From Multimedia Retrieval to Knowledge Management

The authors suggest ways in which speech-based multimedia information retrieval technologies can evolve into full-fledged knowledge management systems in which audio, video, and images contribute as much as textual sources.

Pedro J. Moreno
J-M. Van Thong
Beth Logan
Compaq Cambridge Research Laboratory

Gareth J.F. Jones
University of Exeter

Knowledge management (KM) is generally defined as the capture of an organization's collective know-how and making that expertise easily accessible. This knowledge is typically available in computer-readable form—most often in a structured form such as a relational database, but also in semi-structured form such as formatted textual sources. Standard KM approaches typically organize knowledge in portals, use text search and analysis tools, and rely heavily on text as the medium for transferring knowledge.

Although recent technological advances in multimedia production, storage, and distribution have created new information sources, multimedia use in KM is largely limited to retrieval systems: Users typically follow a search engine paradigm in which a query returns ostensibly relevant multimedia documents but without any attempt to extract knowledge from them.

Using multimedia in KM systems presents many challenges. First, KM systems cannot use multimedia objects in their native form—they must use media-processing algorithms to transform the objects into metadata. The metadata serves as an intermediate representation of the multimedia data that is easier to manipulate and process using standard information retrieval methods.

Second, this transformation of multimedia data introduces uncertainty because no analysis system—such as speech recognizers or face recognizers—is error free, and these tools must extract the knowledge this inherently unstructured data contains and store it in formats that allow easy access and manipulation.

Finally, the volume of data introduces problems of scalability, organization, and user interface. For example, systems that analyze thousands of hours of audio must process such volumes of data quickly and effectively and display it in an intuitive way. Technologies for indexing audio and video can be an integral component of these more sophisticated knowledge extraction systems. Possible avenues for improvement in these systems include more advanced text information retrieval methods and more sophisticated speech technologies.[1,2]

SPEECH TECHNOLOGIES IN INFORMATION RETRIEVAL

Spoken data is a common nontextual source of information available to KM systems. Researchers are focusing on techniques that can analyze speech for indexing and alter information retrieval systems to account for uncertain data. These speech analysis techniques illustrate many of the principles researchers are using for other types of multimedia analyzers.

Spoken document retrieval systems rely on words as the medium of information. These systems use a speech recognizer to transcribe speech or audio so they can apply traditional text information retrieval (IR) techniques.

Unlike text-indexing engines, however, spoken document retrieval must deal with transcription errors. Retrieval systems must compensate for the 20 to 30 percent word error rates that commonly

0018-9162/02/$17.00 © 2002 IEEE

occur when large-vocabulary speech recognizers transcribe unrestricted audio such as broadcast news or informal speech.

IR improvements

Similar to text retrieval, spoken document retrieval is not concerned with uninformative stop words such as *the*, *it*, *a*, and so forth. These short words occur frequently, are poorly articulated, hard to recognize, and in general add little value when searching for relevant documents. Therefore, removing them from the index improves retrieval performance. Similarly, for semantically related words, suffix stripping or stemming to a common root can facilitate matching between different word forms. For example, the words *document*, *documents*, *documentation*, and *documented* can easily confuse a speech recognizer; mapping these words to the common stem *document* typically improves retrieval.

A second IR method for preventing speech recognizer errors combines techniques such as relevance feedback and query expansion. Relevance feedback is a two-pass method. In the first pass, users enter a query and select those hits they consider relevant. In the second pass, the system uses the selected documents to compose a more powerful query. The relevance feedback from the first pass helps to improve retrieval performance in the second pass.

Pseudorelevance feedback removes user intervention by assuming that the top documents from the first pass are relevant and then using the two-pass method. In general, pseudorelevance feedback is not as effective as traditional relevance feedback, but it speeds up the search significantly.

Query expansion uses semantically related terms to expand the query. For example, query expansion might augment the query *George Bush* by adding *White House* and *President*, which the search can extract from offline text collections. Another form of query expansion uses acoustic similarity to account for possible mistakes in the speech recognizer.

Speech recognition improvements

Word-based speech recognition systems use preset vocabularies including 60,000 to 100,000 words.[3] By definition, the system cannot hypothesize words outside this vocabulary. While a vocabulary of 100,000 words includes most spoken words, every document includes a small percentage of out-of-vocabulary (OOV) words that are likely to be content-bearing terms, and not including them has an adverse effect on retrieval performance.

To circumvent this problem, the system can tailor the vocabulary by examining documents related to the task. For example, a speech recognizer used for court hearings could use legal documents to learn the appropriate dictionary words. While these specialized vocabularies can reduce the number of OOV words, they cannot guarantee their elimination.

Word spotting. An alternative to large-vocabulary transcription is word spotting. A word-spotting system has a limited vocabulary that typically includes fewer than 50 keywords selected for a particular task. The word spotter transcribes the audio material that passes through it as a predefined keyword, other speech, audio, or silence. The word-spotting approach is attractive because the systems are simple and have low computational requirements. A disadvantage is that if a new search word is introduced, the word spotter must reprocess the entire document.

Word spotting is an attractive approach because the systems are simple and have low computational requirements.

Subword recognition. The vocabulary limitations of the word-spotting approach have led to a number of open vocabulary-indexing strategies based on subword recognition. Rather than recognizing spoken words, these approaches recognize subword units—typically, phonemes or syllables—from which all words are constructed. The IR system decomposes search terms into their constituent subword strings, then scans the recognized terms for strings corresponding to the search unit.

Retrieval systems can use two approaches to perform a subword match between query terms and documents: *n*-gram matching and approximate- or fuzzy-string matching. In *n*-gram matching, the system extracts fixed-length phoneme sequences from the search words and scans the sequences for these *n*-grams. Approximate string matching substitutes, inserts, or deletes phonemes. These replacements take into account the most likely errors observed on training data. As the possibilities for a match increase, the search will find more relevant documents at the cost of more false positives. Clearly a tradeoff is needed to maximize retrieval performance.

Thus far, we have assumed that the retrieval system represents documents as a linear sequence of (sub)words. Other alternative intermediate representations are graphs or word lattices, which are readily available because large-vocabulary speech recognizers often use them.[4] Lattice nodes repre-

Knowledge Management and Speech Recognition

James Allan
University of Massachusetts

Speech recognition technologies have the potential to play a major role in implementing knowledge management techniques, and they may have a dramatic impact on knowledge management applications as well. For example, a properly integrated KM system could offer an indexable catalog of meetings, phone conversations, and e-mail messages, including cross-reference links to individuals with expertise in the subject the system is mining.

Although speech recognition has been used successfully in some KM-related applications, several problems remain unresolved, including improving the quality of speech recognition for telephone conversations.

Capturing Information

KM generally refers to the techniques an organization uses to capture members' and customers' information and habits and to store that knowledge for later use. Broadly speaking, businesses use these techniques to formalize their management strategies and improve the use of their intellectual assets (http://www.cio.com/archive/110199_think_content.html). These intellectual assets can be either documents or knowledge that is not stored elsewhere—in other words, either information or processes (http://www.sveiby.com.au/articles/KnowledgeManagement.html).

The information retrieval community focuses on using KM as a wrapper around information capture, indexing, and retrieval, with some careful profile crafting—either manually or automatically—to get the right information to users. However, facilities for storing and accessing objects efficiently and meaningfully remain an important KM component.

KM Applications

The simplest KM capability is document storage and retrieval in a variety of formats. Other applications include

- Creating agents that monitor information sources for particular items. These agents provide a customizable query that indicates what type of passing documents it should retrieve.
- Indexing people by the documents they create and store or by self-generated descriptions of their interests to help an organization rapidly locate expertise on a particular topic.
- Representing situations or cases by the documents and people associated with them. This type of indexing operates on the premise that information about previous situations has value when similar situations occur in the future.
- Tracking the information flow within an organization. Observing where new information enters an organization and how it moves can highlight individuals or departments or it can illuminate information-based social structures for sharing knowledge.
- Extracting information automatically from arriving data. Mining Web pages for price information as part of competitive analysis is an example of this application.

These applications deal primarily with text, but they can just as easily work with documents created from spoken information. The spoken word will always be a natural way for people to interact—both within and outside an organization. Because many business transactions take place over the telephone, generating transcripts of telephone conversations has potential value for capturing valuable knowledge for future use.

Speech recognition could also help users interact with online information, whether from a telephone, cell phone, or desktop computer. Some operating systems now ship with small-vocabulary voice recognition systems that use speech to complete some tasks—for example, "close this window." This functionality is not part of KM itself, but it does provide an important entry point into stored information.

Speech recognition could also play an important role in capturing dictated information. Knowledge workers who do not have time to key in useful information might be willing to record it while they are on the move.[1] KM techniques could index the resulting transcripts as retrievable documents.

Technical Challenges

Using an automatic speech recognition system to transcribe captured speech would leverage its potential value as an information source by making it available for use in the same way as written documents. For example, recording a meeting could provide automatically indexable and retrievable information, compared to relying on rapidly created transcribed minutes that only attempt to summarize key points.

A remaining problem is that speech recognition systems are not perfect. High-quality speech recordings—for example, an announcer reading an advertisement in a broadcasting studio—might have a recognition error rate of less than 10 percent. Thus, approximately one in ten words would be incorrectly recognized. In contrast, the error rate for conversational speech, particularly on a telephone, ranges from 30 to 40 percent.[2,3]

Fortunately, current indexing and retrieval technologies are robust in the face of speech recognition errors.[4,5] Even with a 40 percent recognition error rate, the effectiveness of a typical document retrieval system decreases only 10 percent. Several circumstances contribute to this statistic:

- Redundancy provides some resilience to recognition errors. If the spoken text is long enough, the most important words repeat—consider how often "knowledge" appears in this article. The voice recognition software probably will recognize the most important words in repeated uses.
- Even if the software sometimes misses critical words altogether, it usually recognizes other strongly related words. For example, if the word "speech" completely escaped recognition in this article, the words "voice" and "spoken" might not. These words would provide sufficient context for effective use of document retrieval techniques.
- Many unrecognized words are not content bearing. If a transcript omits the word *the*, it is an error. However, *the* is not content bearing, its absence—although contributing to the error rate—is not important for

information storage and retrieval applications.

Any information technology—and any aspect of KM that deals with reasonably sized texts (100 words or more)—is likely to succeed whether it uses written text or text that a speech recognition system generates. Although sufficient recognition errors would make document retrieval fail, speech recognition error rates are low enough even for poor quality speech.[3] Thus, few documents are likely to be inaccessible.

Successes

Despite speech recognition errors, KM has succeeded in both automatic indexing and document retrieval and in topic detection and tracking.

Indexing and document retrieval

Automatic indexing and document retrieval operate on the idea that documents using the same vocabulary discuss the same topics. More sophisticated systems provide elaborate query-processing techniques to make it more likely that a system will find documents relevant to a query.

Document comparison and grouping functions rely upon overlapping words for the core of their success. Some researchers originally felt that errors would complicate speech recognition efforts, but retrieval of spoken documents has been successful because of the inherent redundancy in speech.

Questions remain concerning how well a system will do when it integrates spoken and written documents into a single setting. Anecdotal evidence suggests that retrieval systems are likely to select written documents because the important words appear to occur more frequently, and they are not misrecognized.

Topic detection and tracking

Spoken documents do not appear to cause problems with automatically organizing news stories by the events they describe. Topic detection and tracking (TDT)[6] tasks focus on television and radio news, where speech recognition is the only way to acquire text transcripts—except in the case of television programs that provide closed captioning.

Although KM systems do not currently use TDT, this research program has potential interest for many KM settings. For example, TDT could identify new topics and group news stories by their underlying topics.

TDT research has demonstrated a limited impact when using recognized speech instead of written materials as an information source. As with document retrieval, TDT's robustness is largely attributable to basing comparisons on large stretches of speech in a complete news story, where repetition of words is common.

Unresolved Problems

Recognition errors could be a significant problem in some KM applications. Information retrieval systems are more sensitive to recognition errors in spoken queries than in documents derived from speech-recognition output because a short span of recorded speech lacks redundancy, which offsets recognition errors (http://www.destinationcrm.com/km/dcrm_km_article.asp?id=973).

Mining text for small pieces of information—for example, WhizBang!'s job listing (http://www.flipdog.com)—is also likely to be less robust. Some research shows that at the 40 percent or higher speech recognition error rates for telephone speech, the system's ability to find the names of people, places, and organizations degrades by 80 percent compared to written text (http://svr-www.eng.cam.ac.uk/~ajr/esca99/). Although some studies suggest that the problem may not be quite so severe (http://www.nist.gov/speech/publications/darpa98/html/lm50/lm50.htm), recognition errors clearly have a much stronger impact on finer-grained text analysis tasks.

Having more information does not necessarily mean that the knowledge is readily available. KM requires techniques for extracting knowledge from the data. Also, on a more social than technological level, capturing more information introduces privacy issues that demand attention. If the captured information has enough value, an organization probably can devise methods for handling social problems—for example, providing an "off the record" mode.

Speech recognition clearly plays an important role in using KM for capturing, indexing, and using information. Capturing the information that people exchange in meetings, during telephone conversations, or in casual settings could provide access to more of the underlying knowledge that an organization wants to preserve. The experience with applying several document-based tasks to spoken documents indicates that there is potential for success in this endeavor.

References

1. S. Barth, "Tell It to the Machine: A New Generation of Recorders and Software Make It Easier to Digitize Knowledge," *Knowledge Management*, http://www.global-insight.com/KM-VR.htm.
2. M. Padmanabhan et al., "Evolution of the Performance of Automatic Speech Recognition Algorithms in Transcribing Conversational Telephone Speech," http://www.research.ibm.com/voicemail/pdf/imtc2001.pdf.
3. J.G. Fiscus et al., "2000 NIST Evaluation of Conversational Speech Recognition Over the Telephone," http://www.nist.gov/speech/publications/tw00/html/abstract.htm#cts-10.
4. J.S. Garafolo, C.G.P. Auzanne, and E.M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *Proc. TREC-8* (1999), Dept. Commerce, Nat'l Inst. Standards and Technology, Gaithersburg, Md., 1999, pp. 107-130.
5. J. Allan, *Perspectives on Information Retrieval and Speech*, Lecture Notes in Computer Science, Springer-Verlag, New York, 2002, pp. 1-10.
6. J. Allan, ed., *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic, Boston, 2002.

James Allan *is an assistant professor in the Computer Science Department at the University of Massachusetts and is the assistant director of the Center for Intelligent Information Retrieval. His research interests include topic detection and tracking, automatic information organization, and interactive and multimodal information retrieval. He received a PhD in computer science from Cornell University. Allan is a member of the IEEE Computer Society, the ACM, and ASIST. Contact him at allan@cs.umass.edu.*

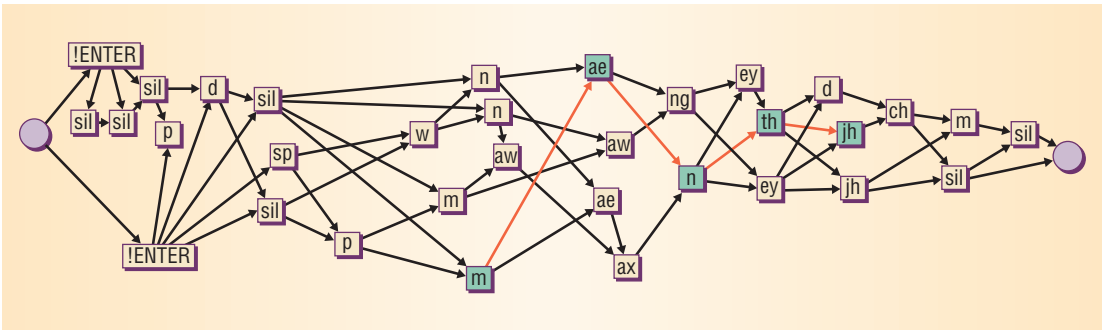

Figure 1. Lattice representation encoding different hypotheses for the word* manage*. Starting with the initial phoneme* m*, the scanner follows the path to the phoneme* ae*, and so on to the end of the word.

sent a hypothesized word or subword, and the connections between nodes represent alternative paths in the lattice. An information retrieval system can use a word lattice for a more effective search in appropriately encoded alternative hypotheses.

Figure 1 shows a lattice representation encoding different hypotheses for the word *manage*. The search scans the lattice for exact phoneme sequences corresponding to the query words. In the lattice in Figure 1, the sequence for *manage* is highlighted. Starting with the initial phoneme *m*, the scanner follows the path to the phoneme *ae*, and so on to the end of the word.

This search can control the lattice depth to give increased or reduced numbers of hypotheses at any point. The deeper the lattice, the more likely it is that the correct phoneme sequence is present. Unfortunately, this also leads to a greater number of false positives. Another disadvantage of this approach is that the lattice representation search cannot use inverse index structures effectively, thus the search costs increase linearly with the repository size.

Before the advent of large-vocabulary transcription systems, subword indexing provided an attractive alternative because it was computationally practical, and it also facilitated development of open-vocabulary indexing systems with only modest amounts of training data. Even today, this approach offers advantages for languages such as German that freely generate new word compounds, increasing the vocabulary size dramatically. In this case, many words will fall outside the vocabulary of a word-based transcription system but will be available via the subword approach.

Speaker adaptation techniques

Speech recognition systems employ various methods to improve their baseline accuracy.[5] Speaker adaptation adjusts the parameters for an individual speaker's acoustic models. For example, commercial dictation systems often adapt a personal set of acoustic models by having users read a number of sentences aloud in an enrollment session. The system can continue to modify the parameters during actual use, gradually producing an additional small improvement in recognition accuracy.

While applying speaker adaptation to a monologue from a single speaker is straightforward, applying it when multiple speakers are taking turns as in a meeting or an interview is more complex. In this case, the retrieval system can only apply speaker adaptation after a preprocessing stage that segments the audio stream to indicate speaker changes and clusters the segments for individual speakers.

Speaker segmentation. The various speaker segmentation strategies include a simple recognition system that isolates speech from nonspeech audio—background music, breath sounds, lip-smacking—and then uses a sequence of subword models to differentiate male and female speakers. The system then makes an additional pass over the data to identify changes between individual speakers. This procedure uses a likelihood ratio test to determine whether the same speaker produces adjacent recognized phonemes.

Speaker clustering. Clustering links together and uniquely labels all segments from the same speaker so the system can apply speaker adaptation methods. Clustering improves the effectiveness of speaker adaptation methods by increasing the amount of data available for retraining the acoustic models.

Speaker identification. Combining speaker adaptation with speaker identification improves recognition accuracy. The system identifies known individual speakers within the audio stream and uses previously trained acoustic models for these speakers to improve recognition accuracy.

MULTIMEDIA RETRIEVAL SYSTEMS

Over the past decade, these IR technologies have been incorporated into a number of research multimedia retrieval systems.

CMU's Informedia project

Carnegie Mellon University's pioneer Informedia multimedia retrieval project focused equally on speech- and video-processing technologies.[6] Informedia used closed-caption transcriptions extracted from CNN broadcasts to build a text index. For CNN programs without closed-caption transcriptions, the project used the Sphinx III speech recognizer. Informedia used several video analysis modules to extract facial features, text appearing

on the screen, video shot boundaries, and so on and then combined these information sources into a final index.

The successor project, Informedia II, continues to focus on novel audio- and video-processing technologies while devoting attention to the problem of presenting and organizing multimedia content in an effective and user-friendly way.

Cambridge University projects

The Video Mail Retrieval (VMR) and Multimedia Document Retrieval (MDR) projects at Cambridge University explored a variety of techniques for retrieving spoken documents. VMR demonstrated successful message retrieval using an interactive open-vocabulary search with phone lattice indexing on a collection of five hours of spoken messages in the Medusa networked multimedia system.[4] Experimental results showed that combining this indexing method with speaker-independent acoustic models produced retrieval precision performance of approximately 75 percent of the performance achieved using a perfect manual transcription.

MDR, the successor to VMR, focused on developing effective techniques for information retrieval from large collections of news broadcasts.[3] The MDR project used the Hidden Markov Model Tool Kit large-vocabulary speech recognition system for indexing. MDR combined the indexing output with retrieval enrichment and enhancement methods to achieve retrieval precision comparable to that obtained using manual document transcriptions.

IBM's CueVideo

IBM Almaden Research Center's CueVideo consists of a client-server video retrieval and browsing system and an automatic multimedia-indexing system. Like Informedia, the video retrieval system automatically detects shot boundaries, generates a shot table, and extracts representative key frames to generate compact, browsable input video summaries.

For audio processing, IBM's ViaVoice Recognition system uses acoustic and language models specially tuned for broadcast news and then performs text analysis and information retrieval. This process creates multiple searchable speech indexes, including an inverted word index, a phonetic index, and a phrase glossary index. The system generates a phonetic transcription of the input audio and selects overlapping triphone and quadphone sequences as subword index terms.[7] Then it augments this phone sequence representation with additional phone sequences derived from the transcription.

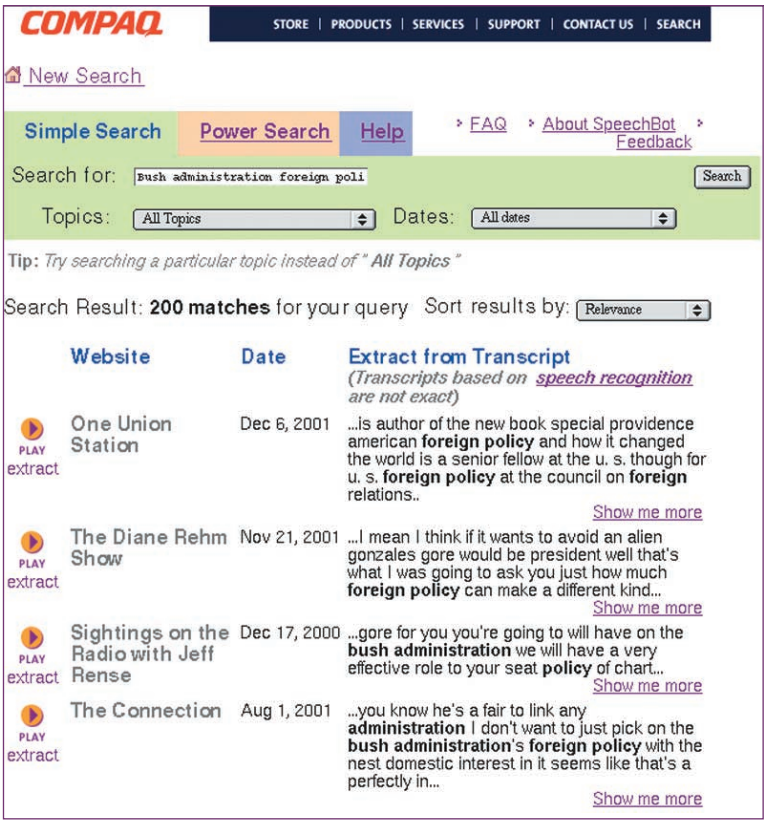

Figure 2. SpeechBot search result. The search uses word transcriptions to provide a catalog of audio and video documents that feeds the user interface a list of documents matching user queries. The user can click on a* play extract *button to play the multimedia stream that pronounces the query words.

Compaq's SpeechBot

SpeechBot from Compaq's Cambridge Research Laboratory is a general tool for audio and video indexing.[8] SpeechBot handles large volumes of speech recognition and user query data. The system fetches audio and video documents from the Web or an intranet and uses a large-vocabulary, continuous speech recognition system to process the audio data.

Figure 2 shows a typical SpeechBot search result: By clicking on the *play extract* button, the user can play the multimedia stream in which the query words are pronounced.

If an audio transcription is available, the system can replace the speech recognition module with an aligner module that provides time marks for each spoken word. SpeechBot uses word transcriptions to provide a catalog of audio and video documents that feed the user interface a list of documents matching user queries. The indexer also retrieves the word locations of the matches within a document. It uses IR techniques to calculate and sort the matches according to relevance.

SpeechBot does not serve content. Rather, it keeps a link to the original document much like traditional search engines such as AltaVista. This

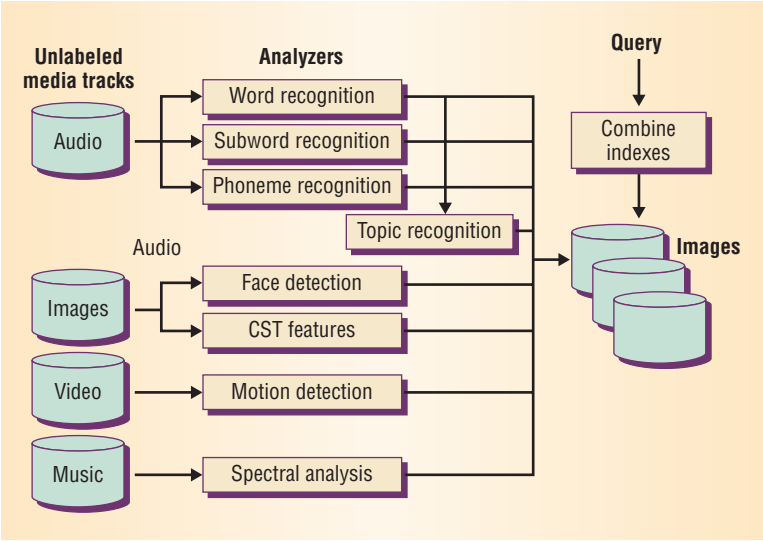

Figure 3. Multimedia analyzers. The system applies three analyzers to the audio stream: a conventional large-vocabulary, continuous speech recognizer; a syllable recognizer; and a phonetic recognizer that analyzes the audio signal.

search index has been running since December 1999 (http://www.speechbot.com), and it currently indexes more than 13,000 hours of Web audio and video content. SpeechBot achieves 90 percent precision on the top five hits returned by the system and close to 80 percent for the top 20 hits.

BEYOND MULTIMEDIA RETRIEVAL

Clearly, current multimedia retrieval systems—though useful and strong first steps in the right direction—are still far from being KM systems. To evolve into KM systems, multimedia retrieval systems need several improvements.

Multimedia analyzers

Many annotations are available for use during multimedia-content production. For example, video editors and producers have a short summary describing each video shot, they know its origin, beginning and ending time, and production time. They often create their own annotations, and sometimes they even produce closed captions before or during the video broadcast. Unfortunately, annotations of this type are often lost, they represent only a partial description of the multimedia details, and producing them is expensive because they require human intervention. Clearly, there is a need for more automated data analysis.

Because of multimedia's multitrack nature—which includes video, audio, and text—information retrieval systems may use several additive and complementary analyzers. These analyzers are additive in that each one works on different tracks, and they are complementary in that the system can apply multiple analyzers to the same track.

Figure 3 shows several possibilities for analyzing different tracks. In this example, the system applies three analyzers to the audio stream. The first is a conventional large-vocabulary, continuous speech recognizer. The second is a syllable recognizer that uses an associated trigram-syllable language model to automatically recognize speech. Finally, a phonetic recognizer analyzes the audio signal. Each of these speech recognizers works at different time resolutions with different constraints, providing a different view of the audio-speech signal.

Other forms of audio analysis use speaker recognition technology to identify the speaker or audio scene analysis to classify the audio as belonging to broad categories such as clean audio, telephone bandwidth, or background music. Systems such as BoogeeBot[9] represent a first step toward attempting to identify music segments and perhaps even to find similar sounding songs.

IR systems can also use face recognition algorithms that extract color, shape, or texture (CST) features to analyze video keyframes so that query-by-example methods can find similar images later on. In general, these systems use specific detectors—motion detectors, shot detectors, keyframe extractors, and text analyzers—to annotate the video stream and provide enriched metadata.

Unsolved problems

Current speech recognition analyzers cannot hypothesize out-of-vocabulary words, which often hold the most significance. One approach to solving this problem combines word and subword recognizers to hypothesize OOVs. Furthermore, subword recognizers also provide an intermediate representation that the system can analyze after the fact and insert into the word index at a later time.

Another approach combines analyzers to improve performance. For example, the analyzer can search the word or syllable recognizer's textual output for particular topics and segment the audio—via its textual representation—into coherent stories. Similarly, the system can combine a face recognizer with a speaker detector to improve identification.

Methods as simple as majority-voting schemes can combine all these knowledge sources to help reduce the errors that individual methods introduce. More sophisticated techniques based on data-fusion methods or Bayesian combination of knowledge sources provide further opportunity for improvement.

Representing information

Metadata is an intermediate representation describing multimedia content, whether simple

hand-generated textual annotations or specific features a content analyzer extracts. Metadata describes the document's nature and the relations among the document's different parts. The analyzer can search or index this compact representation to retrieve information. Examples of metadata include word or phoneme transcriptions of spoken content, topic segmentation, or CST feature vectors for images.

Because metadata represents new descriptive information, the question is how to represent and store this data. De facto standards such as XML facilitate indexing and quick access for information retrieval purposes as well as easy data exchange between system components.

Several initiatives define standards for multimedia content representation. Examples include

- SMIL—With the XML-based Synchronized Multimedia Integration Language, users can describe the temporal behavior and layout of multimedia presentations and coordinate their timing.
- METS—Promoted by the Library of Congress, the Metadata Encoding and Transmission Standard provides a standard set of multimedia attributes.
- DCMI—The Dublin Core Metadata Initiative provides recommendations and standards that primarily focus on the traditional publishing industry but also include many elements for use in describing multimedia documents.
- MPEG-7—The standard from the Motion Picture Experts Group for describing and searching audio and visual content supports a broad range of applications.

Figure 4 shows an example of how multiple analyzers provide different views of a multimedia document.

Indexing metadata

Complex querying and information retrieval requires storing metadata in a database. Current database technology such as Oracle is expanding to model XML data representations, and new native XML database technology is emerging. Tamino (http://www.softwareag.com/tamino/) and XYZFind (http://www.xyzfind.com/) are two examples of native XML databases.

Indexing metadata in its native form is a challenging problem—largely due to the inherent inaccuracy of the information that content analysis generates, but also because of the metadata's nature.

```
- <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 showing how multiple analyzers provide different views of a multimedia document. The different views assist in indexing and information retrieval and facilitate data exchange between system components.

If the metadata has a textual form, retrieval systems can use text-based indexes such as those that index the Web. If, on the other hand, the metadata is a high-dimensional feature vector extracted from multimedia, such as CST features, using an indexing approach is more difficult.

Often, a metadata search is limited to approaches that scan the entire database. Query-by-example image search engines such as IBM's QBIC are the most commonly used approach to this problem.[10]

TOWARD KNOWLEDGE MANAGEMENT

These architectures provide the tools for taking the initial steps to build effective KM systems. For example, research teams at IBM, CMU, and other institutions are investigating methods for analyzing spoken discourse to provide meeting support. Meetings generate a wealth of information that is usually lost. Relying on note takers to capture a meeting is a time-consuming and often fallible process. Additionally, meeting participants may not have easy access to all the information they require.

Systems such as IBM's MeetingMiner[11] use speech recognition to capture, analyze, and transcribe audio recordings of meetings. These transcriptions can bring important information to the participants' attention. For example, an analyzer can search a patent database and alert the participants in a technical meeting to similar intellectual property that a competitor owns. 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.

Researchers at CMU[12] have built a similar system for tracking, categorizing, and summarizing meetings. CMU's system includes 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 nonverbal cues based on video analysis to rapidly review records of human interaction.

Multimedia data introduces several challenges to knowledge management systems, including the uncertainties associated with media analyzers and the need for good scalability and effective user interfaces. Architectures capable of handling system complexity will also play a crucial role in deploying multimedia-based KM solutions.

Nonetheless, the prospects for fully exploiting multimedia content are promising. The experience gained in developing multimedia retrieval systems such as SpeechBot shows that even with current limitations in speech recognition technology, analyzers can achieve good performance when searching multimedia sources. Given current trends in audio and video analysis, multimedia storage and distribution over the Internet, developments in XML representations, and integration with knowledge portals, we expect multimedia data to become truly pervasive and as important, if not more so, than textual sources in KM systems. ■

References

1. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.
2. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Upper Saddle River, N.J., 1993.
3. J.S. Garafolo, C.G.P. Auzanne, and E.M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *Proc. 6th RIAO Conf.* (Content-Based Multimedia Information Access), Center for the Advanced Study of Information Systems, Paris, 2000, pp. 1-20.
4. M.G. Brown et al., "Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval," *Proc. ACM Multimedia*, ACM Press, New York, 1996, pp. 307-316.
5. F. Kubala et al., "Integrated Technologies for Indexing Spoken Language," *Comm. ACM*, vol.43, no. 2, 2000, pp. 48-56.
6. H.D. Wactlar et al., "Lessons Learned from Building a Terabyte Digital Video Library," *Computer*, Feb. 1999, pp. 66-73.
7. S. Srinivasan and D. Petkovic, "Phonetic Confusion Matrix-Based Spoken Document Retrieval," *Proc. 23rd Int'l Conf. Information Retrieval,* ACM Press, New York, 2000, pp. 81-87.
8. J-M. Van Thong et al., "SpeechBot: A Speech Recognition-Based Audio Indexing System for the Web," *Proc. 6th RIAO Conf.* (Content-Based Multimedia Information Access), Center for the Advanced Study of Information Systems, Paris, 2000, pp. 106-115.
9. B. Logan and A. Salomon, "A Music Similarity Function Based on Signal Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, Piscataway, N.J., 2001, pp. 952-955.
10. J. Ashley and M. Flickner et al., "The Query by Image Content (QBIC) System," *Proc. ACM Sigmod Conf.*, ACM Press, New York, 1995, p. 475.
11. E.W. Brown et al., "Towards Speech as a Knowledge Resource," *IBM Systems J.*, vol. 40, no. 4, 2001, pp. 985-1001.
12. A. Waibel et al., "Meeting Browser: Tracking and Summarizing Meetings," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Morgan Kaufmann, Lansdowne, Va., 1998, pp. 281-286.

Pedro J. Moreno *is a senior member of the technical staff at Compaq Computer Corp. Cambridge Research Laboratory. His research interests are multimedia indexing, speech recognition, and applications of machine learning. He received a PhD in electrical and computer engineering from Carnegie Mellon University. He is a member of the IEEE and ESCA. Contact him at pedro.moreno@compaq.com.*

J-M. Van Thong *is a senior member of the technical staff at Compaq Computer Corp. Cambridge Research Laboratory. His research interests are speech recognition, multimedia indexing, and user interfaces. He received a PhD in computer science from the Pierre and Marie Curie University in Paris. Contact him at jm.vanthong@compaq.com.*

Beth Logan *is a member of the technical staff at Compaq Computer Corp. Cambridge Research Laboratory. Her research interests are acoustic modeling and indexing of speech and music. She received a PhD in electrical engineering from the University of Cambridge. She is a member of the IEEE. Contact her at beth.logan@compaq.com.*

Gareth J.F. Jones *is a lecturer in the Department of Computer Science, University of Exeter. His research interests include multimedia, cross-language, and context-aware retrieval. He received a PhD in electrical and electronic engineering from the University of Bristol. He is a member of the Institute of Electrical Engineers. Contact him at G.J.F.Jones@exeter.ac.uk.*