
- Tollerant, decentralized, web philosophy with no assumption on the availability of the actual sound source (direct extraction as a last resource).
- No assumption on the origin and format of an existing MPEG-7 descriptions. Will transparently adapt from different hopsizes, missing LLDs (by cross prediction) and different ways of representing data (E.G Spectral Envelope vs Basis + Projection, [1])
- Can transparently use different LLDs to improve the quality of each requested LLD with regards to common distortions caused by popular net distribution formats (e.g. MP3)
- Generates results which are properly interoperable on the semantic web by using both RDF and an appropriate OWL ontology.

The architecture presented here is therefore a framework which naturally allows to develop Semantic Web enabled applications using existing MPEG-7 algorithms. Chapter 7 shows an example of such an application.

2. Architectural overview

The proposed architecture (as currently implemented, the MPEG7DB project [7]) is depicted in Figure 1. URIs are both used as references to the audio files and as subject of the annotations produced in standard RDF/OWL format.

The MPEG7 Audio Compact Type (ACT) DB will fetch MPEG-7 descriptions of the selected URIs as described in chapter 3. The MPEG-7 ACT type used inside the DB will both be memory and computationally efficient (as opposed to the original XML structure) and abstract from the fine details of the LLDs (such as Hopsize and the many possible alternative of the scalar and vector series MPEG-7 descriptors [1]).

The projection block, described in chapter 4 will provide projections of the MPEG7ACT to metric spaces as needed by the specific application built upon this framework. In doing so, it will abstract from the specific structure of the original MPEG-7 description by using a recursive adaptation technique.

Once a set of uniform projections had been obtained for descriptions within the database, classic computational intelligence can be applied to fulfill the wanted task. Although fully functional, the computational intelligence blocks included in the framework (classificators and approximators including some based on Multi Layer Perceptrons Neural Networks) and used in the example application (chapter []), could be substituted with any vector based technique as needed.

Finally, once higher level results have been inferred (e.g , piece with uri "file://c:/my%20legal%20music/foo.mp3" belongs to genre "punk ballade") they can be saved into the provided semantic containers which will, hiding all the complexity, provide RDF annotations using terms given in an appropriate ontology (in OWL notation). Before outputting the annotation stream, the system will make sure that local URIs (e.g. "file://foo.mp3") are converted

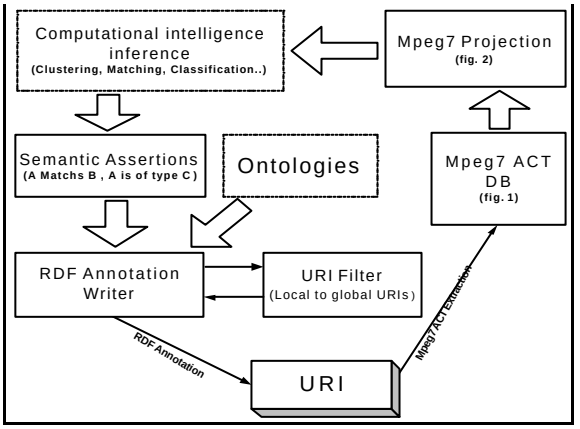


Figure 1 The overall structure of the proposed architecture. URIs are used to indicate which audiofiles are going to be considered in the database and are the subjects of the output annotations.

into globally meaningful formats like binary hash based URIs (e.g. hash "urn:md5: " , "ed2k:// " , etc.).

3. From URI to MPEG7-ACT

When the database component is given a URI indicating a new audio clip to index, it will first try to locate an appropriate MPEG-7 resource describing it. At this point it is possible to insert several alternative model of metadata research among which, calls to Web Services, queries on distributed P2P systems (P2P Exchange of metadata being a very promising field of study [18]) or lookup in a local storage or cache.

If a preliminary search fails to locate the MPEG-7 file, a similar mechanism will attempt to fetch the actual audio file (if the URI turns out to be a resolvable URL) and process it with the included, co-developed MPEG7ENC library[6]. Once retrieved, the schema valid MPEG-7 is parsed recursively so that the basic "stripes" of data belonging to Low Level Descriptors are mapped into flat, name indexed, array structures. These will not only serve as a convenient and compact container, but also provide abstraction from some of the basic feedom of description allowed by MPEG-7. Among these, the MPEG7 ACT type provides the basic time interpolation/integration capabilities to handle that LLDs could have many different sampling periods (even inside the same description block! See the "scaling" in [1]) and different grouping operators applied (e.g AudioWaveformType could be expressed as Raw values or as 2 grouped series of "Max" and "Min"). Currently, the MPEG7 ACTs are only indexed by the URI that they describe, but it is in this object that can fit specific indexes of interest to the final application, such as for example in [].

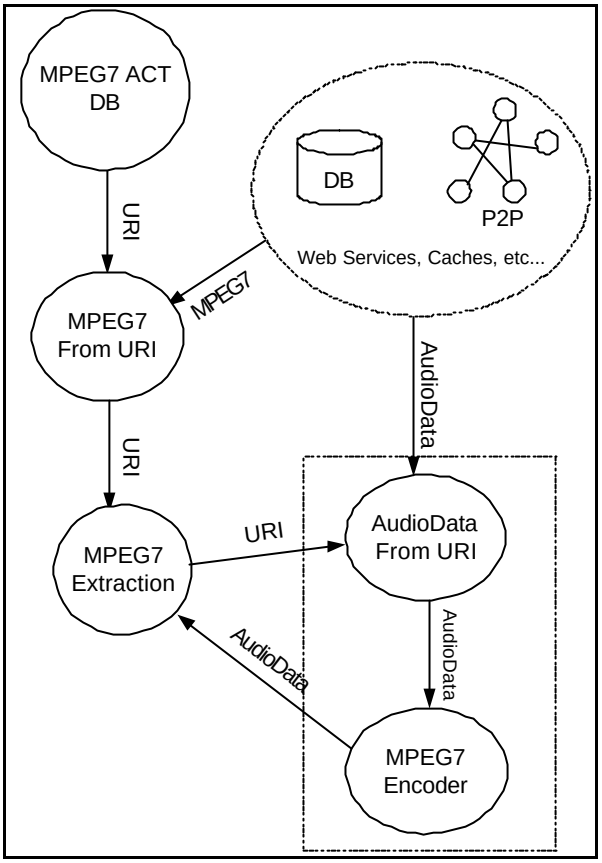


Figure 2 *The MPEG-7 Audio Compact Type database contains compact MPEG-7 representations associated with the URI of the original audio data. New audio segments are inserted simply by providing a corresponding URI. The database will first try to locate an existing MPEG-7 (either locally or querying remote sources) but it will also be able to fetch the original audio and encode it if needed.*

4. Projection into a metric space

To exploiting the benefits of computational intelligence (e.g. neural networks) and perform clustering, matching, comparisons and classifications each mpeg-7 resource will have to be projected to a single, fixed dimension vector in a consistent and mathematically justified way. In the next sections we discuss how we perform this task by the use of projection operators and input adaptation structures

4.1 Recursive adaptation

The following step is better understood as if driven by a "feature space request". A "feature space" deemed appropriate for the desired computational intelligence task will be composed of couples (one per dimension) of feature names and functions capable of projecting a series of scalars (or vectors) into a single scalar value. Among these, the frameworks provides a full set of classical statistical operators (mean, variance, higher data moments, median, percentiles etc..) that can be cascaded with others "pre processings" such as a derivative or a filter. Since MPEG-7 coming from different sources and processes, could have different available features and not necessarily those that we have selected as application "feature space", a recursive cross prediction of the missing ones takes place. This process is described in Figure 3.

For each of the required features, the system has a series of "registered providers". Each one of this is bound to provide the specified feature given a list of prerequisite features it will rely on. For each possible feature a "obvious feature provider" is present that will simply deliver the original data from the MPEG7ACT if this is available. Other registered providers will generically

contain cross prediction algorithms to estimate the requested feature from others which are found to be somehow related. Some cross feature estimation based on transformations and signal processing techniques currently implemented include:

- AudioTempoType DS from AudioPower DS and or from dfdfdf
- AudioPower DS from the Spectral Envelope DS
- SpectralEnvelope DS from the SpectralBasis+SpectralProjection DS (as specified in the MPEG-7 document)

It is also interesting to notice that, when a direct algorithm was not available, cross prediction based on neural networks proves to be, for a selected number of features, a viable alternative. In figure ThRISTO the we see the performance of a Multi Layer Perceptron (MLP) based on a single hidden layer and trained with data from different audio clips of different genre.

Each of the registered "feature provider" is also bound to provide an a priori estimation of extraction quality. By defition, the "obvious feature provider" will return 1 as quality while cross prediction techniques will return less. Interestingly, at this point, it is also possible to specify providers that will be able to provide a feature with a quality greater than 1. This is the case when the MPEG-7 has been extracted from lossy audio formats.

To measure the impact of lossy compression on the MPEG-7 LLD quality we define S/N ratio in dB as:

$$\text{S/N ratio} = \frac{1}{N}\sum_{x=0}^{N-1} 10\log_{10}\left(\frac{\sum_{n=1}^{r} x_n^2}{\sum_{n=1}^{r}(x_n - \bar{x}_n)^2}\right) \quad (1)$$

Where $\bar{x}_n$ is the MPEG-7 LLD extracted from the lossy source, x_n the same LLD from the clean (PCM) source, r is the dimension of the descriptor and N is the number of frames (e.g. 1 frame per 10ms hopsize as MPEG-7 specifies as default).

It has been shown in [19] that the noise so introduced is not negligible with S/N ratio, in case of 128kbps MP3 encoded music, in the area of 33dBs for AudioWaveForm and 39dBs for AudioPower.

To increase the S/N ratio we experimented with "feature providers" based on MLPs. A single hidden layer network was used, trained with four different LLDs derived from audio belonging to different categories and encoded in MP3 format. As target value, the neural network was given the input and the "clean" (PCM derived) version of one of them as targets. Gradient descent momentum technique and regularization by testing on a separated population was applied in the learning phase. Results are shown in table x and show an improvement of 1 to 3 dBs in the S/N ratio.

As longs as loops are prevented, there is no reason why a "feature provider" (other than the "objvious feature provider") should access directly the original MPEG7 act type while providing an estimate of a requested LLD.
The resulting recursive adaptation algorithm is represented in the top part of Figure 3. Lets put as an example that the feature AudioPower (AP) is requested but only AudioSpectrum Basis+Projections (ASB+ASP) are available in the original MPEG-7 stream. Since the "obvious feature provider" for AP will fail, the algorithm will try to resort to the next available adapter (ASE to AP) which then in turn will invoke the "ASB+ASP to ASE" feature provider. It is normal that the adaptation chain will necessarily introduce some form of uncertainty. To cope with this, each feature provider is requested to provide an estimate of its ability to reconstruct the requested feature given its own prerequisites. By recursive multiplication of the "qualities" provided by the plugins in the adaptation chain for each specific request we are able to determine to best trajectory into the adaptation cascade and ultimately provide a quality measure.
to its own ada. In Since the power can be estimated from the spectral envelope, the ASE to AP adapter will be invoked. To operate the adapter will request feature ASE from the algorithm again but since ASE is not actThe adapter will need to have the ASE find the next available

Altre feature other than mpeg7-

5. Once in a vector space

Once obtained, the projection vectors representing the mpeg-7 files in the db can be processed using a variety of well known techniques. The MPEG7DB project already provides tools such as neural networks, statistical classifiers and approximators, but vectors could also be readily exported for use in any external engine.

6. Reaching the Semantic Web with models and results

Once higher level semantic information has been extracted (such as the belonging of a particular piece of music to a particular genre), these are likely to be very informative and terse enough to be tractable with the current semantic web technologies (RDF reasoners and

alike). To allow this, we have created an ad hoc ontology [10] describing semantic concepts as “classification” , “feature sets” “machine inferred groupings”) and use it to produce an RDF stream describing what we have obtained and how. Among the interesting things , the ability to describe the “feature sets” themselves in terms of operators and mpeg7 features. This enables interoperability of the results not only as such (for example by annotating that a given mp3 belongs to a category that a user called “rock ballade”) but would also allows another user to import and run the same classification scheme remotely on his own data..

7. An experimental SW-MPEG7 application

All the work here presented has been implemented in Java (see [5] on why this is also computationally acceptable) and is available for review, suggestions and collaborative enhancement in the free software/open source model [7].

As an example application we used the described framework to classify the quality of speech content audio files. During the training phase, the database was divided into seven directories corresponding to different audio qualities, and the classifier (a MLP Neural Network) was trained with simply by assigning, as class label, the name of the directory containing the file.

To be able to write the results in an RDF format, we defined a simple ontology [] (using OWL language), describing concept like “generic class of objects ”,“belonging to a particular class” and “matching between objects”. Also if is possible to use a foundational ontology (CIC or Dublin Core) to describe concept like these, we decided to define a simple new one that matches closer our purposes. In fact we want to define every “semantic” relation between object (audio clips in this case) as a Closed World Inference, that is characterized by some attributes indicating the accuracy of every result and describing the set of audio clips from which every inference was deducted.

8. Conclusions

Even in this early experimentation phase the software system performs as expected and succeeds in bridging multimedia to the semantic web by “distilling” from distributed sources of verbose and heterogeneous audio representations concise and semantically relevant information along with the relative machine readable ontology. Further works will be about integrating in the feature space also aspects derived by the higher level descriptors such those about speech and melody contour or those as described in [9].

9. Acknowledgments

10. References

[1] ISO/IEC JTC1/SC29/WG11 N4031. MPEG-7 (2001)

[2] MPEG-7 XM http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html

[3] ISO/IEC JTC1/SC29/WG11 N5527, MPEG-7 Profiles under Consideration, March 2003, Pattaya, Thailand.

[4] “An analysis of XML database Solutions for the management of MPEG-7 media descriptions” Utz Westermann, Wolfgang Klas. ACM Computing Surveys (CSUR) Dec. 2003.

[5] “Java and Numerical Computing” Ronald F. Boisvert, Jose Moreira, Michael Philippsen, and Roldan Pozo. IEEE Com-puting in Science and Engineering March/April 2001

[6] MPEG7AUDIOENC – http://sf.net/projects/mpeg7audioenc , Holger Crysandt, Giovanni Tummarello

[7] MPEG7AUDIODB – http://sf.net/projects/mpeg7audiodb

[8] “Enhancing the semantic interoperability through a core ontology”, Jane Hunter. IEEE Transactions on circuits and systems for video technologies, special issue. Feb 2003.

[9] “Digital Media Knowledge Management with MPEG-7” Ralf Klamma, Marc Spaniol, Matthias Jarke. WWW2003, Budapest.

[10] http://semanticweb.deit.univpm.it/ontologies/CWMpeg7Inference.owl , Ontology for inferences on sets mainly aimed at computational intelligence algorithm

[11] ISO/IEC JTC1/SC29/WG11N5525 MPEG-7 Overview, Revision 9, March 2003

[12] “Application of Temporal Descriptors to Musical Instrument Sound Recognition, Journal of Intelligent Information Systems”, ALICJA A. WIECZORKOWSKA, JAKUB WR OBLEWSKI, PIOTR SYNAK 21 (1): 71-93, July 2003

[13] “Web Architecture from 50,000 feet” , revised 2002 Tim Berener Lee et Al. http://www.w3.org/DesignIssues/Architecture.html

[14] A hierarchical approach to automatic musical genre classification. [DAFX06], Juan José Burred and Alexander Lerch, 6th International Conference on Digital Audio Effects, London, september 2003

[15] “Applying Neural Network On Content Based Audio Classification”, Xi Shao,Changsheng Xu,Mohan S Kankanhalli, IEEE Pacific-Rim Conference On Multimedia (PCM03),Singapore,2003

[16] “The Cuidado Music Browser”, Pachet, F. Laburthe, A. Aucouturier, JJ., CBMI 03, Rennes (Fr).

[17] “Content based identification of Audio Material using MPEG-7 low level description” Proceedings of the Int. Symp of music information retrival 2001. Allamanche, E. Herre, J . Helmuth, O. Frba, B. Kasten, T and Cremer, M.

[18] http://p2p.semanticweb.org related to p2p and the exchange of metadata

[19] “An Examination of practical information manipulation using the MPEG-7 low level Audio Descriptors” J. Lukasiak, D. Stirling, M.A. Jackson, N. Harders. 1st Workshop on the Internet, Telecommunications and Signal Processing

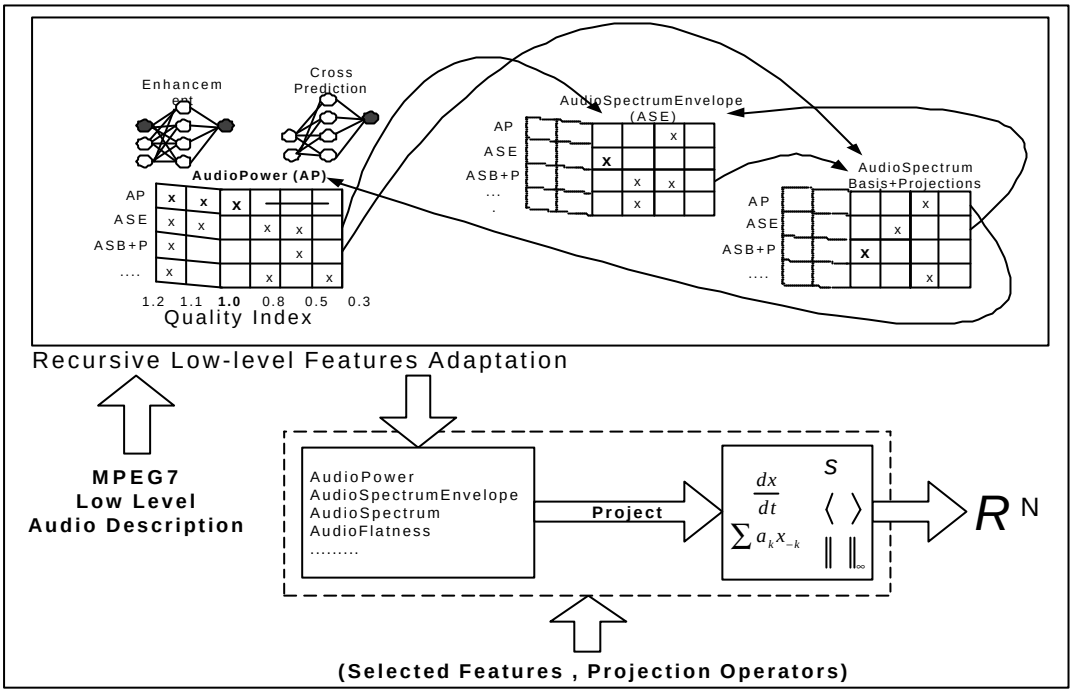
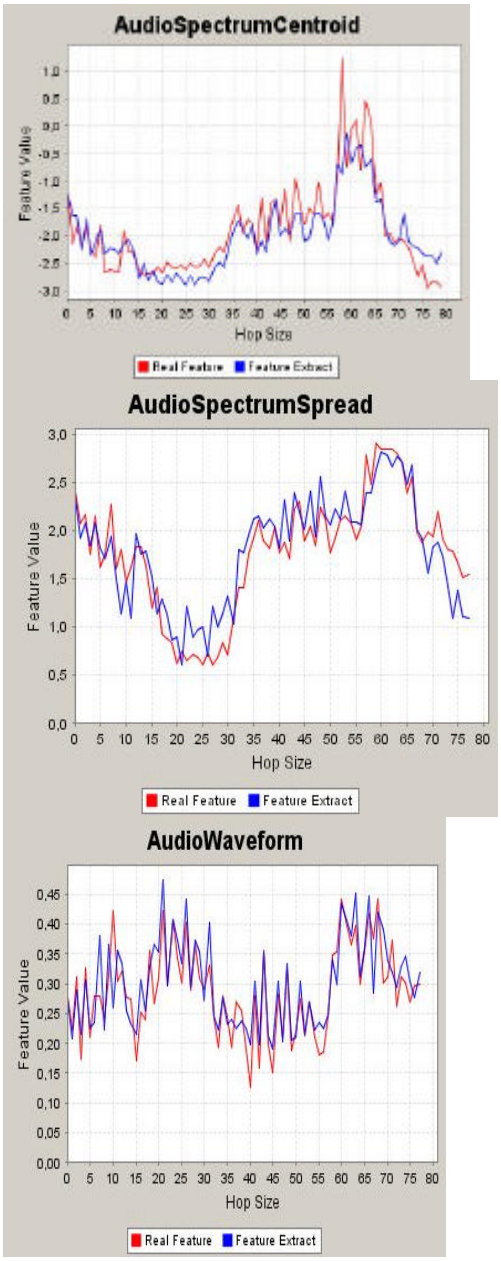


Figure 3 A feature space is specified in terms of selected features (e.g. “Audio power”, “Spectral spread”, etc..) and Projection Operators (e.g. “max”, “mean” , possibly after filtering or differentiation) and MPEG7 act data to be projected is processed through a recursive system both capable of cross predicting a missing feature and providing improved version of an existing one (if the sources are encoded with lossy formats).



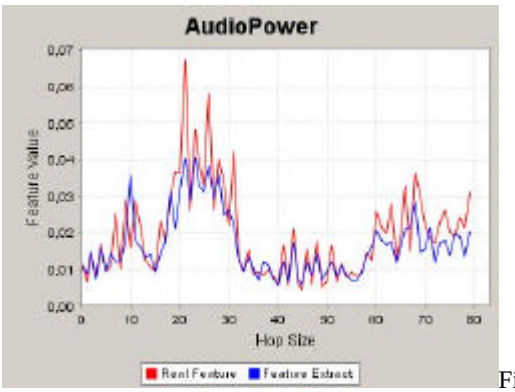


Figure 4 Cross estimation of MPEG-7 descriptors by means of Multi Layer Perceptron. For each feature the three others here represented are given as source of estimation.