From Multimedia Retrieval to Knowledge Management

Pedro J. Moreno JM Van Thong Beth Logan

Cambridge Research Laboratory

Technical Report Series

CRL 2002/02

March 2002

COMPAQ

From Multimedia Retrieval to Knowledge Management

Pedro J. Moreno JM Van Thong
Beth Logan
Cambridge Research Laboratory
Compaq Computer Corporation
Cambridge MA 02142-1612

March 2002

Abstract

We explore how current traditional applications in multimedia indexing can evolve into fully-fledged knowledge management systems in which multimedia content, audio, video and images, are first class citizens and contribute as much as textual sources. We start by describing a current application for indexing audio and video from the Web, the *SpeechBot* web index, and continue by exploring possible uses and expansions of this technology for knowledge management. We describe the main problems the use of audio and video data introduce to knowledge management and suggest ways to compensate for them.

Authors email: {Pedro.Moreno, JM.VanThong, Beth.Logan}@compaq.com

CRL Technical reports are available on the CRL's web page at
http://crl.research.compaq.com.

Compaq Computer Corporation
Cambridge Research Laboratory
One Cambridge Center
Cambridge, Massachusetts 02142 USA

1 Introduction

Knowledge management (KM) is in general defined as capturing and organizing the expertise, experience and collective "know how" of an organization and making this knowledge easily accessible. This knowledge is typically available in computer-readable form and most techniques for knowledge management expect the data to be in a structured form such a as relational database, or at least in semi-structured form such as formated textual sources. The standard KM approach typically organizes knowledge in portals, uses text search and analysis tools, and in general relies heavily on text as the medium to transfer knowledge.

In recent years, technological advances in the storage, distribution, and production of multimedia have created a new source of information for knowledge management systems. However, at present, multimedia is at most indexed and used in retrieval systems. Users follow a typical search engine approach where a query is introduced and segments of multimedia documents deemed similar to the query are returned. There is no attempt to extract knowledge from these documents.

The use of multimedia in KM systems presents many challenges. First, it cannot be used in its native form. It must be analyzed and transformed into a format that a KM system can use. Typically media processing algorithms or *analyzers* extract knowledge and transform it into metadata. This metadata is an intermediate representation of the multimedia data that is easier to manipulate and process using standard information retrieval methods.

The second challenge is that this transformation or analysis of multimedia data introduces uncertainty. No analysis system is perfect or error free. Speech recognizers, speaker recognizers, topic analyzers and in general any analyzer will make errors. Third, as the data in its native form is unstructured, our analysis tools must extract and infer the hidden structure and knowledge behind the data. Furthermore, the inferred structure must be stored in formats that allows easy access and manipulation. Finally, the volume of data introduces problems of scalability, organization and user interface. For example, when analyzing thousands of hours of audio, we need to design systems able to process such volumes of data quickly and effectively and to display this information to users in an intuitive way.

This article describes some solutions to these challenges and then provides examples of how current multimedia indexing systems can be part of more sophisticated knowledge extraction systems.

2 Multimedia Retrieval Systems: the SpeechBot architecture

In multimedia retrieval systems, speech recognition may be used to produce a textual transcription. A time-coded index built from the transcription allows users to query by keywords. A typical example of such a system is *SpeechBot* [8]. *SpeechBot* is a general tool for audio and video indexing. The system is designed to handle large volumes of data both in speech recognition processing and in the number of user queries. *SpeechBot* fetches audio and video documents from the Web or Intranet and builds an index

from that data. It does not serve content, but rather keeps a link to the original document similar in approach to traditional search engines such as AltaVista. The search index is available and running on the Web[1]. Figure 1 shows a typical result from a search. By clicking on the "play extract" button the user can play the multimedia stream at the time location in which the query words are pronounced.

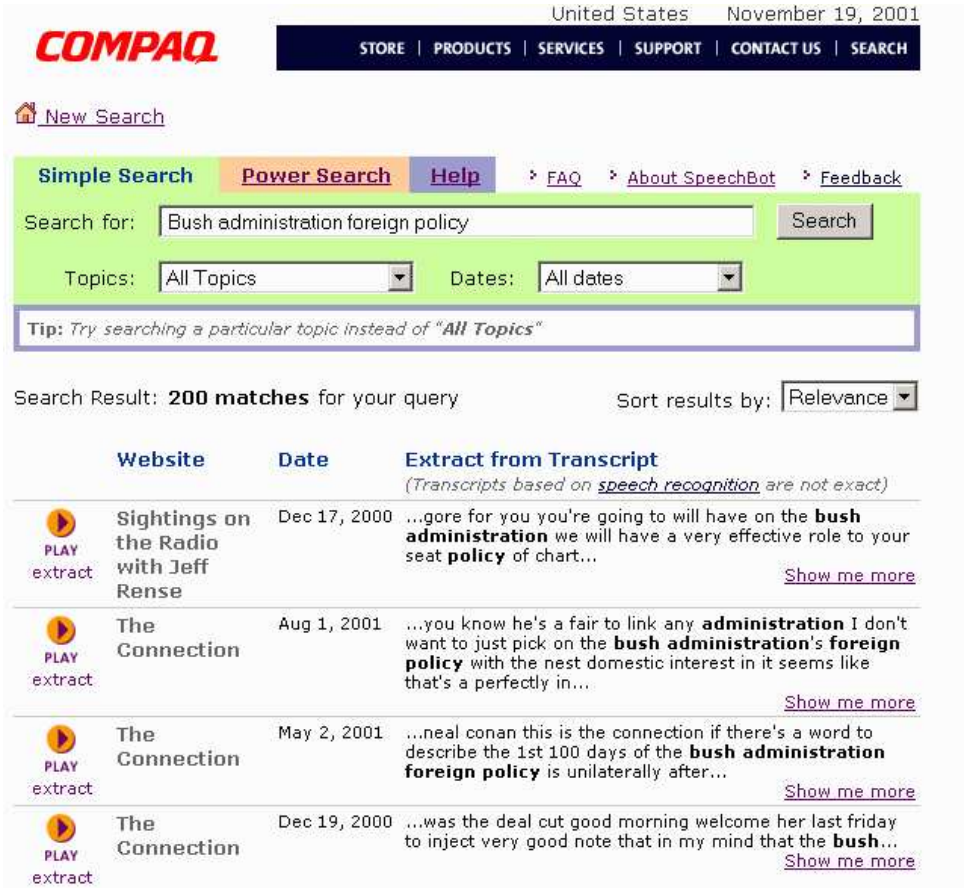

Figure 1: SpeechBot search page example

Figure 2 presents the system architecture at a conceptual level. The system consists of the following modules: the transcoders, the speech decoders, the metadata database, the indexer, and the UI manager.

[1] *http://www.speechbot.com*

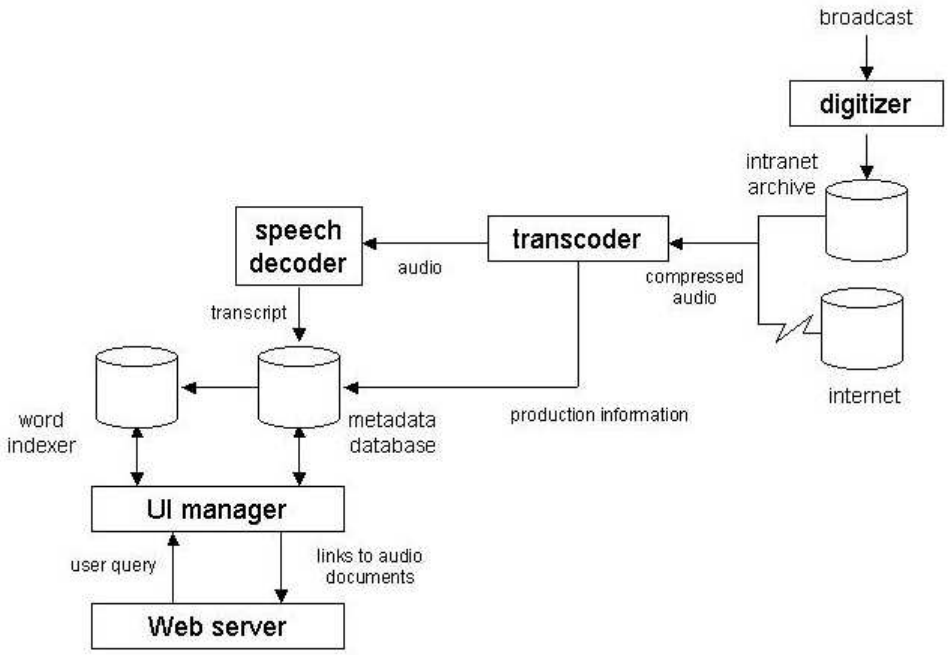

Figure 2: Overall architecture of the system

2.1 Transcoders

Broadcasts and archives are processed by the *transcoders*. These digitize the video and audio and capture information related to each recording. When fetching the documents from the Internet, the transcoders download each file and extract and store the production information as metadata. The metadata, when available, contains information about the downloaded file such as the sample and frame rate, copyright, the story title, and possibly a short description. The audio and video content is stored on a temporary local repository and converted into an uncompressed unified format.

2.2 Speech Recognition

SpeechBot, like similar multimedia indexing systems, uses a large vocabulary continuous speech recognition package using state-of-the-art mixture Gaussian, triphone based, Hidden Markov Models (HMMs). The acoustic and language models were trained on audio and textual sources of broadcast news data. The speech recognizer produces a textual transcription, or annotation stream, from the downloaded audio files. The annotation stream consists of the start time and end time of each word in the transcript and the word itself. The audio files are segmented in such a way that the speech recognition can be performed in parallel on different portions of the document.

A farm of workstations tied together in a queuing system recognizes each portion of the document. Thus, even if the speech decoder is not real-time, it can still achieve sub real-time throughput. When all the portions are recognized, the results are assembled to create a fully annotated document. If a transcription is already available for a section of audio, the speech recognition module may be replaced with an aligner module. Its role is to provide time marks for each word of the input text.

2.3 Indexer

The indexer provides an efficient catalogue of audio and video documents based on the transcription produced by the speech decoder. As well as supplying the user interface with a list of documents that match a user's query, the indexer also retrieves the word location of these matches within the documents.

The indexer sorts the matches according to relevance. Relevance is defined using the term frequency inverse document frequency (tf/idf) metric [7], adjusted for the proximity of the terms within the document. This sorting is performed on both the list of documents returned and the list of match locations.

2.4 Metadata database

The metadata database stores the semantic information produced by the transcoders and speech recognition modules. It maintains a mapping between word locations in the index to 10 second long text clips and the time the clip occurs in the multimedia document. The user interface uses this information to construct the query response pages displayed to the user. The use of a central repository for shared information allows a robust distributed architecture which can scale on demand.

2.5 User Interface manager

The user interface manager connects to the indexer to process the queries, retrieves the most relevant documents and presents them to the user. For each document, a brief 10 seconds excerpt extracted from the metadata database is displayed which allows the user to see at a glance what the document is about.

3 Beyond Multimedia Retrieval

Clearly a system such as *SpeechBot* while useful is still far from a KM system. It represents a very first step in the right direction. To improve its performance and to evolve into a knowledge management system several changes and improvements are needed.

3.1 Multimedia Analyzers

During the production of multimedia content many annotations are available. For example, in video production the beginning and ending time of each of the scene shots,

its production time, the origin of the source, the production company, a small summary describing the shot are known. Editors and producers often create their own annotations and sometimes even close captions are produced before or during the broadcast of the video. Unfortunately, annotations of this type are often lost. Also, they represent only a partial description of the multimedia as they might miss important details. Finally, they are expensive to produce since they require human intervention. Clearly there is a need for a more automatic analysis of the data.

Because of the multi-track nature of multimedia, where we can find video, audio, and text tracks, several analyzers may be used. Each one is additive and complementary, additive in the sense that each one works on different tracks and complementary in the sense that multiple analyzers can also be applied to the same track.

3.1.1 Analyzer examples

Figure 3 shows several possibilities for analyzing the different tracks.

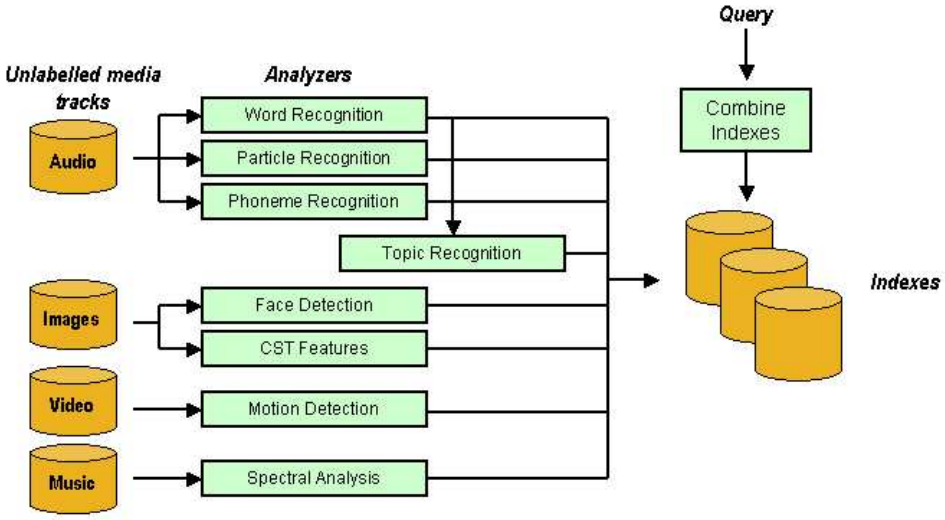

Figure 3: Multiple Analyzers are applied to different multimedia tracks

In this example, three analyzers are applied to the audio stream. The first one is a conventional large vocabulary continuous speech recognizer. The second one is a particle recognizer, where each particle is a subword unit that can vary from a single phone to several phones long. The particle recognizer uses an associated trigram particle language model and the particles are automatically learned from textual data. Finally a phonetic recognizer is also applied to the audio signal. Each of these speech recognizers works at different time resolutions and with different constraints hence providing a different view of the audio speech signal.

Other forms of audio analysis could use speaker recognition technology to identify

who is speaking or audio scene analysis to classify the audio as belonging to broad categories such as clean, telephone bandwidth or having music in the background. For those segments recognized as music, we could also attempt to identify the music and perhaps even find similar sounding songs. Systems such as *BoogeeBot* [6] represent first steps in this direction.

The images from the multimedia stream (keyframes) can be effectively analyzed with face recognition algorithms. CST (Color, Shape, Texture) features can be extracted for each image and stored in a metadata database. Query by example methods can be used to find similar images later on. Video streams can also be annotated with motion detectors, shot detectors and keyframe extractors. The text displayed in the video can also be recognized and transcribed. In general, specific detectors can be applied to the video stream to improve and enrich its annotation.

3.1.2 Combining Analyzers

As mentioned in the introduction, the analyzers are not perfect and will introduce errors. There is always an associated uncertainty with their annotations.

The speech recognition analyzers provide a good example. A significant percentage of the word error is due to out of vocabulary words (OOVs), *i.e.* words not present in the speech recognizer dictionary but present in the audio stream. These words will never be hypothesized since the recognizer is not aware of them. Often it is these words the most important to recognize but by their very nature they are also quite hard to predict [10].

The role of the particle and phonetic recognizers is to complement and aid a word recognizer. Since any word can be expressed as a combination of particles and phonemes, particle and phonetic recognizers do not suffer from the OOV problem and can compensate for the limitations of word recognizers. With a simple transformation of query words into particle or phonetic sequences we can search for any word. In addition they provide us with an intermediate representation of the audio that can be further explored. For example, new words not present in the word recognizer can be searched for in the particle/phoneme metadata representation at a later time and then they can be reinserted into the word index.

We can also combine analyzers to improve performance. For example, the textual output of the word/particle recognizers can be searched for particular topics. Segments of audio (via its textual representation) can be segmented into coherent stories and new topics not previously seen can be identified [4, 11]. A face recognizer can be combined with a speaker detector to improve identification.

Methods as simple as majority voting schemes can help us in combining all these knowledge sources. More sophisticated techniques based on data fusion methods or Bayesian combination of knowledge sources provide further opportunities for improvements. In summary, the combination of all these sources of knowledge can help us to reduce the errors introduced by individual methods.

3.2 Representation of Information

Metadata is the intermediate representation that describes the content of multimedia documents. The term is used in its most general sense; it could be textual annotations generated by hand, or some specific features extracted by a content analyzer. The metadata describes the structure and the nature of the document content, and the relations between the different parts of the document. It is a compact representation that can be search or indexed to retrieve information. Examples of metadata include: word or phoneme transcriptions of spoken content, CST feature vectors for images, or topic segmentation and class. The metadata itself can be used to generate new metadata.

As it represents new descriptive information, the problem to solve is how to represent and store this new data and how to use it to efficiently retrieve information.

3.2.1 Metadata encoding and storage

The output of the different content analyzers can be combined into an XML representation that organizes the extracted knowledge in an easy manner. Below we show a possible example. This representation allows indexing and quick access for information retrieval purposes. The use of XML adds the benefit of an industry standard representation easy to exchange between system components. Figure 4 shows an example of how multiple analyzers provide different views of a multimedia document.

There are several initiatives to define a standard for multimedia content representation. Time coded multimedia metadata can be represented with SMIL, an XML based language that can be used to describe the temporal behavior and layout of multimedia presentations. The Library of Congress is also promoting the development of a standard set of multimedia attributes, called *METS* (Metadata Encoding and Transmission Standard). The Dublin Core Metadata Initiative (DCMI) provides recommendations and standards, mostly focused on the traditional publishing industry, but many elements can be reused for the description of multimedia documents. Finally, the MPEG-7 standard includes several elements for content description, such as spoken content in intermediate forms.

The metadata must be stored in a database to allow complex querying and retrieval of information. Current database technology such as *Oracle* is expanding to model XML data representations. In addition new native XML database technology is emerging [1, 2].

3.2.2 Indexing metadata

Indexing metadata in its native form is a challenging problem, mostly because of the inherent inaccuracy of the information generated by content analysis, and the nature of the metadata itself. If the metadata has a textual form, then a text-based index, such as those used for indexing the Web can be used. If on the other hand the metadata is a high dimensional feature vector extracted from the multimedia, such as CST feature vectors from images, an indexing approach is much harder. Often we are limited to search approaches that must scan the whole database. Query by example image search engines such as IBM's *QBIC* [3] are the current and most used approach to this problem.

```
- <mmdoc>
    <origin source="WBUR 90.9 FM, Boston" id="wbur 90.9 fm, boston"
      copyright="Copyright 2001 WBUR, Boston" />
    <media mimetype="ra" bitrate="16000" duration="3081.38" />
    <show name="Here and Now" airdate="2001-06-26" id="hn062601" />
  - <content>
      <clip id="24" begin="6.28" end="16.65" type="word"
        author="calista.v.3.0">president bush meets with israeli prime minister
        ariel sharon at the white house today aids say the president will stress
        the importance of .....</clip>
      <clip id="24" begin="6.28" end="16.65" type="particle"
        author="calista_part.v1.0">P_ R_EH_Z_ AH_D_ EH_N_T__w
        B_UH_SH__w M_IY_ T_S__w W_IH_ DH__w IH_Z_ R_EY_ L_IY__w P_
        R_AY_M__w M_IH_N_ AH_S_ T_ER__w EH_R_ IY_AH_L__w SH_AE_R_
        AH_N__w AE_T__w DH_ .....</clip>
      <clip id="24" begin="6.28" end="16.65" type="phone"
        author="calista_phone.v2.4">P R EH Z AH D EH N T B UH SH M IY T S W IH
        DH IH Z R EY L IY P R AY M M IH N AH S T ER EH R IY AH L SH AE R AH N
        AE T DH AH HH W AY T HH AW S T AH D EY EY D Z S EY DH AH P R EH
        Z .....</clip>
    - <topic id="24" author="guayabal.v1.2">
        <hyp value="israel" score="0.82" />
        <hyp value="war" score="0.13" />
      </topic>
    </content>
  </mmdoc>
```

Figure 4: XML representation of multimedia document as seen by different analyzers

3.3 Toward Knowledge Management

With architectures such as the one described above we can build effective KM systems that take advantage of multimedia data. For example, researchers at IBM, CMU and other institutions are investigating the analysis of spoken discourse for meeting support [5, 9].

Meetings generate a wealth of information that is usually lost. We typically rely on note takers to capture the experience of the meeting. This is a time consuming process that probably forgets important facts. Additionally, during the course of a meeting, participants may not have easy access to all the information they require. IBM's *MeetingMiner* attempts to solve these problems. *MeetingMiner* captures and analyzes the meeting audio, transcribing it using speech recognition. It can thus provide a searchable record of the event. In addition, it can bring important information to the participants' attention. For example, in a technical meeting it can search a patent database and alert the participants to potentially similar intellectual property owned by competitors. Although still in an early stage and limited by the challenging speech recognition environment of meetings, this project is a promising example of a true multimedia KM system.

A similar system was built by researchers at CMU [9]. They built a system capable of tracking, categorizing and summarizing meetings. It is composed of a speech recognizer, a summarizer, a tool to detect salient and novel turns in the meeting, a discourse component that identifies speech turns and an analyzer of non-verbal cues based on video analysis. Their prototype allows users to rapidly review records of human interaction using an architecture similar to the one described in this paper.

4 Conclusions

Multimedia data introduces several challenges to KM systems. Uncertainty is introduced by media analyzers. The need for good scalability and effective user interfaces present other problems. Continuous research will be needed in each of these areas. Effective architectures able to handle the complexity of such systems will also play a crucial role in the deployment of multimedia based KM solutions.

The prospects for fully exploiting multimedia content are promising. Several situations where multimedia is not exploited effectively come to mind. We have given some details of the IBM *MeetingMiner* and the CMU meeting browser projects. But there are more interesting opportunities. For example, the wealth of audio data available in telephone consumer departments offers opportunities for analysis and knowledge extractions. The detection of trends in demand and complaints are intriguing opportunities for KM systems.

Our experience developing the *SpeechBot* system shows that even with today's current limitations in speech recognition technology, good performance can be obtained when searching multimedia sources. When this is combined with the current trends in audio and video analysis, in multimedia storage and distribution over the Internet, the developments in XML representations, and the integration with knowledge portals, there is hope that multimedia data will become truly pervasive and eventually as important if not more than textual sources in building knowledge management systems.

5 Acknowledgments

The *SpeechBot* project has benefited from the insights and work of many people. We first would like to thank Scott Blackwell, for his web master support. We also would like to acknowledge the past contributions of Edward Whittaker, Mike Swain, Dave Goddeau, Blair Fidler, Katrina Maffey, Matthew Moores and Chris Weikart.

References

[1] http://www.xyzfind.com.

[2] http://www.tamino.com.

[3] J. Ashley and M. Flickner. The query by image content (qbic) system. In *ACM SIGMOD Conference, May, 1996.*, 1996.

[4] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177, 1999.

[5] E. W. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, and A. Amir. Towards Speech as a Knowledge Resource. *IBM Systems Journal*, 40(4), 2001.

[6] B. Logan. A music similarity function based on signal analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2001.

[7] G. Salton and M. J. McGill. In *Introduction to Modern Information Retrieval. McGraw-Hill*, 1983.

[8] J. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain. Speechbot: a speech recognition based audio indexing system for the web. In *Proc. of the 6th RIAO Conference, Paris, April 2000.*, 2000.

[9] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA*.

[10] E. Whittaker. Temporal adaptation of language models. In *Proceedings of the ISCA Workshop on Adaptation Methods for Speech Recognition*, 2001.

[11] J. Yamron, I. Carp, L. Gillick, and S. Lowe. A hidden markov model approach to text segmentation and event tracking. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

CRL 2002/02
March 2002

From Multimedia Retrieval to Knowledge Management

Pedro J. Moreno JM Van Thong Beth Logan

COMPAQ