Overview of MPEG-7 Audio

Schuyler Quackenbush, *Member, IEEE*, and Adam Lindsay, *Associate Member, IEEE*

***Abstract*—MPEG-7 is a new ISO standard that facilitates searching for media content much as current text-based search engines ease retrieval of HTML content. This paper gives an overview of the MPEG-7 audio standard, in terms of the applications it might support, its structure, the process by which it was developed, and its specific descriptors and description schemes.**

***Index Terms*—Audio, meta-data, MPEG, search and retrieval.**

I. Introduction

THE Internet, as a global network, can be viewed as a giant depository of data that can be searched, sorted and sifted in order to extract the desired information. The emergence of hypertext markup language (HTML), hypertext transfer protocol (HTTP), web browsers, and finally search engines have made this wealth of information accessible to anyone with a desktop PC.

However, nontext-based documents, such as MPEG-1 [1], MPEG-2 [2], [3], or MPEG-4 [4], [5] compressed audio media files are not well served by these languages, search engines, or search appliances. First, audio content is not expressed in a character format, making textual search impossible. Second, consider how one typically listens to audio content: sometimes via a home stereo system, but much more often via a portable device, such as a radio or portable compact disc player. In either case, the means of accessing audio content is via some few buttons and perhaps a small display, not a PC keyboard and display. What is needed is standardized means of representing data about media that also supports a number of compact search modalities.

During its beginnings in 1996, MPEG-7 [6] was motivated in part by MPEG's past successes in making the efficient transmission of multimedia content feasible, and by the explosion of networked resources that were making access to these resources a reality. Whereas textual search engines indexed web pages on the Internet using well-known information retrieval techniques, no such standard techniques were available for audio files. There is clearly a need to index network-based media, be they compressed or PCM. However, new methods of indexing should also serve physical archives of media, even including a home music collection.

MPEG-7 attempts to address this need. It is a standard for content-based media description. It is independent of the coding format of the media (i.e., is not limited to MPEG-compressed formats, or even digital formats) and is independent of the physical location of the media. The breadth of representation in MPEG-7 supports not only simple, text-based queries (e.g., find all songs by artist A), but also complex content-based queries (e.g., find all songs that sound like the melody I am humming) [7]. We also note that once this meta-data is available, many more applications than just search will be available to the user.

This paper gives an overview of the emerging MPEG-7 standard for audio meta-data, ISO/IEC 15938-4 Information Technology—Multimedia Content Description Interface—Part 4: Audio. At the time of writing, this work is at the Committee Draft (CD) stage, and is expected to be an International Standard (IS) in September 2001. However, for simplicity, this work in progress will be referred to as the MPEG-7 audio standard. Section II outlines the structure of the standard, and the relationship between the audio part and the rest of the standard. Section III gives some examples what audio applications that are based on the MPEG-7 standard might look like. Section IV describes the process used to develop the audio portion of the MPEG-7 standard. Finally, Section V gives an overview of the MPEG-7 audio Descriptors and Description Schemes.

II. Structure of the Standard

MPEG-7 standardizes a representation of meta-data associated with media content. Unlike previous MPEG standards, MPEG-7 does not standardize an encoder or extraction process, or a decoder or a query or matching process. What is common amongst all MPEG standards is the standardization of a medium of exchange: normative representations and semantics of meta-data in the case of MPEG-7, and bit streams in the case of MPEG-1 or MPEG-2.

The MPEG-7 audio standard is composed of *Descriptors* and *Description Schemes*. Descriptors are instantiations of meta-data that may be associated with a single temporal interval (perhaps the entire signal) in an audio media signal, or with a (perhaps periodic) set of intervals in the media signal. Another way of looking at it is that a Descriptor is a semantic atom—a single unit of description regarding a feature of content. A spectral envelope is an example of an audio Descriptor.

A Description Scheme is an association of a set of Descriptors and their relationship that addresses a particular task or application. A Description Scheme is a tree structure of Descriptors and possibly other Description Schemes.

The MPEG-7 audio standard provides Description Schemes as fundamental constructs and as parts of application-driven tools. In the understanding that the standardized set of tools will not fit every application, however, there is a Description Definition Language (DDL) that allows for new Description Schemes to be written for specific applications. This is the mechanism by which the standard may be extended.

For each Descriptor and Description Scheme defined in the audio standard, there is a normative definition, expressed in the DDL, an equivalent expression as a C++ class and text describing extraction and use of that component. In addition, there

Manuscript received September 2000; revised March 2, 2001. The work of A. Lindsay was supported in part by the U.K. EPSRC/DTI project RETRIEVE.

S. Quackenbush is with AT&T Labs, Florham Park, NJ 07932 USA (e-mail: srq@research.att.com).

A. Lindsay is with the Department of Computing, the University of Lancaster, Lancaster LA1 4YW, U.K. (e-mail: atl@comp.lancs.ac.uk).

Publisher Item Identifier S 1051-8215(01)04990-4.

1051–8215/01$10.00 © 2001 IEEE

is informative source code embodying both an exemplary extraction method and an exemplary query and matching method, both of which may operate on the C++ class associated with that Descriptor or Description Scheme.

III. Applications

One of the best ways to understand the standard is through example applications.

A. Query by Humming

Consider a streaming audio service in which there is a database of compressed MPEG-4 audio on one (or more) MPEG-4 media servers and an associated database of MPEG-7 meta-data on a MPEG-7 query server. For each MPEG-4 compressed song indexed, the MPEG-7 database stores a complete representation of the song's melody and some mechanism to link the meta-data to the associated media, for example a URL. The database could also store the song title, artist, song length, musical genre and style, and musical "mood."

The user presses the query button on his wireless hand-held access device and hums a tune. The sampled signal waveform is transmitted to the query server, where a process is run that extracts its MPEG-7 melody meta-data. That meta-data is the query target, and the MPEG-7 server searches the MPEG-7 melody database for matches. The top few matches are transmitted back to the access device and displayed as song title and artist. The user can hit a button to immediately start streaming from the MPEG-4 database to the wireless access device, where the media is decoded and played.

The user could follow-up the initial query with the more powerful query (for example via button and menu selection) to play more songs "by that artist," or "in the same genre and style" as that song, or just "with the same mood" as that song.

B. Query for Spoken Content

Consider a telephone message service with associated MPEG-7 meta-data. For each message recorded in a person's message queue, a speech recognition engine processes the voicemail and puts the corresponding text annotation into the MPEG-7 database. The database also stores calling number (from which some estimation of the caller can be obtained), time of day of call and length of call.

The user is anxiously expecting to receive a specific call. He could call up the message service, where the recognition engine could process his voice-based query to form a textual MPEG-7 query. Where the recognition engine fails to convert to text (for example, due to a proper name being outside of the dictionary), it may use a phonemic representation. The resident MPEG-7 query server could then match the keywords in the query with the words in the message annotation and also with the MPEG-7 fields representing the caller. For example: "Get me all the messages from my boss about the Tallahassee account!"

C. Assisted Consumer-Level Audio Editing

Imagine that an audio signal has been processed to extract meta-data to facilitate interactive applications. For example, many low-level audio features such as power, loudness, spectral centroid, and spectral envelope could be used in a hypothetical consumer-level audio editor. One could use these Descriptors for display, such as loudness, in the place of the waveform. With some knowledge of standard editing rules within a genre, a sophisticated program could use spectral features to equalize channels and thereby automatically enhance the clarity of a mix.

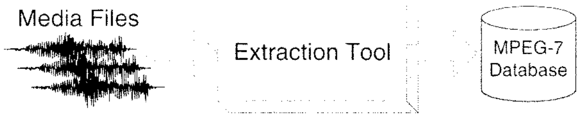

Fig. 1. MPEG-7 extraction process.

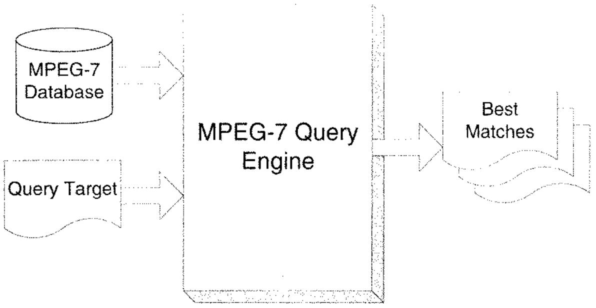

Fig. 2. MPEG-7 query matching pocess.

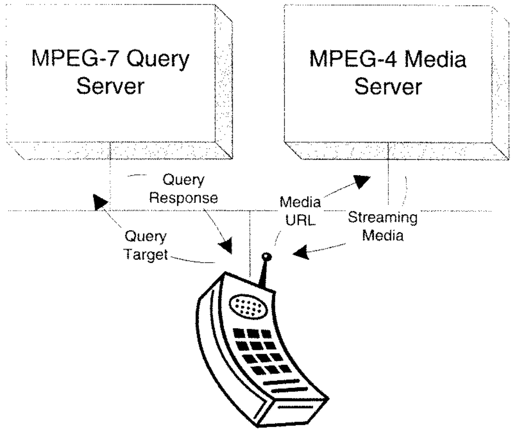

Fig. 3. MPEG-7 client server architecture.

D. Extraction and Query Paradigm

The first two examples shared a common architecture. The first step in developing the application is to extract the MPEG-7 meta-data from a set of media files, as is shown in Fig. 1. This could be done once for a largely static set of media files (such as in the query by humming example), or done repeatedly (as in the phone message example).

The next step is the query process, as shown in Fig. 2. In this process, the query is converted to an MPEG-7 query target, typically by the same extraction engine, although that is not shown. The query engine then compares the query target with all the MPEG-7 instantiations in the database via a suitable distance function and returns the best match(es). Media-locator information that is part of the meta-data can then be used to locate the desired media data.

Fig. 3 shows how this might be combined to form a client server MPEG-7 application, just as was described in the query by humming application.

IV. Core Experiment Process

In order to ensure a fair and orderly development of the standard, a rather strict procedure was developed to guide the evaluation of proposed technology and promotion of that technology into the standard. The process is a “Core Experiment,” as this experimental process constitutes the core of the MPEG-7 development work.

To include a new Descriptor into the MPEG-7 audio standard, the following three steps had to be completed:

1) proposal;
2) challenge;
3) integration.

Each will be described in turn.

A. Proposal

In the proposal stage, internal verification and prescreening has to be done by the proponent. A proposal must contain:

1) information on whether the new Descriptor is planned to be an addition to the standard or if it is planned to replace or modify an existing Descriptor in the standard;
2) a technical description that is accurate enough that someone skilled in the art can duplicate the main function of the Descriptor;
3) the eXperimentation Model (XM, or the developing reference software) framework changes that would be necessary to implement the proposal;
4) text for the normative (and possibly informative) parts of the standards document;
5) evidence for the merit of the Descriptor (e.g., potential market, studies of user preference, examples of how it might be used in products or services). The scope of the technology will be taken into account.

B. Challenge

In the challenge phase, there is testing and reviewing of the technique by other parties. A proposal must be independently evaluated by at least one third party. The objective of choosing a challenge format is to expose the shortcomings of each proposed technology. Although it is assumed that proposals will work well in the vast majority of cases, the challenge is to identify the degenerate cases in which the proposed technology does not work, and to choose which “description space” is more complete and/or desirable. The proponent may be part of the test and review process.

Standard figures of merit will be allowed and are encouraged where they are appropriate. Where a standard method is not available, the “stimulus triplet” method is proposed as a simple means that is sufficient to determine the utility of MPEG-7 technology. The stimulus triplet is a set of three pieces of content, denoted A, B, and C. Within the stated scope of the Descriptor, A is perceptually (as ranked by a human listener) more similar to B than to C. However, if A is computationally (as computed based on the challenged Descriptor values) more similar to C than to B, then such a triplet would successfully expose a weakness in the challenged Descriptor. Therefore, by successively mapping out the weaknesses in the feature-spaces of competing proposals, the parties collaborate to produce evidence for relative merit.

C. Integration Phase

After the acceptance of the core experiment results, source code implementing extraction and matching aspects of the new Descriptor must be submitted. The extraction code or executables performing extraction are then integrated into the XM. The matching code must be integrated into the XM so that a full query can be posed. The normative expression (i.e., the DDL representation and any C++ code) of the core experiment technology, along with descriptive text, is placed in the standards document.

V. Overview of Audio Descriptors

The MPEG-7 audio standard comprises six main technologies that can be divided roughly into two classes: low-level or generic tools, and application-specific tools. The audio description framework (which includes the Scale Tree and low-level Descriptors) and the uniform silence Segment may apply to any audio signal, and therefore are in the class of generic tools. The sound effect description tools, instrumental timbre description tools, spoken content description, and melodic Descriptors restrict their application domain as a means to afford more descriptive power, and so are in the class of application-specific tools.

A. MPEG-7 Audio Description Framework

Several technologies combine to form the low-level Audio description framework. One of the foundations is the Scale Tree, which allows (generally temporal) series of Descriptors to be represented in a scalable way. The basic idea is to enhance a traditional time-series representation of a feature with the ability to extract a meaningful summary (by decimation or statistics) that can be parameterized to fit within any available volume of storage. The basic data type that forms the series can be a scalar, vector, or matrix.

The other method of aggregating data for the purpose of description draws heavily (by means of inheritance) from the generic multimedia Description Scheme called “Segment.” For the purpose of MPEG-7 audio, an audio Segment is a temporal interval to which all attached Descriptors apply. A Segment is a recursive structure, so an audio stream may be hierarchically decomposed [8]. All audio Descriptors and Description Schemes, including all of the application-based tools and the Scale Tree, may be attached to a given audio segment.

The low-level Descriptors that fit into this foundation composed by the Audio Segment and Scale Tree include temporal envelope, spectral envelope, harmonicity, spectral centroid, and fundamental frequency.

B. Silence Segment

A very simple but useful tool is the MPEG-7 silence Descriptor. It attaches the simple semantic of “silence” (i.e., no significant sound) to an audio Segment. It also includes a simple indicator of the level of silence via the minimal temporal threshold. It may be used to aid further segmentation of the audio stream, or as a hint not to process a Segment.

C. Sound Effects Description Tools

The sound effects Descriptors and Description Schemes are a collection of tools for indexing and categorization of general sound effects. Support for automatic sound effect identification and indexing is included as well as tools for specifying a taxonomy of sound classes and tools for specifying an ontology of sound recognizers. Such recognizers may be used to automatically index and segment sound tracks.

A more computational retrieval representation is also available within the sound effect tools. Although, to many who wish simply to identify sound effects, the spectral basis decomposition may seem like an unnecessary intermediate result. However, given a properly trained set of sound effect parameters, the tools allow for sound-effect recognition in sound mixtures. The spectral basis is essentially a temporal envelope crossed with a spectral envelope, with several of these bases typically being combined to form a model of a sound.

D. Musical Instrument Timbre Description Tools

Timbre Descriptors aim at describing perceptual features of instrument sounds. Timbre is currently defined in the literature [9] as the perceptual features that make two sounds having the same pitch and loudness sound different. The Timbre Description Scheme describes these perceptual features with a reduced set of Descriptors. The Descriptors relate to notions such as "attack," "brightness," or "richness" of a sound.

There are four classes of musical instrument sounds: 1) harmonic, sustained, coherent sounds; 2) nonharmonic, sustained, coherent sounds; 3) percussive, nonsustained sounds; and 4) noncoherent, sustained sounds. Within these four classes, two classes are well-detailed within the MPEG-7 standard, and have been the subject of Core Experiment development. They are the harmonic, coherent, sustained sounds and nonsustained, percussive sounds. The other two classes were deemed to be of lower priority due to their relative rarity, but may still make an appearance in the standard.

E. Spoken Content Description Tools

The Spoken Content Description Scheme is predicated on the idea that speech recognition systems are currently imperfect. Rather than being a simple textual transcript of what is spoken (although it can accommodate such a thing), the Description Scheme consists of combined word and phone lattices for each speaker in an audio stream. By combining the lattices, the problem of out-of-vocabulary words is greatly alleviated and retrieval may still be carried out when the original decoding was in error. The Description Scheme can be used for two broad classes of retrieval scenario: indexing into and retrieval of an audio stream, and indexing of multimedia objects annotated with speech.

F. Melody Contour Description Scheme

The Melody Contour Description Scheme is a compact representation for melodic information, which allows for efficient and robust melodic similarity matching, for example, in query-by-humming. The Melody Contour Description Scheme uses a five-step contour (representing the scale-step interval difference between adjacent notes), in which intervals are quantized. The Melody Contour Description Scheme also represents basic rhythmic information by storing the number of the nearest whole-beat of each note, information that can dramatically increase the accuracy of query matches.

There is also a more expanded version of the Melody Description Scheme being developed collaboratively, which trades some of the terseness of the Melody Contour Description Scheme for greater precision and robustness.

G. Other Parts of the Standard

The Audio part of the standard was designed to work well with the other parts of MPEG-7, especially Part 5-MDS. To avoid redundant functionality within the standard, the audio sub-group relies on the efforts of other groups for completing the set of tools that an audio archivist might need. We have already discussed the use of the DDL as a standard way of extending the standard. We have also seen how the Audio Framework utilizes the general segment from the MDS.

Anyone describing content may wish to attach typical Descriptors such as title, composer, and year of recording to a piece of music. These Descriptors, which are in current practice in libraries and archives, are covered in detail in the MDS. Similarly, one may try to describe musical genre as a hierarchical ontology, or describe musical instrument from a list of controlled terms. The mechanisms by which one can create these ontologies and dictionaries are in the MDS as the Controlled Term datatype and the Classification Scheme Description Scheme. By putting all of these parts together, one can transfer an existing archive to a unified framework under the MPEG-7 standard, and then enhance it further with specialized, signal-processing derived audio Description Schemes.

VI. Summary

MPEG-7 addresses the critical need for media indexing and search tools, and several example applications that would build upon MPEG-7 media metadata have been described. We have given an overview of the process by which the MPEG-7 Audio standard was developed, and of the Descriptors and Description Schemes in the MPEG-7 Audio standard. These already support a rich application space; however, the DDL permits the set of Description Schemes to be extended, thus ensuring relevance in applications that are yet to be envisioned.

The MPEG-7 standard provides a framework for describing audio signals and audio archives using not only library practices from the 19th and 20th Centuries, but also using new, signal-based criteria that can be automatically extracted at any desired temporal (or possibly spectral) resolution. In doing so, MPEG-7 provides a new description technology for the 21st century.

References

[1] *Information Technology—Coding of Audio-Visual Objects—Part 3: Audio*, ISO/IEC 111172-3, 1993.

[2] *Information Technology—Coding of Audio-Visual Objects—Part 3: Audio*, ISO/IEC 13818-3, 1995.

[3] *Information Technology—Coding of Audio-Visual Objects—Part 7: Advanced Audio Coding*, ISO/IEC 13818-7, 1997.

[4] *Information Technology—Coding of Audio-Visual Objects—Part 3: Audio*, ISO/IEC 14496-3, 1999.

[5] *Information Technology—Coding of Audio-Visual Objects—Part 3: Audio*, ISO/IEC 14496-3:1999/AMD1, 2000.

[6] *Information Technology—Multimedia Content Description Interface—Part 4: Audio*, ISO/IEC CD 15938-4, 2001.

[7] A. T. Lindsay and W. Kriechbaum, "There's more than one way to hear it: Multiple representations of music in MPEG-7," *J. New Music Res.*, vol. 28, no. 4, pp. 364–372, 1999.

[8] A. T. Lindsay, S. Srinivasan, J. P. A. Charlesworth, P. N. Garner, and W. Kriechbaum, "Representation and linking mechanisms for audio in MPEG-7," *Signal Processing: Image Commun.*, vol. 16, pp. 193–209, 2000.

[9] American National Standards Institute, *Psychoacoustical Terminology*. New York: American National Standards Institute, 1973, vol. S3.20, p. 56.

Schuyler Quackenbush (S'71–M'75) received the B.S. degree from Princeton University, Princeton, NJ, in 1975, the M.S. degree in 1980, and the Ph.D. degree in 1985, both in electrical engineering from Georgia Institute of Technology, Atlanta.

He spent four years in industry as a Design Engineer prior to his graduate tudies. For the latter half of 1985, he was a Staff Research Associate at Georgia Tech. He joined AT&T Bell Laboratories, Murray Hill, NJ, in 1986 as Member of Technical Staff in the Digital Signal Processing Research Department. In 1996, he joined the Speech and Audio Research Department, AT&T Laboratories, Florham Park, NJ, as Principal Technical Staff Member. His research interests are in speech and audio coding algorithms, real-time implementation of signal processing algorithms, and signal processing hardware. He is the author of more than 30 publications in these areas, including one book, *Objective Measures of Speech Quality* (Englewood Cliffs, NJ: Prentice-Hall, 1998). He is active in the area of standardization of audio coding algorithms, was one of the authors of the ISO/IEC MPEG Advanced Audio Coding standard, and is currently the chair of the ISO/MPEG Audio subgroup.

Dr. Quackenbush is the Chair of the IEEE Technical Committee on Audio and Electroacoustics and was the General Chair of the 1995 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is also a member of the AES.

Adam Lindsay (A'00) received the B.S. degrees in cognitive science and music in 1994, and the M.S. degree in media arts and sciences in 1996, all from Massachusetts Institute of Technology, Cambridge.

He is a member of Research Staff in the Department of Computing at Lancaster University, Lancaster, U.K. He was previously the Principal Investigator in multimedia representation for the Belgian research company Starlab. During 1996, he was one of the charter researchers for Riverland Research, Brussels, Belgium. Following his interests in the description of audio-visual material, he has emerged to be one of the leaders in MPEG-7 standardization, focusing on applications, audio, systems, and philosophy.