AN APPROACH TO THE SEMANTIC MODELING OF AUDIO DATABASES

Mustafa Sert
Başkent University
Department of Computer Engineering
06530 Ankara, TURKEY

Buyurman Baykal
Middle East Technical University
Department of Electrical and Electronics Engineering
06531 Ankara, TURKEY

Keywords: Audio databases, audio modeling, content-based retrieval, MPEG-7

Abstract: The modeling of multimedia databases for multimedia information systems is a complicated task. The designer has to model the structure and the dynamic behavior of multimedia objects, as well as the interactions between them. In this paper, we present a data model for audio database applications in the context of MPEG-7. The model is based on the object-oriented paradigm and as well as low-level and high-level signal features, which are standardized within the MPEG-7 framework, thus enabling interoperability of data resources. The model consists of two parts: a structural model, which provides a structural view of raw audio data, and an interpretation model, which allows semantic labels to be associated with audio data. We make use of an object-oriented approach to capture the audio events and objects in our model. Compared to similar models, particular attention is paid to integration issues of the model with commercial database management systems. Temporal relations between audio objects and events are also considered in this study.

1 INTRODUCTION

With the advances in information technology, the amount of multimedia data captured, produced and stored is increasing rapidly. In addition to this, multimedia content is widely used in many applications in today's world, and hence, a need for organizing this data and accessing it from repositories with vast amount of information has been appeared. In addition to video and image, audio content-based retrieval, content analysis and classification have a wide range of applications in the entertainment industry, audio archive management, commercial musical usage, and surveillance. As stated in (Petkovic and Jonker, 2000), the fast increase in the amount of auditory data caused audio to draw more attention as a multimedia data type and revealed an important problem; hence new methods should be developed to manage them because existing data management techniques do not provide sufficient support for audio data type. However as pointed out in (Ghafoor, 1994), one of the challenging issues that the researchers have to encounter is the development of such models which capture the characteristics of that data type.

Audio content-based retrieval requires many adjustments in a multimedia database management system compared to a traditional database management system. Traditional database management systems are not suitable to manage auditory data since audio data has its own characteristics, which differentiate it from simple textual or numerical data. Traditional database management systems must be enhanced with new capabilities to handle audio data, as a data type to query. The first step in achieving this goal can be to create a multimedia data model and incorporate it into the existing database architectures. As stated in (Grosky, 1997), a multimedia data model has different properties relative to a traditional data model. Such models should be able to capture and represent various types of information about multimedia objects, their structures, operations and properties, as well as real-world objects and relationships among them. The defined model can then be used for retrieval and querying of audio with the extracted information.

This paper proposes a data model in the mentioned fields of a multimedia database management system in the context of audio databases, and organized as

Sert M. and Baykal B. (2004).
AN APPROACH TO THE SEMANTIC MODELING OF AUDIO DATABASES.
In *Proceedings of the First International Conference on E-Business and Telecommunication Networks*, pages 385-390
DOI: 10.5220/0001401503850390
Copyright © SciTePress

follows. In Section 2, related approaches are explored. Our model is explained in Section 3. In Section 4, integration issues of the model with commercial database management systems are discussed. The following section gives some practical examples to provide an understanding of possible applications of the model. Finally, our conclusions with the further issues are presented.

2 RELATED WORK

Content-based retrieval of multimedia data has been explored in several studies. Early attempts have addressed the problem of retrieval of images. Afterwards, the problem of video retrieval has attracted much more attention. Audio, however, is generally studied with regard to video retrieval, and not much has been done on this issue. As stated in (Gudivada and Raghavan, 1995), we can broadly classify the various approaches into three categories: keyword based, feature based, and concept based approaches. In keyword based approaches, which is the simplest way to model the content, is by using free text manual annotation. In feature based approaches, a set of features are extracted from the multimedia data, and represented in a suitable form. In the latter case, application domain knowledge is used to interpret an object's content and may require user intervention.

Systems in the first category are mostly based on textual data, hence a traditional database management system which provides support for object retrieval is adequate for this purpose. In the latter categories, further considerations are needed in the context of modeling in multimedia database systems.

In the literature, there are a few works on content-based retrieval systems for auditory data both commercially and academically. However, most of these systems support only segmentation and classification of audio data, that is, the signal processing aspects of the audio. As query languages are very dependent on the underlying data model, our survey will take account of some of the multimedia query languages.

One specific technique in content-based audio retrieval is query-by-humming. The approach in (A. Ghias, 1995) defined the sequence of relative differences in the pitch to represent the melody contour and adopted the string matching method to search similar songs.

In the content-based retrieval (CBR) work of the Musclefish Company (E. Wold, 1996), they took statistical values (including means, variances, and autocorrelations) of several time and frequency-domain measurements to represent perceptual features like loudness, brightness, bandwidth, and pitch. As merely statistical values are used, this method is only suitable for sounds with a single timbre.

A music and audio retrieval system was proposed in (Foote, 1997), where the Mel-frequency coefficients were taken as features, and a tree-structured classifier was built for retrieval.

In (A. Woudstra, 1998), an architecture for modeling and retrieving audiovisual information is proposed. The proposed system presents a general framework for modeling multimedia information and discusses the application of that framework to the specific area of soccer video clips.

In (L. Lu, 2003) an SVM-based approach to content-based classification and segmentation of audio streams is presented for audio/video analysis. In this approach, an audio clip is classified into one of the five classes: pure speech, non-pure speech, music, environment sound, and silence. However, there is no underlying database model for content-based audio retrieval in the system.

(J.Z. Li, 1997) describes a general multimedia query language, called MOQL, based on ODGMs' Object Query Language (OQL). Their approach is to extend the current standard query language, OQL, to facilitate the incorporation of MOQL into existing object-oriented database management systems. However, as stated in (J.Z. Li, 1997), further work needs to be done to investigate the support for audio media and to establish the expressiveness of MOQL.

There are other audio data models, which are explored in the context of video (G. Amato, 1998), (A. Hampapur, 1997), (R. Weiss, 1994). However, since the main purpose is video, less attention is paid on the audio component.

The main contribution of this work lies on the following. We mainly emphasized on the audio component. Particular attention is given to the integration issues of the model with commercial database management systems, and finally, we believe that the interoperability of the model is enabled by utilizing the signal features, which have been standardized in MPEG-7 framework.

3 AUDIO DATA MODEL

In this section, we present our audio data model, its components and some details on representation of audio data. As identified in the work on MPEG-7 (John R. Smith, 2000), audio-visual content can be described at many levels such as structure, semantics, features and meta-data. At this stage, MPEG-7 takes place by standardizing a core set of descriptors and description schemes to enable indexing, retrieval of audio-visual data, and interoperability of the data resources (MPEG-7, 1999). A descriptor (D) is used to represent a feature that characterizes the audio-visual

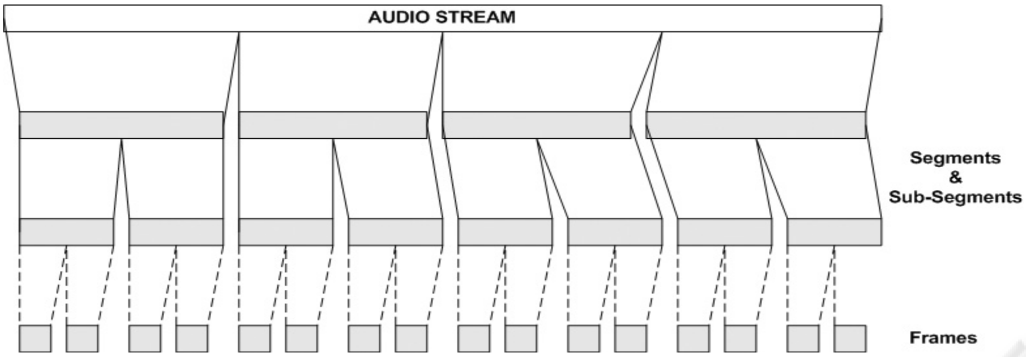

Figure 1: Segmentation of an audio for Audio DS.

content, while a descriptor scheme (DS) is used to specify the structure and semantics of a relation between its components, such as descriptors and description schemes. Descriptors in MPEG-7 deal with low-level features of a multimedia data (e.g., audio, video), such as color, motion, audio energy, and so forth. On the other hand, descriptor schemes deal with high level features, such as semantic description of objects and events. In this context, by separating the distinct tasks of conceptual, logical and physical modeling as in database design, we separate the content description process into two levels, namely structural and interpretation.

3.1 Modeling the Structure and Concept

At the lowest level of representation, an audio data is an unstructured piece of information as a sequence of sample values (raw object) that can be represented in the time domain or the frequency domain. Different features can be extracted from these two representations, however this is not our goal in this study. Interested readers are referred to (MPEG-7, 2001) for low and high level features of audio data. A raw object contains a large amount of significant information and should be managed by using an explicit representation. Our model stands for this purpose, and includes hierarchical structures, as well as object-oriented methodologies for identifying the possible conceptual entries in a raw object.

Our data model consists of two parts: A structural model, which provides a structural view of raw audio data, and an interpretation model, which allows semantic labels to be associated with audio data. Structural and semantic information of an audio are described by MPEG-7 meta-data in order to enable indexing and retrieval of an audio data. Structural modeling includes 17 low-level MPEG-7 features (e.g., AudioFundamentalFrequencyType, AudioWaveformType, AudioPowerType) to describe an audio. Semantic modeling consist of identifying audio entities (objects) and their relations. We make use of an object-oriented approach to capture audio events and objects in an audio. We have defined an *audio object* as a sound source, and that any kind of behavior of that object is an *event*. Events develop in time by an object and also have a duration property. As the main function of an object is describing sound sources, it is possible to distinguish different levels for describing audio objects. Some generic source objects can be a musical instrument, speech voice/owner, environmental sound, and sound effects. Similarly, as event is the temporal behavior of some audio object along or around a certain time; crying, shouting, dialogs between persons, and a musical note can be considered as events. As a consequence, these two components together are very useful for querying at the semantic level.

In our model, we have extended the idea which is presented in proposal (P. Salembier, 1999). What we present here is a first step towards that. We are utilizing a *frame-based* view at the bottom level of representation instead of a *scene-oriented* view. In addition, we have classified an audio into some common types such as *speech* (sp), *music* (mu), *sound* (so), and *combination of all-mixed* (mi). These classes are immediate descendants of Audio DS with the *is a* relation type. For instance, *a music is an audio, a speech is an audio,* and so forth. An Audio TOC is constructed for all these types, in other words, every component of an audio has an Audio TOC which also shows the *aggregation* relation. Class *mixed* is defined to handle the audio pieces which are not classified into other three classes due to signal characteristics. The inclusion of these classes to the model is important for several reasons such as different audio types have different significance to different applications, the audio type or class information itself may be very useful for some applications, and the search space after classification is reduced to a particular audio class during the retrieval process.

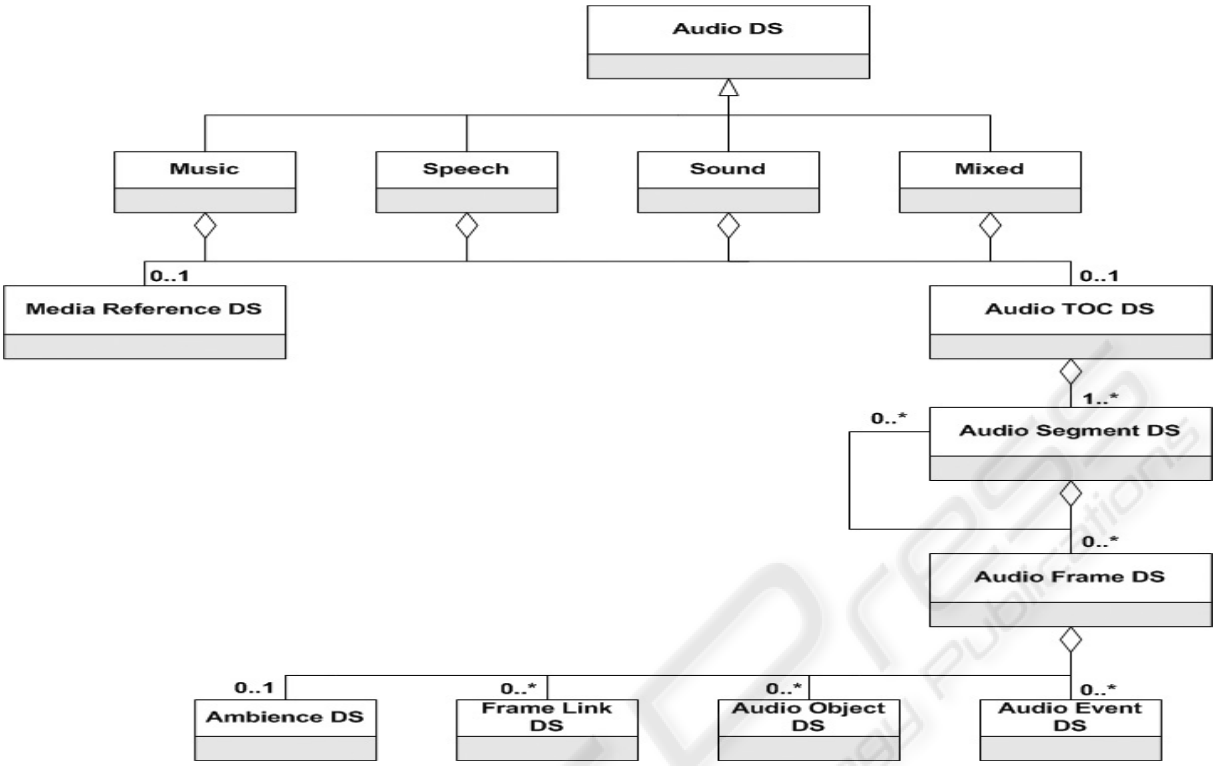

Figure 2: UML diagram of the overall model (not exhaustive).

The overall structure of the generic audio DS is constructed as follows. An audio entity is progressively partitioned to temporal segments and subsegments (Fig. 1). This process is continued until the segments cannot be sub-segmented any further. At this stage, we call the bottom-level elements as frames.

An audio segment may contain any number of descendant segments and frames, which both have their own begin and end times. As stated in (Adam T. Lindsay, 2000), this abstraction provides a view called audio table of contents (TOC), which is very similar to a table of contents in a book. The overall view of the model (but not exhaustively), in the context of MPEG-7 description tools is presented in Fig. 2.

Audio DS provides a general framework for the description of audio and is composed of labels and descriptors to identify the audio to be described.

The *Media Reference DS* holds two descriptors concerning with the media, one for the begin and end time (Time DS), while the other one is to identify the media.

Audio Segment DS is a specialized instance of Segment DS. Segment DS is an abstract type, and defines the properties of segments, such as Audio Segment DS. The Audio Segment DS is utilized to describe a temporal interval or segment of an audio. In order to describe the structural relations among segments, segment relation description tools should be used. In the context of temporality, we make use of the thirteen interval relations as indicated in (Allen, 1983), such as *before*, *after*, *overlaps*, *during*, *starts*, *finishes*, and *meets*.

An *Audio Frame DS* is also a temporal portion of the audio stream that have one or more characteristics different from other frames in the stream. Audio frames may contain a list of components such as *audio objects* and *audio events*.

The semantic information about an audio object and event are handled by the *Audio Object* and *Audio Event* DSs, respectively. Each object and event are described with their attributes such as start-time, end-time, duration and high-level information to facilitate understanding of an audio content. These two components together are also very useful for querying at the semantic level.

Ambience DS, which is an optional DS, is used to describe some information about the entire audio frame to be able to distinguish it from others, while *frame link DS* is used to capture the relations between audio frames. There may be any number of frame links and any number of relationships.

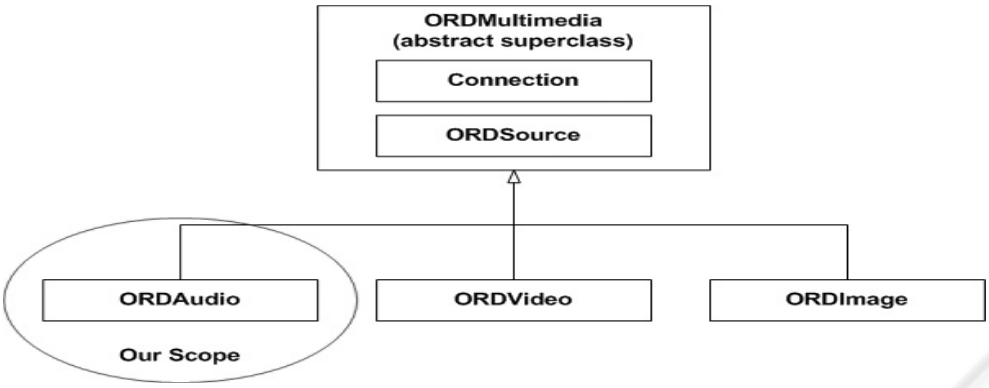

Figure 3: Oracle's multimedia object data types.

3.2 Query Examples

This section provides some query examples and potential applications of the proposed model. Proposed system supports both the Query-by-Example (QBE) and semantic (textual) queries. QBE queries are performed by providing an example query object, while semantic queries are expressed in the form of object and event concepts. Example queries are,

- Retrieve audio piece(s) that are *similar* to **A**
- Retrieve audio piece(s) in which object **O** *appears*
- Retrieve audio piece(s) in which event **E** *appears*
- Retrieve audio piece(s) in which object **O1** *appears before* object **O2**
- Retrieve audio piece(s) in which event **E1** *appears after* event **E2**

where A, O, and E represent an audio instance, an object, and an event, respectively. In the examples, *before* and *after* are temporal predicates. Other temporal predicates (e.g., *starts*, *finishes*, *overlaps*) are also supported. In addition, the queries can also be expressed in the form of *conjunctive* and *disjunctive* queries.

4 INTEGRATION ISSUES

Several database management system vendors have included characteristic features of object-oriented databases to relational database management systems. In particular, many database vendors, such as IBM and Oracle provide the capability of handling object data by embedding content-based retrieval prototypes. For instance, IBM's DB2 provides content-based retrieval for images and video, and its prototype has been employed from the research called QBIC (M. Flinker, 1995), while Oracle makes use of the prototype called Virage (A. Hampapur, 1997). Although both systems reasonably support the content-based retrieval of image and video, they do not provide the same capability for audio (Sert and Baykal, 2003).

We have mainly emphasized to the Oracle database management system for integration issues of our model. Oracle is an object relational database management system (Oracle, 2000). This means that, in addition to its traditional role in the safe and efficient management of relational data, it provides support for the definition of object types, including the data associated with objects and the operations (methods) that can be performed on them. This mechanism is established in the object-oriented paradigm, thus enabling complex objects, such as digitized audio, image, and video to the databases.

Within Oracle, multimedia data is handled by the ORD* data types (ORDAudio, ORDImage, and ORDVideo), which are provided by the technology called interMedia. All three data types derived from the abstract object type called ORDSource. This hierarchy is shown in Fig. 3. Interested readers may refer to (Sert and Baykal, 2003) and (Oracle, 2000) for content-based retrieval capability of the database. However, this feature is a lack for auditory data. Therefore, Oracle provides some methods to extend this feature for ORD* data types. ORD* data types make possible the following features:

- Manipulating multimedia data sources
- Extracting attributes from multimedia data (partially)
- Content-based retrieval of image and video

These data types can be extended to support audio and video data processing, as well as content-based retrieval of auditory data. In order to achieve these goals, we apply the following procedures:

- Design of the new/extended data source (model)

- Implementation of the new/extended data source (model)
- Installation of the new module as a plug-in by using the ORDPLUGINS schema
- Adjustment of the privileges of new plug-in

5 CONCLUSIONS

Considerable research has been conducted on video and audio data modeling in recent years. However, to the best of our knowledge, most of them were application of specific approaches. With this motivation, in this paper, we described the audio modeling constructs and presented how an audio information can be modeled in the context of MPEG-7 descriptors and description schemes, in order to provide interoperability in world-wide scale. Finally, we have explored the integration issues of the model in commercial database management systems. Since proposed model exposes the structure of a generic audio description scheme, as a result, it can be used in various audio applications as an underlying data model to handle the audio characteristics and their semantics.

Our future work lies on two directions: (a) encapsulation of the proposed data model to constitute a composite audio data type, (b)implementation of a symbolic query language to query it.

REFERENCES

A. Ghias, J. Logan, e. a. (1995). Query-by-humming-musical information retrieval in an audio database. In *ACM Multimedia Conference*. Proc. ACM.

A. Hampapur, e. a. (1997). Virage video engine. *SPIE*, 3022.

A. Woudstra, e. a. (1998). Modeling and retrieving audio-visual information. *LNCS*, 1508.

Adam T. Lindsay, e. a. (2000). Representation and linking mechanism for audio in mpeg-7. *Signal Processing: Image Communication*, 16:193–209.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of ACM*, 26(11):832–843.

E. Wold, T. Blum, e. a. (1996). Content-based-classification, search, and retrieval of audio. *IEEE Multimedia*, pages 27–36.

Foote, J. (1997). Content-based retrieval of music and audio. In *Proceedings of SPIE'97*.

G. Amato, e. a. (1998). An approach to a content-based retrieval of multimedia data. *Multimedia Tools and Applications*, 7(1/2):5–36.

Ghafoor, A. (1994). Multimedia database course notes. In *ACM Multimedia Conference*.

Grosky, W. (1997). Managing multimedia information in database systems. *Communications of the ACM*, 40(12):73–80.

Gudivada, V. and Raghavan, V. (1995). Content-based image retrieval systems: Guest editors' introduction. *IEEE Computer*, pages 18–22.

John R. Smith, A. B. B. (2000). Conceptual modeling of audio-visual content. In *IEEE International Conference on Multimedia and Expo (II)*, pages 915–. IEEE Press.

J.Z. Li, M.T. Ozsu, e. a. (1997). MOQL: A Multimedia Object Query Language. In *The 3rd International Workshop on Multimedia Information Systems*.

L. Lu, H.J. Zhang, e. a. (2003). Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8:482–492.

M. Flinker, e. a. (1995). Query by image and video content: The qbic system. *IEEE Computer*, 28:23–32.

MPEG-7 (1999). Mpeg-7 requirements document v.8, iso/iec jtc1/sc29/wg11/n2727. Technical report, Seoul Meeting.

MPEG-7 (2001). Multimedia content description interface - part4: Audio, iso/iec jtc1/sc29n. Technical report, MPEG-7.

Oracle (2000). User's guide and reference: Oracle intermedia audio, image, and video. Technical report, Oracle.

P. Salembier, e. a. (1999). Video ds. Proposal P185, P186, MPEG-7 Lancaster Meeting.

Petkovic, M. and Jonker, W. (2000). An overview of data models and query languages for content-based video retrieval. In *International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*.

R. Weiss, e. a. (1994). Content-based access to algebraic video. In *Proc. of Int. Conf. on Multimedia Computing and Systems*, pages 140–151. IEEE Press.

Sert, M. and Baykal, B. (2003). A web model for querying, storing, and processing multimedia content. In *IKS'03, International Conference on Information and Knowledge Sharing*. ACTA Press.