Toward speech as a knowledge resource

by E. W. Brown
S. Srinivasan
A. Coden
D. Ponceleon
J. W. Cooper
A. Amir

Speech is a tantalizing mode of human communication. On the one hand, humans understand speech with ease and use speech to express complex ideas, information, and knowledge. On the other hand, automatic speech recognition with computers is very hard, and extracting knowledge from speech is even harder. Nevertheless, the potential reward for solving this problem drives us to pursue it. Before we can exploit speech as a knowledge resource, however, we must understand the current state of the art in speech recognition and the relevant, successful applications of speech recognition in the related areas of multimedia indexing and search. In this paper we advocate the study of speech as a knowledge resource, provide a brief introduction to the state of the art in speech recognition, describe a number of systems that use speech recognition to enable multimedia analysis, indexing, and search, and present a number of exploratory applications of speech recognition that move toward the goal of exploiting speech as a knowledge resource.

Today's arsenal of tools to address the knowledge management problem typically includes relational databases, text search engines, document clustering and classification tools, and knowledge portals. Relational database technology has existed since the early 1970s and provides rich, robust access to structured data—data easily represented as fields in tables (e.g., employee records with name, birth date, address, and salary fields). Text search technology has been around for nearly as long, but has only recently come into mainstream prominence as a tool for finding documents on the World Wide Web. Document clustering and classification tools go beyond text search by automatically organizing collections of documents into subgroups of related documents, allowing a user to see a summary of the information contained in the document collection and quickly navigate to documents of interest. Knowledge portals provide a single interface to all of these data management and access tools and may additionally provide access to other enterprise applications.[1]

Although these tools provide a tremendous amount of functionality and sophisticated features, they work primarily with *explicit knowledge*, or knowledge that can be formally expressed in procedures, rules, manuals, documents, etc., and easily transmitted from one person to another. An equally important kind of knowledge in any organization is *tacit knowledge*, or knowledge that is based on the personal experience, know-how, values, and beliefs of the individuals who possess the knowledge. Tacit knowledge does not easily lend itself to structured data representations or even articulation in documents, limiting the ability of structured data and document management tools to support tacit knowledge management.

For tacit knowledge to be useful to anyone beyond the individual who possesses it, it must be transferred to other individuals through the process of socialization, or converted to explicit knowledge through the process of externalization.[2] Socialization occurs when individuals get together and share their expe-

riences and expertise through discussions, while externalization occurs when expertise and experience are codified in the form of rules, procedures, or even documents that record those experiences with written language. Socialization and externalization are complex, cognitive, human interaction activities, and we are a long way from automating these activities with a computer. To significantly improve upon current knowledge management tools, however, we must move in that direction.

Toward that end we propose the following approach: use microphones to record spoken discourse during situations where tacit knowledge is likely to be revealed, and apply automatic speech recognition to convert the spoken discourse into a text transcript. For example, a computer services consulting group might meet after the close of a customer engagement to discuss what went right, what went wrong, what was learned, and how the next engagement might be conducted better. At a minimum, automatically creating a text transcript of this meeting is a crude approximation of the externalization process whereby experience is codified in a written document. In the future, text analysis and natural language processing techniques might be used to automatically summarize the transcript and distill higher-level concepts from the discourse.

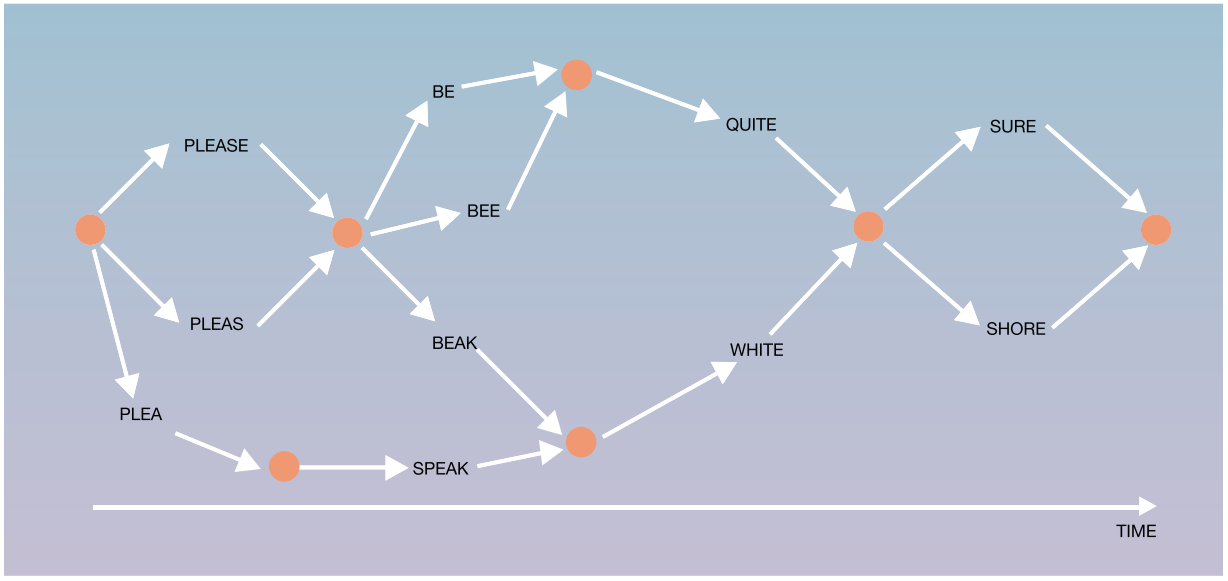
The viability of this approach depends on the ability of automatic speech recognition to generate sufficiently accurate transcripts in a variety of acoustic environments and discourse situations. In this paper, we explore the capabilities of speech recognition in a variety of applications related to knowledge management. We begin with a brief tutorial on speech recognition and a review of the various kinds of speech recognition applications. Next, we review a number of specific research projects that apply automatic speech recognition (among other technologies) to the problem of video analysis, indexing, and retrieval. We then describe applications of automatic speech recognition that are more exploratory, covering an area loosely called "speech mining." This work is representative of the overall approach we recommend for solving the problem of managing tacit knowledge. Finally, we offer some concluding remarks.

Speech recognition applications and concepts

Speech recognition applications may be seen as belonging to one of three categories: dictation or document creation systems, navigation or transactional systems (e.g., automated voice response systems), and multimedia indexing systems. In dictation systems, the words spoken by a user are transcribed verbatim into text to create documents such as personal letters, business correspondence, etc. In navigation systems, the words spoken by a user may be used to follow links on the Web or to navigate around an application. In transactional systems, the words spoken by a user are used to conduct a transaction such as a stock purchase, banking transaction, etc. In multimedia indexing systems, speech is used to transcribe the audio (possibly extracted from the video) into text, and subsequently, information retrieval techniques are applied to create an index with time offsets into the audio. Advances in technology are making significant progress toward the goal of allowing any individual to speak naturally to a computer on any topic and have the computer accurately understand what was said. However, we are not there yet. Even state-of-the-art continuous speech recognition systems require the user to speak clearly, enunciate each syllable properly, and have his or her thoughts in order before starting. Factors inhibiting the pervasive use of speech technology today include the lack of general-purpose, high-accuracy continuous-speech recognition, lack of systems that support the synergistic use of speech input with other forms of input, and challenges associated with designing speech user interfaces that can increase user productivity while being tolerant of speech recognition inaccuracies.

Speech recognition systems are typically based on hidden Markov models (HMMs),[3] which are used to represent speech events (e.g., a word) statistically, and where model parameters are trained on a large corpus of speech data. Given a trained set of HMMs there exists an efficient algorithm for finding the most likely word sequence when presented with unknown speech data. The recognition vocabulary and vocabulary size play a key role in determining the accuracy of a system. A vocabulary defines the set of words or phrases that can be recognized by a speech engine. A small vocabulary system may limit itself to a few hundred words, whereas a large vocabulary system may consist of tens of thousands of words. Large vocabulary speech recognition systems typically use a subword approach where phonetic subword models are built instead of an explicit model for each word in the large vocabulary. Such systems also use a statistical language model that defines likely word sequences in a particular domain to provide statistical information on word sequences. The

Figure 1 Recognition of phrase “please be quite sure” in a word lattice representation



language model assists the speech engine in recognizing speech by biasing the output toward high probability word sequences. Together, vocabularies and language models are used in the selection of the best match for a word by the speech recognition engine. Dictation systems are typically large vocabulary applications, whereas navigation and transactional systems are typically small vocabulary applications. Multimedia indexing systems could be either large vocabulary or small vocabulary applications.

Speech recognition systems provide the most probable decoding of the acoustic signal as the recognition output, but keep multiple hypotheses that are considered during the recognition. The multiple hypotheses at a given time, often known as N-best word lists, provide grounds for additional information that may be used by an application. Recognition systems generally have no means to distinguish between correct and incorrect transcriptions, and a word lattice representation (a directed acyclic graph) is often used to consider all hypothesized word sequences within the context. Figure 1 shows a word lattice representation for the hypothetical recognition of the phrase “please be quite sure” together with the multiple hypotheses considered during recognition. The nodes represent points in time, and the arcs represent the hypothesized word with an associated confidence level (not shown in the figure). The path with the highest confidence level is generally provided as the final recognized result, often known as the 1-best word list. The N-best word lists are typically used by speech recognition applications to improve the usability and performance of applications such as dictation systems and multimedia indexing systems.

Speech recognition accuracy is typically represented in terms of word error rate (WER), defined to be the sum of word insertion, substitution, and deletion errors divided by the total number of correctly decoded words. The WER can vary dramatically depending on the nature of the audio recordings. Recent algorithmic advancements for a large vocabulary speech recognition task known as Voicemail (conversational telephone speech from a single speaker) have resulted in a lowered WER of 28 percent.[4] The WER for a large vocabulary speech recognition task known as Broadcast News on prepared speech (as opposed to spontaneous speech) from anchors in a studio is reported to be around 19 percent.[5] These numbers reflect speech recognition benchmark evaluations on selected evaluation data. In general, for a wide variety of real-world speech data that includes combinations of speech with background noise, degraded acoustics, and non-native speakers on a real-time speech recognition system, the WER can vary between 35 and 65 percent.[6–9] Retrieval on transcripts with

WER of 10–30 percent have been reported to yield an average precision of 0.6–0.7 on test collections of a few tens of hours. However, for real-world audio with high WER of 60–70 percent, the precision and recall have been reported to drop dramatically to 0.17 and 0.26, respectively.[6] Much of the work in this field was promoted by the Text REtrieval Conference (TREC), supported by NIST (National Institute of Standards and Technology). It has promoted the research in spoken document retrieval by establishing benchmarks, organizing conferences, and creating an international competitive-collaborative research environment.[7,8,10,11]

In summary, the specific speech applications such as dictation, transactional, and indexing applications are driven by distinct requirements for vocabularies and vocabulary sizes, and by recognition accuracy. Given the state of the art in recognition technology, some applications lend themselves more favorably to the successful use of speech recognition.

Dictation applications. Systems that support the synergistic use of speech and direct manipulation using keyboard or mouse hold the appeal of improved usability by providing a more natural interface to the computer. Members of the medical, legal, and journalism professions have been using speech recognition to create their documents in order to eliminate transcription time and improve productivity. Dictation applications are aimed at text entry or document creation, such as dictation for electronic mail or word processing. Depending on the specific application, a particular language model may be selected in order to obtain the best accuracy. Within a language domain such as general English, perplexity is a measurement of the number of equally likely word choices given a sequence of words. In a high perplexity domain such as general English, it is more difficult to predict a word given its preceding words in a sentence, due to the large number of equally likely words that may follow. In contrast, the radiology domain, for example, has a lower perplexity than general English, which leads to higher accuracy. As a result, practitioners of health care and other specialized areas in medicine have looked favorably at integrating the use of speech recognition, since eliminating transcription time and cost matches their drive to improve productivity. This class of applications has the most stringent requirements for recognition accuracy of 95 percent or higher, since a human user has to manually correct recognition errors. Cost and time benefits due to eliminating transcription costs may be achieved only if the accuracy is high enough to require minimal error correction. A key speech interface design issue has to do with correction of recognition errors in order to maximize throughput.[12]

Navigation or transactional applications. Navigation or transactional systems such as telephony applications typically use a constrained set of words or grammars for a particular application. Speech grammars are an extension of the single words or simple phrases supported by vocabularies. A speech grammar is a structured collection of words and phrases bound together by rules. These rules define the set of speech tokens that can be recognized by the speech engine at a given point in time. Call centers and help-desk functions such as airline schedules, telephone help centers, and trading companies have started to take advantage of speech recognition technology by developing small-vocabulary applications intended to guide a user through a selection process based on certain keywords. Such applications are naturally more resilient to speech recognition errors, since they are primarily intended to navigate through a structured set of questions leading to one of a small number of states. The switching of states is triggered by the user speaking one of several trigger words valid in a given state. The requirements for high speech-recognition accuracy are not as stringent, because a well-designed constrained vocabulary is capable of achieving the desired goal—switching state in order to navigate through a transaction.

Multimedia indexing applications. Indexing applications provide the ability to retrieve relevant audio or video segments when presented with a textual query. This is usually done by indexing the output of speech recognition using words or subword units as index terms. This has been referred to as "spoken document retrieval" in scientific forums. A well-known issue in spoken document retrieval is the concept of in-vocabulary terms and out-of-vocabulary terms. Since the speech engine matches the acoustics from the speech input to words in the vocabularies, only words in the vocabulary are capable of being recognized. Words in a vocabulary are recognized based entirely on their pronunciations. Words can have multiple pronunciations; for example, "the" will have at least two pronunciations, "thee" and "thuh." Punctuation symbols are often associated with several different verbal representations; for example, the symbol "." may be referred to as "period," "point," or "dot." In addition, a word not in the vocabulary will often be erroneously recognized as

an in-vocabulary word that is phonetically similar to the out-of-vocabulary word.

To address these problems, researchers have explored the use of subword representations based on phonemes as index terms, with varying degrees of success.[13–16] A phoneme is defined to be any of the abstract units of the phonetic system of a language. Phonemes correspond to a set of similar speech sounds, which are perceived to be a single distinctive sound in the language. The accuracy of phone recognition is limited, particularly in the case of short words.[14,15] However, for the purpose of retrieval of out-of-vocabulary words or where the confidence level associated with the recognized words is low, there is considerable benefit in combining phonetic information with the words for indexing. The use of phonetic retrieval also makes "sounds-like" retrieval applications possible.

Multimedia indexing research projects

Retrieval of structured data is often based on the values of entries in the fields that format these records; in contrast, retrieval of unstructured multimedia data requires content-based retrieval whereby the contents of the data are examined for the presence or the absence of an object, of words or phrases, or of a visual action.

Multimedia retrieval approaches have been classified into expression-based and semantic approaches. Expression-based techniques rely on an example or a physical description of the information that is sought, and semantic approaches rely on the actual content of the media.[17] Another characterization of multimedia retrieval is based on query formulation and data representation. A query-by-example technique, such as an image query to search an image database[18] or an audio query to retrieve a segment of music, exemplifies an expression-based technique whereby the query is specified in the same format as the data. In contrast, the choice of a different format for query formulation from the data representation, such as a physical description of the required information or a textual query to search an audio or video database, exemplifies a semantic technique. Clearly, the ultimate goal in multimedia retrieval is to achieve semantic retrieval using fully automated techniques. In recent times, perhaps the most significant technological advance toward this semantic ideal has been the application of automatic speech-recognition technology to multimedia content that contains speech in the audio portion.

Several researchers have been working on multimedia indexing techniques for extraction of semantic information from unstructured video. This includes a wide range of topics from computer vision, pattern recognition, video analysis and summarization, speech recognition, natural language understanding, and information retrieval.[13,19–22] One approach to video retrieval is to apply image-retrieval techniques to key frames extracted from the video. In this approach, an image is posted as a query, and similar images are retrieved. There are two reasons why this approach, in general, has not yet become popular. First, in most practical situations the user does not have such an image available to formulate the query. Second, the state of the art in content-based image retrieval has not yet reached the semantic level desired by most users. Rather, it is typically done in a feature space, such as color histograms, color layout, color blobs, texture, and shape.[18] A more popular approach to video retrieval is to search the audio transcript using the familiar metaphor of free text search.[13] In this case, speech recognition is applied to the audio track, and a time-aligned transcript is generated. The indexed transcript provides direct access to the semantic information in the video.

Although searching the audio using free text proves to be almost as efficient as text searching, browsing the video is much more time consuming than browsing the text. This is because the user has to play and listen to each of the retrieved videos, one by one. This is unlike the situation with text, where a quick glance at the result page is often sufficient to filter the information. In the case of video, it is more efficient to browse the visual part, e.g., the video storyboard. A few pages of storyboard, each showing ten or more key frames, can cover one hour of presentation. Usually these show the slides presented during the talk.

We next describe briefly three representative research projects on multimedia retrieval that include a strong speech retrieval component. In addition, we mention SpeechBot: a search engine for audio and video content that currently has indexed 8427 hours of content.[23]

Informedia. The Informedia research project[21,24] has created a terabyte digital library in which automatically derived descriptors for the video are used for indexing, segmenting, and accessing the library contents. It combines speech recognition, image processing, and natural language understanding techniques for processing video automatically to produce a video

"skim," which reduces viewing time without losing content. It offers three ways to view the search results: poster frames, filmstrips, and skims. The poster frame view presents search results in poster frame format with each frame representing a video "paragraph." The filmstrip view reduces the need to view each video paragraph in its entirety by providing storyboard pages for quick viewing. The most relevant subsections of the video paragraph are displayed as key scenes, and key words are clearly marked. Combined word and phonetic retrieval has also been explored in the Informedia project[16] where an inverted index for a phonetic transcript comprising phonetic substrings of three to six phones have been used. During retrieval, the word document index and phonetic transcription index are searched in parallel and the results are merged. Experiments on a corpus of about 500 ABC News and CNN (Cable News Network) stories using combined word and phone indexes resulted in an average precision of .67 with an overall performance of 84.6 percent of that of a full-text retrieval system. However, for real-world audio with high WER of 70–80 percent, the precision and recall have been reported to drop dramatically to 0.17 and 0.26, respectively.[6]

Medusa. Cambridge University, in collaboration with Olivetti Research Laboratory (ORL), has developed the Medusa networked multimedia system, which is in use on a high-speed ATM (asynchronous transfer mode) network for a video mail application based on word-spotting using a 35-word indexing vocabulary chosen *a priori* for the specific domain. They have developed retrieval methods based on spotting keywords in the audio sound track by integrating speech recognition methods and information retrieval technology to yield a practical audio and video retrieval system.[25] They used the phone-lattice scanning approach, where the speech recognition system generates a generalized subword, or phone lattice. Spoken words can be decomposed into a sequence of phone units and the precomputed lattices are scanned for phone strings corresponding to the query word. Subsequently, large vocabulary speech recognition-based indexing has been combined with phone-lattice scanning to yield 82–85 percent relative precision compared to perfect text retrieval, and has proved to be better than either method alone.[13] These results are based on a test collection of 300 messages used in a video mail retrieval application.

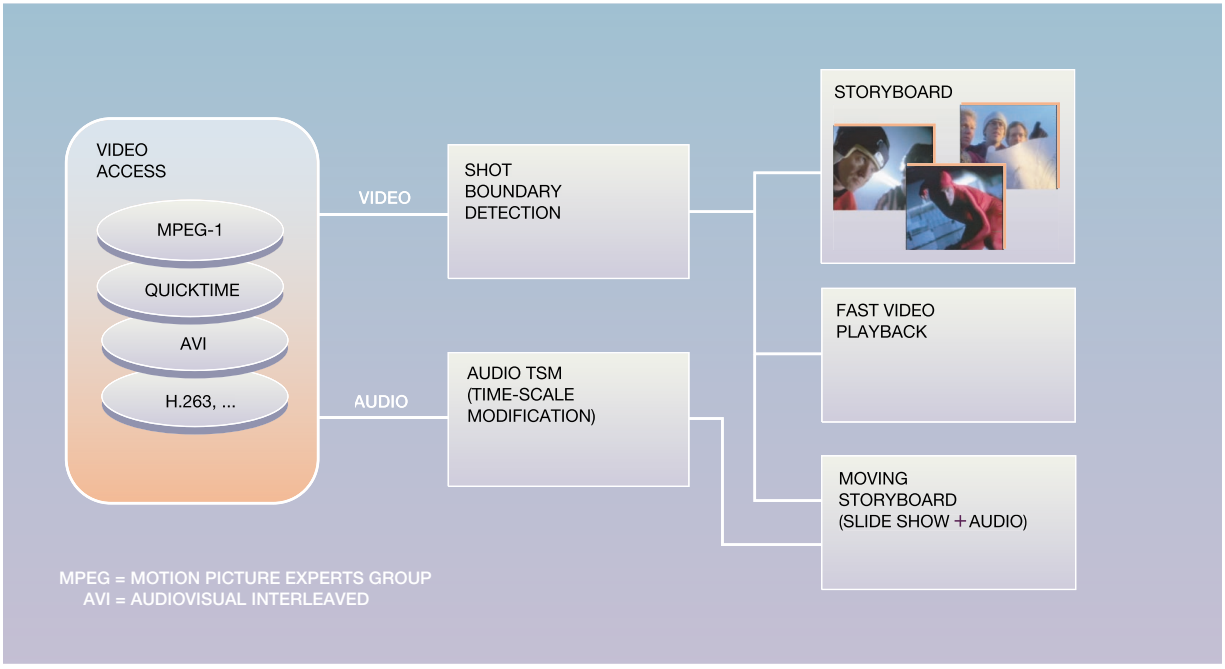
CueVideo. CueVideo is a research project at IBM Almaden Research Center that consists of an automatic multimedia indexing system and a client/server video retrieval and browsing system. Their approach to multimedia retrieval is "*Search the speech, browse the video*." The video and audio are two parallel media streams of information that are related by a common time line. Therefore, they take advantage of the two parallel streams, using the audio stream for search and the video stream for quick visual browsing in a complementary manner to provide the desired video search functionality. The video indexing automatically detects shot boundaries, generates a shots table, and extracts representative key frames as JPEG (Joint Photographic Experts Group) files from each of the shots. Several browsable video summaries are generated using the indexing architecture shown in Figure 2.

The audio processing starts with speech recognition (using the IBM Speech Recognition system with Broadcast News models[4]) followed by text analysis and information retrieval. Several searchable speech indexes are created, including an inverted word index, a phonetic index, and a phrase glossary index. Another segment of the audio processing generates the time-scale modification (TSM) audio in desired speedup rates for the fast and slow moving storyboards.[26] A phonetic transcription of the input audio is generated and overlapping triphone and quadphone sequences are selected as subword index terms.[27] This phoneme sequence representation is augmented with additional phone sequences derived from the observed phonemes in the transcription and in the phoneme confusion matrix. Their results are based on a test collection assembled from a corporate training data set that is being used in a realistic environment for distributed learning. For query terms that are out-of-vocabulary and are greater than four characters long, phonetic retrieval in one hour of reasonable quality speech (35 percent WER for the in-vocabulary terms) results in an average recall of 88 percent and an average precision of 69 percent. For higher WERs, retrieval performance is lower. For in-vocabulary query terms, phonetic retrieval results in an average recall improvement of 17 percent with an average precision loss of 17 percent, even for high WERs, over word-based retrieval.

Speech mining research projects

The multimedia indexing and retrieval applications discussed thus far are aimed primarily at domains where the data are professionally produced audio or video stored on tape or disk and the main problem is to provide users with the ability to quickly find a particular clip, fact, or piece of information. Some

Figure 2 CueVideo indexing architecture



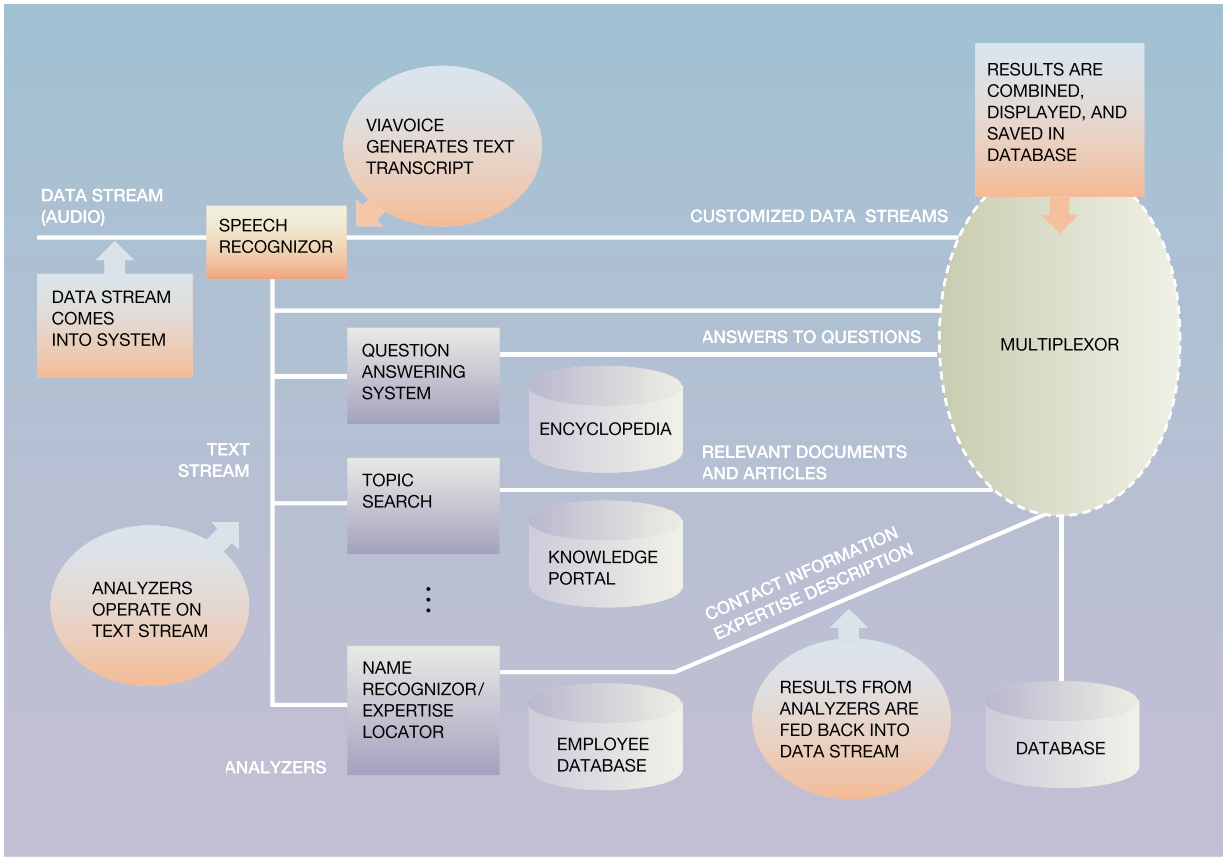
systems provide additional functionality, including automatic summarization, clustering, classification, and organization of the recorded information. These systems typically operate off line, require 2 to 10 times real time to analyze the recorded data, and generally assume high-quality produced recordings.

Although these systems can be invaluable for managing recordings, such as broadcast news or corporate training videos, they are not directly amenable to another rich, abundant source of information, namely spoken discourse. Spoken discourse includes any spoken conversation between two or more individuals that takes place anywhere, anytime. Spoken discourse is a rich source of tacit information, and often the only form in which the tacit information is instantiated. Until very recently, it was nearly impossible to capture and exploit this information. Recent advances in speech recognition technology, however, have allowed the exploration and development of a number of interesting technologies and applications that attempt to capture spoken discourse, convert the discourse to text, apply text analysis to the discourse, and exploit the knowledge discovered in the discourse. In the rest of this section, we explore some of these leading-edge applications and discuss their future prospects.

A recent research focus at the IBM Thomas J. Watson Research Center has been on how to capture spoken discourse and analyze it in real time to both extract knowledge from the discourse and provide additional, related knowledge to the discourse participants. The need for this capability is driven by three separate applications: meeting support, data broadcasting, and call mining. Meeting support in this context includes the ability to understand the current meeting discussion and automatically provide related information to the meeting participants. Data broadcasting is the process of exploiting the unused bandwidth in a television broadcast to send arbitrary data with the television program. In particular, the data should be related to the current television program and provide an enhanced viewing experience for the user, enriching the audio and visual television program with related facts, articles, and references. Call mining is an effort to analyze and index the telephone calls made or taken by customer service representatives at call centers and help desks. We consider these applications in more detail below.

WASABI. The first two seemingly unrelated applications (meeting support and data broadcast) share a common need for the ability to analyze speech in

Figure 3 WASABI architecture



real time and automatically discover relevant, collateral information. Toward that end, the researchers at Watson have built a generic framework for analyzing speech, called WASABI (Watson Automatic Stream Analysis for Broadcast Information).[28] WASABI takes speech audio as input, converts the audio stream into text using a speech recognition system, applies a variety of analyzers to the text stream to identify information elements, automatically generates queries from these information elements, and extracts data from the search results that are relevant to the current discourse. The overall architecture is shown in Figure 3.

The input to the system is the raw data stream, i.e., the captured audio of the discourse. The speech recognition system, IBM ViaVoice*, converts the audio into a text stream, which WASABI feeds to one or more *analyzers*. An analyzer performs a text analysis procedure on its input and produces an output that may be fed to another analyzer or multiplexed back into the original data stream. The task performed by each analyzer depends on the application in which the framework is applied. One of the more important tasks performed by an analyzer is to automatically create a query from the input, use the query to search a relevant knowledge repository, and extract relevant information from the search results that will enhance the input data stream.

Regardless of the task, the analysis must take place in real time, or fast enough to keep up with the incoming data stream. The final enriched data stream may be presented to the user in an appropriate user interface, or archived to a data store for indexing and future reference.

MeetingMiner. Meetings are an obvious source of knowledge in any organization. They occur regularly and in a variety of settings, and often suffer from two

problems. First, depending on the formality of the meeting, the content of the meeting will be captured with limited success. Meetings that are more formal may have actual minutes created by a designated scribe. Less formal meetings, however, will be captured only by the notes taken by the meeting participants, or worse, the collective memory of the participants. The second problem common to most meetings is that during the meeting the participants do not have convenient access to all of the knowledge resources that might be used to facilitate or enrich the meeting. These resources might be as simple as someone's phone number, or they might be more complex, such as a project database that identifies who is working on what and where expertise on a particular topic can be found.

The WASABI framework is being applied to solve these problems in a system called MeetingMiner. MeetingMiner is essentially an agent that passively captures and analyzes the meeting discussion and periodically becomes an active participant in the meeting, whenever it finds information that it determines is highly pertinent to the current discussion. The main input to the system is an audio stream generated by one or more microphones that capture the spoken discourse of the meeting. The audio stream is converted to a text transcript by the speech recognition system, and the text transcript is processed by the meeting analyzers.

The current set of meeting analyzers includes a named entity recognizer, a topic tracker, and a question identifier. The named entity recognizer identifies proper names in the text, such as people, places, and organizations. It is based on the algorithms developed in the Textract system,[29] and it uses a combination of lexical clues (e.g., capitalization patterns, punctuation, etc.) and dictionary lookups to identify named entities. The system uses the identified names of people to search an employee database and retrieve address, phone number, group affiliation, project responsibilities, and expertise information. This information is assembled in a custom address book as the meeting progresses, providing instant access to information about any individual who may be mentioned during the course of the meeting.

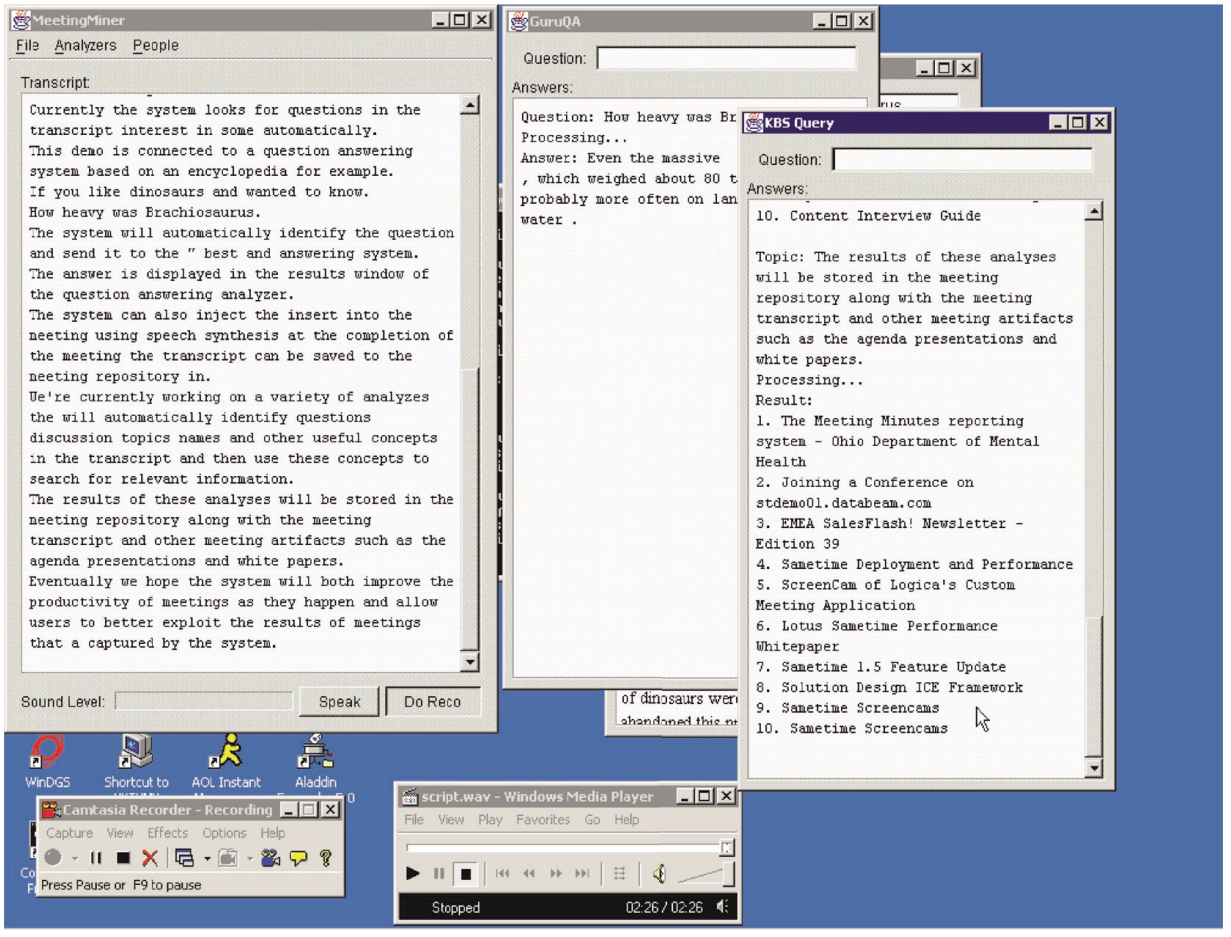
The topic tracker uses a combination of automatic text classification and statistical term frequency analysis to identify keywords in the text. The text classification system analyzes a sliding window of sentences and classifies the window content into a predefined taxonomy of topics. The current topic combined with the identified keywords is used to search related knowledge repositories and provide a continuously updated "hit list" of objects that may be relevant to the meeting. For example, during a technical design meeting, the topic tracker might send topic and keyword queries to a database of U.S. patents and alert the meeting participants whenever a highly relevant patent is found. This capability can enhance the design session with relevant information and help prevent time being spent on solutions that already exist, or worse, are owned by competitors.

The question identifier uses regular expressions to identify various kinds of questions in the text transcript. In particular, the system identifies *who*, *what*, *when*, *where*, *why*, and *how* questions and feeds them to a question-answering system.[30] The question-answering system parses the question and returns a concise phrase or small number of sentences that answer the question. This capability works best with questions seeking factual answers (e.g., "When are the budget numbers due?"), as opposed to more open-ended questions (e.g., "How can we improve profitability?"). Of course, the effectiveness of the question-answering component is limited by the information available for answering questions. Prager et al.[30] describe a system that automatically processes free-text documents and creates a question-answering system based on those documents, allowing any relevant collection of documents (corporate memos, e-mail messages, reports, etc.) to be included in the question-answering system.

The current MeetingMiner prototype is a Java** application that uses the Java** Speech API (application programming interface) to communicate with the speech recognition engine. A screen shot of the prototype is shown in Figure 4. The main window in the prototype displays the transcript from the current meeting, generated in real time. Each analyzer displays its results in a separate window, customized for the particular kind of information being displayed. Analyzers that generate time-sensitive data may use audible alarms (chimes or synthesized speech) to alert meeting participants to the existence of the highly relevant information. Additionally, the meeting transcript and analyzer output are archived in a format that can be indexed, searched, and "played back" later (with all of the analysis results synchronized with the playback).

The MeetingMiner system is still in the early stages of development and testing. A significant hurdle fac-

Figure 4 Screen from MeetingMiner prototype



ing the system is the performance of the speech recognition engine in the particularly challenging environment of a meeting, where audio quality is questionable, the discourse is broken and less grammatically correct, and there are multiple speakers, possibly (most likely) trying to speak simultaneously. Audio quality can be addressed to a certain extent by custom meeting rooms designed with attention to acoustics. Ideally, though, the system would place minimal requirements on the recording environment, allowing it to be portable and as nonintrusive to the meeting participants as possible. Speaker-independent speech recognition should improve as more effort is made to support speech recognition in these challenging multispeaker environments. In the meantime, progress is being made using separate speech models custom trained for each meeting participant.

Interest in the problem of capturing and supporting meetings is growing. Waibel et al.[31] from Carnegie Mellon University (CMU) describe a system for capturing, indexing, searching, and browsing meetings. Their work focuses on building speech recognition models suitable for the speaking modes found in meetings and applying postprocessing steps on the speech recognition transcripts to generate summaries. They also explore the use of visual cues captured by video camera to aid in tracking the discussion and to provide enhanced browsing capabilities. Researchers at Bolt, Beranek and Newman (BBN, now part of Verizon) have built a prototype system called Rough'n'Ready,[32] which uses speech recognition, speaker identification, topic detection, and named entity extraction to process and index video based on the audio track. All of the BBN technologies are based on hidden Markov models.

Rough'n'Ready was originally designed to index and organize broadcast news programs, but it could easily be adapted to support meetings. Both Rough'n'Ready and the system from CMU, however, process the audio off line after the meeting has completed and ignore the potential benefits of on-line analyses.

Of course, interest in supporting meetings predates the recent possibility of using speech recognition and automatic analysis. Moran et al.[33] have explored extensively the issues involved in capturing and exploiting meeting content. The same group has also investigated the use of electronic whiteboards in meetings to support collaboration and capture.[34] A more formal approach to meeting capture and analysis is taken by the IBIS system,[35,36] which defines a methodology for manually recording meetings and classifying statements made so that the decision-making process employed during a meeting can be more easily traced and understood. Ultimately this work on more formal methods of meeting analysis and the results of user interface studies must be incorporated into the new automatic meeting analysis systems.

Data broadcast. The second use of the WASABI framework is to support *data broadcasting*, which is the process of transmitting arbitrary data in the unused bandwidth of a television broadcast.[28] A high-definition television (HDTV) channel has over 1.5 megabits/second of bandwidth available to send data in addition to the audio and video program. This bandwidth can be used to send any data that the receiver is capable of processing, though a particularly interesting use of the bandwidth is to send collateral information related to the currently broadcast television program. For example, if the current broadcast is a news program, the data channel might contain text versions of related stories, biographies of people mentioned in the news, geographical information for places mentioned, or links (URLs, or uniform resource locators) to World Wide Web pages that contain additional information. This is clearly useful in a knowledge management context where information (e.g., news, training, reports, etc.) is distributed in the form of video programs. The WASABI analysis can automatically enrich the video with related facts and data from knowledge repositories of particular interest to the end user.

The overall processing flow for data broadcasting is very similar to that used in the MeetingMiner system. This is shown in Figure 5. The audio/video source is fed to a collection of real-time feature extractors that extract text and possibly visual features. The system sends these features to the event analyzers, which classify the features into topics, extract named entities (e.g., names of persons, places, dates, etc.), and combine these events into a data structure called the *knowledge chain*. The knowledge chain is essentially a linked list that assembles all of the events on a time line.

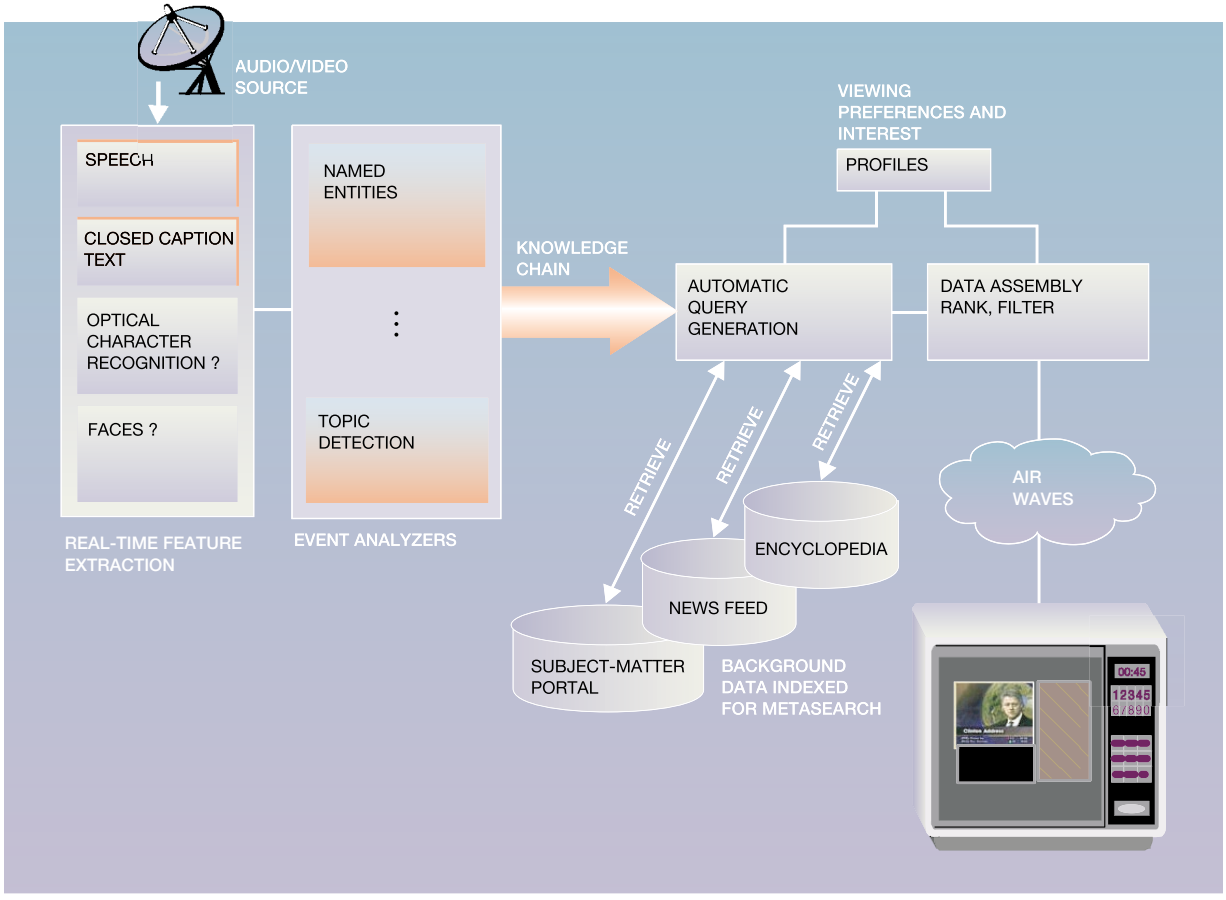
Once the knowledge chain has been created, the next step is to find the collateral information that will be broadcast with the program. This is done by automatically generating queries based on the events recorded in the knowledge chain. Profiles (either personal or application-specific) can be used to guide the query generation. The results from these queries are then assembled, ranked, and sent to the multiplexer, which inserts the results into the broadcast stream. The combined audio, video, and data channels are then broadcast to a receiver (e.g., a set-top box or a TV tuner card in a PC) and displayed in a user interface that shows the audio/video program along with the collateral information. The receiver may additionally have the capability to buffer or store the program, allowing it to be interrupted while the user explores the collateral information in more detail.

SAMSA. The SAMSA (Speech Analysis, Mining and Summary Application) project[37] is an experiment in the actual mining of outbound telephone sales calls using text mining techniques. Telephone calls from financial consultants were recorded on digital recording tape over a period of several weeks, and then processed using the IBM Research version of the speech recognition engine optimized for telephone calls and speaker-independent recognition.[4] From about 11 000 call units, including empty calls, 529 calls of substantive length for speech recognition processing were selected.

The TALENT text mining processes have previously been described.[38] Briefly, the text mining system recognizes multiword names,[29] locations, and organizations,[39] and detects relationships between terms based on proximity and by common English language patterns, such as appositives and parentheticals. In addition, a context thesaurus can be constructed that allows free-text indexing of sentences surrounding each major term. This is similar to and inspired by the Phrase Finder system.[40]

These calls presented a fairly difficult series of problems because of the wide variety of client voices to

Figure 5 WASABI data broadcasting architecture



be recognized, as well as strong regional accents of many financial consultants. The word error rate was easily 60–70 percent after this speech recognition processing. However, it was found that some post-processing of these transcripts made text mining considerably more fruitful.

For example, the raw data from the speech recognition engine included timing information, and word certainty analysis made it possible to insert punctuation and eliminate words of low certainty. Punctuation was particularly important to the text mining system, since it prevents the system from forming incorrect multiword phrases across sentence boundaries. For the same reason, the low-certainty words were not removed but were replaced with "z" words to prevent incorrect multiword discovery.

Once these documents had been postprocessed as described above, they were processed using the TALENT text mining systems and indexed using a conventional search engine, both for document content and to construct a context thesaurus index. It was then possible to construct a simple client/server query system using JavaServer** pages and a DB2* data repository. A typical client search page from this system is shown in Figure 6.

The results were surprisingly good. The context thesaurus provided a number of extremely good terms to refine and focus the query, and all of the retrieved documents contained the query terms. Since the call documents had no titles, the titles were constructed using the (here fictitious) consultants' names, any person name discovered in the call, and a number representing the call date and extension.

Since these calls did indeed have fairly low word recognition rates, a display of the actual call text would be disconcerting and misleading. However, a display

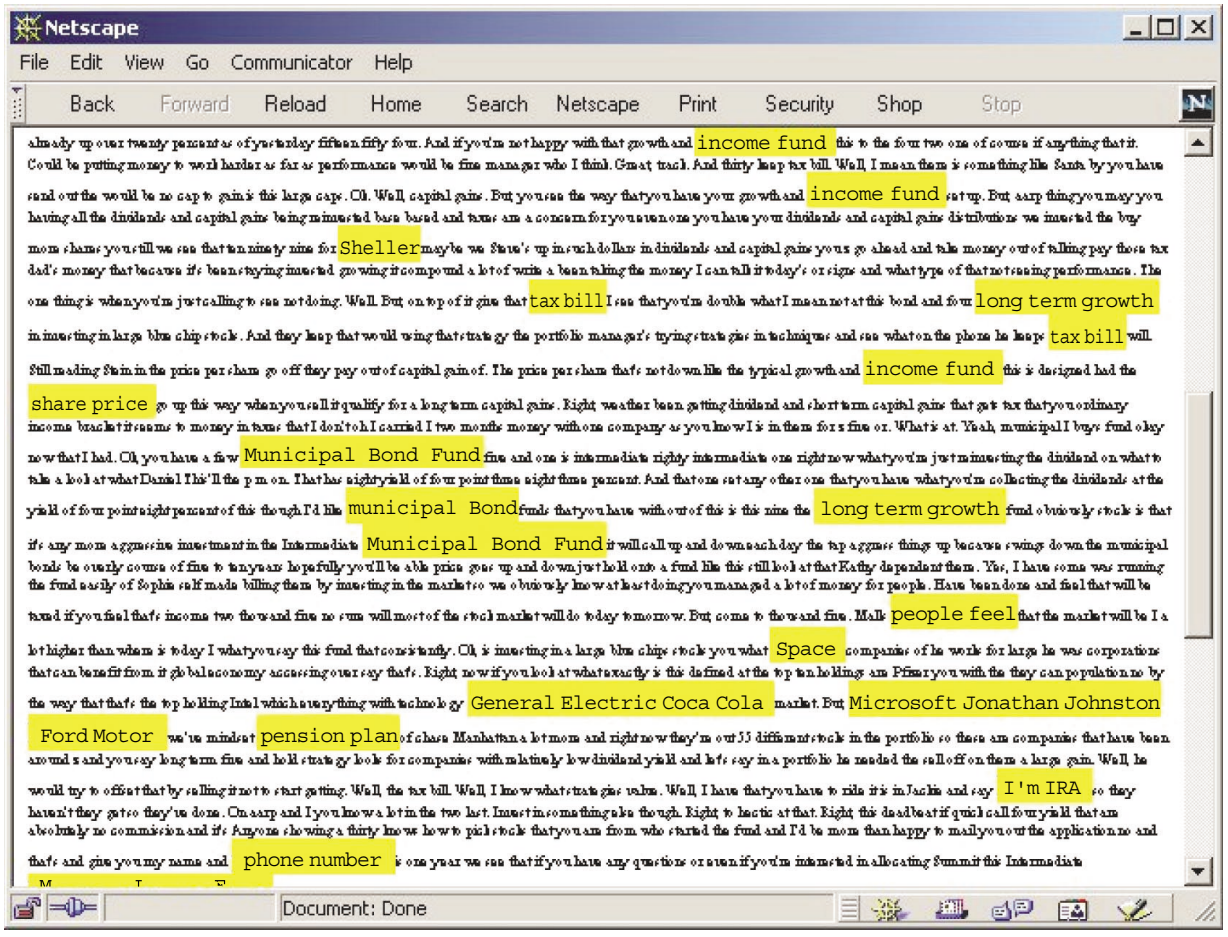
Figure 6 The speech mining user interface, showing the context thesaurus terms (left) and call documents (right)

system was developed using DHTML (Dynamic HyperText Markup Language) that displayed only the salient recognized multiword terms in a readable size font. The remainder were rendered as small as possible, as shown in Figure 7.

In addition, this display system was arranged so that each highlighted term was hyperlinked to a JavaScript** function that calls the browser's audio player to begin playback of the audio file from the time point in the call where the term was displayed. This, then, provided a convenient method of playing back the call around the terms of interest.

Our conclusions so far from these experiments indicate that while it is not possible to make accurate

Figure 7 A call and playback display



transcripts of such telephone calls, it is quite possible to recognize a large number of salient terms and index them for searching. Further, the number of salient multiword terms this mining process discovers compared to the number found manually is quite high (greater than 65 percent) even when the WER is as high as 70 percent. This, then, provides call center supervisors and sales analysts with a method of indexing and searching the vast quantity of call information that is accumulated each month, allowing them to discover sales trends and improve sales training and customer response techniques.

Conclusions

Successful knowledge management applications will ultimately need to solve the problem of capturing and exploiting tacit knowledge. Current solutions that focus on structured data and semistructured documents are limited to the knowledge that users are willing (or forced) to document in these forms. Even when users wish to document their knowledge, they may be unable to articulate the tacit knowledge that they use daily in an almost subconscious manner. This leaves us with only the words and actions of the user as an explicit representation of the tacit knowledge they possess. Fortunately, we can easily capture words and actions with audio and video recordings. The challenge is to process these recordings and extract the knowledge in a form amenable to further analysis and reuse.

The most promising approach to accomplishing this task is the use of automatic speech recognition to

convert recorded audio into text transcripts. In this paper, we have explored the basics of speech recognition as well as a variety of speech recognition applications. The most successful applications to date involve relatively controlled environments with high-quality audio recordings or highly constrained vocabularies and grammars. Examples of such applications include the automated voice response systems, dictation systems, and multimedia indexing systems discussed in the second and third sections.

The successful application of speech recognition to the knowledge management problem, however, requires that the technology work well in the face of a variety of obstacles, including poor-quality audio, random background noise, and arbitrary discourse. The exploratory speech mining systems described in the fourth section are investigating these more challenging environments and looking at ways to enhance speech recognition with other technologies, such as text analysis and natural language processing. The ultimate success of these systems will depend on both better speech recognition technology and clever application of text analysis techniques.

The prospects for fully exploiting tacit knowledge are promising. The systems presented here show great progress in building software solutions that can process and exploit recorded audio and video. When this is combined with trends in pervasive computing, where audio and video capture devices are becoming smaller, cheaper, and more powerful, we have a compelling story for the future. Much work remains in this important problem area, but the preliminary results look encouraging.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of Sun Microsystems, Inc.

Cited references

1. R. Mack, Y. Ravin, and R. J. Byrd, "Knowledge Portals and the Emerging Digital Knowledge Workplace," *IBM Systems Journal* **40**, No. 4, 925–955 (2001, this issue).
2. I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company*, Oxford University Press, New York (1995).
3. L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of IEEE* **77**, No. 2, 257–286 (February 1989).
4. G. Saon, G. Zweig, J. Huang, B. Kingsbury, and L. Mangu, "Evolution of the Performance of Automatic Speech Recognition Algorithms in Transcribing Conversational Telephone Speech," *Proceedings, Instrumentation and Measurement Technology Conference*, Budapest, Hungary (May 21–23, 2001).
5. S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen, "Automatic Transcription of Broadcast News," to appear in *Speech Communication* (2001).
6. A. G. Hauptmann, "Speech Recognition in the Informedia Digital Video Library: Uses and Limitations," *Proceedings, 7th IEEE International Conference on Tools with AI*, Washington, DC (November 5–8, 1995).
7. S. E. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. C. Woodland, "Spoken Document Retrieval for TREC-7 at Cambridge University," *NIST Special Publication 500-242: The Seventh Text Retrieval Conference* (*TREC-7*), Gaithersburg, MD (November 9–11, 1998), pp. 191–200.
8. A. Singhal, J. Coi, D. Hindle, D. Lewis, and F. Pereira, "AT&T at TREC-7," *NIST Special Publication 500-242: The Seventh Text Retrieval Conference TREC-7*, Gaithersburg, MD (November 9–11, 1998).
9. S. Srinivasan, D. Petkovic, D. Ponceleon, and M. Viswanathan, "Query Expansion for Imperfect Speech: Applications in Distributed Learning," *Proceedings, IEEE Workshop on Content-Based Access of Image and Video Libraries*, Hilton Head, SC (June 12, 2000).
10. J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *NIST Special Publication 500-246: The Eighth Text REtrieval Conference* (*TREC 8*), Gaithersburg, MD (November 17–19, 1999), pp. 107–130.
11. E. M. Voorhees and D. Harman, "Overview of the Sixth Text Retrieval Conference (TREC-6)," *Information Processing and Management* **36**, No. 1, 3–35 (January 2000).
12. J. Lai and J. Vergo, "MedSpeak: Report Creation with Continuous Speech Recognition," *Proceedings, ACM Conference on Human Factors in Computer Systems*, Atlanta, GA (March 22–27, 1997), pp. 431–438.
13. G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Retrieving Spoken Documents by Combining Multiple Index Sources," *Proceedings, ACM SIGIR International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland (August 18–22, 1996), pp. 30–38.
14. K. Ng and V. Zue, "Phonetic Recognition for Spoken Document Retrieval," *Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA (May 12–15, 1998), pp. 325–328.
15. M. Wechsler, E. Munteanu, and P. Schäuble, "New Techniques for Open Vocabulary Spoken Document Retrieval," *Proceedings, ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia (August 24–28, 1998), pp. 20–27.
16. M. Witbrock and A. Hauptmann, "Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents," *Proceedings, ACM International Conference on Digital Libraries*, Philadelphia, PA (September 27–29, 1997), pp. 30–36.
17. S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox, "Multimedia Search: An Authoring Perspective," *Proceedings, First International Workshop on Image Databases and Multimedia Search*, Amsterdam (August 22–23, 1996).
18. P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," *Multimedia Tools and Applications* **3**, No. 3, 179–202 (1996).
19. J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "Virage Image Search Engine: An Open Framework for Image Management," *Proceedings, Storage and Retrieval for Still Images and Video Databases* (*SPIE*), (February 1996). Available at http://www.virage.com.

20. S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," *Proceedings, ACM Multimedia*, Seattle, WA (November 9–13, 1997), pp. 313–324.
21. H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann, "Lessons Learned from Building a Terabyte Digital Video Library," *IEEE Computer* **32**, No. 2 (February 1999).
22. H. Zhang, P. Aigrain, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," *Multimedia Tools and Applications* **3**, 179–202 (1996).
23. See http://speechbot.research.compaq.com/.
24. M. G. Christel, M. A. Smith, C. R. Taylor, and D. B. Winkler, "Evolving Video Skims into Useful Multimedia Abstractions," *Proceedings, ACM Conference on Human Factors in Computer Systems*, Los Angeles, CA (April 18–23, 1998), pp. 171–178.
25. G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Video Mail Retrieval: The Effect of Word Spotting Accuracy on Precision," *Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, Detroit, MI (May 8–12, 1995), pp. 309–312.
26. A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, "Using Audio Time Scale Modification for Video Browsing," *Proceedings, Hawaii International Conference on Multimedia* (*HICSS-33*), Wailua, Maui, HI (January 4–7, 2000).
27. S. Srinivasan and D. Petkovic, "Phonetic Confusion Matrix Based Spoken Document Retrieval," *Proceedings, ACM SIGIR International Conference on Research and Development in Information Retrieval*, Athens, Greece (July 2000).
28. A. Coden and E. Brown, "Speech Transcript Analysis for Automatic Search," *Proceedings, Hawaii International Conference on System Science*, Maui, HI (January 3–6, 2001).
29. Y. Ravin, N. Wacholder, and M. Choi, "Disambiguation of Names in Text," *Proceedings, ACL Conference on Applied Natural Language Processing*, Washington, DC (March 31–April 3, 1997), pp. 202–208.
30. J. Prager, E. Brown, A. Coden, and D. Radev, "Question-Answering by Predictive Annotation," *Proceedings, ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece (July 24–28, 2000), pp. 184–191.
31. A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting Browser: Tracking and Summarizing Meetings," *Proceedings, DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA (February 8–11, 1998).
32. F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated Technologies for Indexing Spoken Language," *Communications of the ACM* **43**, No. 2, 48–56 (February 2000).
33. T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle, and P. Zellweger, " 'I'll Get That off the Audio': A Case Study of Salvaging Multimedia Meeting Records," *Proceedings, ACM International Conference on Human Factors in Computer Systems*, Atlanta, GA (March 22–27, 1997), pp. 202–209.
34. E. R. Pederson, K. McCall, T. P. Moran, and F. G. Halasz, "Tivoli: An Electronic Whiteboard for Informal Workgroup Meetings," *Proceedings, ACM International Conference on Human Factors in Computing Systems*, Amsterdam, Netherlands (April 24–29, 1993), pp. 391–398. Reprinted in *Readings in Computer-Human Interaction: Toward the Year 2000*, R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, Editors, Morgan Kaufmann Publishers, San Francisco, CA (1995), pp. 509–516.
35. H. Rittel and W. Kunz, *Issues as Elements of Information Systems*, Report S-78-2, Institut für Grundlagen der Planung I.A., University of Stuttgart, Germany (1970).
36. J. Conklin and M. Begeman, "gIBIS: A Hypertext Tool for Exploratory Policy Discussion," *Proceedings, Conference on Computer-Supported Cooperative Work*, Portland, OR (September 26–28, 1988), pp. 140–152.
37. J. W. Cooper, M. Viswanathan, and Z. Kazi, "Samsa: A Speech Analysis, Mining and Summary Application for Outbound Telephone Calls," *Proceedings, Hawaii International Conference on System Science*, Maui, HI (January 3–6, 2001).
38. J. W. Cooper and R. J. Byrd, "Lexical Navigation: Visually Prompted Query Expansion and Refinement," *Proceedings, ACM International Conference on Digital Libraries*, Philadelphia, PA (July 23–26, 1997), pp. 237–246.
39. J. S. Justeson and S. Katz, "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text," *Natural Language Engineering* **1**, 9–27 (1995).
40. J. Xu and W. B. Croft, "Query Expansion Using Local and Global Document Analysis," *Proceedings, ACM SIGIR International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland (August 18–22, 1996), pp. 4–11.

Accepted for publication June 8, 2001.

Eric W. Brown *IBM Research Division, Thomas J. Watson Research Center, Box 704, Yorktown Heights, New York 10598 (electronic mail: ewb@us.ibm.com).* Dr. Brown is a research staff member at the IBM Thomas J. Watson Research Center. He earned his Ph.D. degree in computer science in 1996 from the University of Massachusetts, Amherst. His research interests include performance issues in information retrieval systems, parallel and distributed text search, text categorization, question answering, and applications of automatic speech recognition to knowledge management problems. Dr. Brown has published a number of database and information retrieval papers and patents and is a member of the ACM.

Savitha Srinivasan *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electronic mail: savitha@almaden.ibm.com).* Ms. Srinivasan is the manager of the Multimedia Knowledge Discovery group at IBM Almaden Research Center. She received an M.S. degree in computer science and joined the speech group at the Thomas J. Watson Research Center in 1990 as a speech applications researcher. She worked on the design and development of IBM's speech recognition products, such as the IBM Speech Server Series and Medspeak Radiology. She moved to IBM Almaden Research in 1997 and has since then been working on applications of speech recognition to multimedia indexing. She is the author of several publications and patents in speech recognition-related applications and multimedia information retrieval.

Anni Coden *IBM Research Division, Thomas J. Watson Research Center, Box 704, Yorktown Heights, New York 10598 (electronic mail: anni@us.ibm.com).* Dr. Coden is a research staff member at the Thomas J. Watson Research Center. Her most recent work is focused on applying speech recognition and text analysis to automatically analyze broadcast video and find relevant, collateral information. Her other recent activities include question answering and video indexing and search. Dr. Coden holds multiple patents in the areas of multimedia search technologies and speech

technology. She received her Ph.D. degree in computer science from the Massachusetts Institute of Technology in 1981 and, after spending a year at MIT as a research scientist, joined IBM.

Dulce Ponceleon *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electronic mail: dulce@almaden.ibm.com).* Dr. Ponceleon received her M.S. and Ph.D. degrees in computer science from Stanford University. In 1991 she joined the Advanced Technology Group at Apple Computer, where she worked on information retrieval, video and audio compression and technologies for QuickTime, and computer graphics. She was a key contributor to the first software-only videoconferencing system. In 1997 she joined IBM Almaden Research, where she has worked on video content analysis, video summarization, visualizations for video, multimedia indexing, and application of text analysis techniques and related technologies to automatic speech recognition transcripts. She has participated in the ISO MPEG-7 standardization efforts. She has numerous patents and publications in video and audio compression, information retrieval, multimedia indexing and retrieval, numerical linear algebra and nonlinear programming.

James W. Cooper *IBM Research Division, Thomas J. Watson Research Center, Box 704, Yorktown Heights, New York 10598 (electronic mail: jwcnmr@us.ibm.com).* Dr. Cooper is a research staff member at the IBM Thomas J. Watson Research Center. His interests include Java technology, object-oriented programming, design patterns, information retrieval and text mining technologies, and user interface design. Dr. Cooper is a columnist of *JavaPro* magazine and the author of 14 books; he holds several patents.

Arnon Amir *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (electronic mail: arnon@almaden.ibm.com).* Dr. Amir received his B.Sc. degree *magna cum laude* in electrical and computer engineering from the Ben Gurion University, Israel, in 1989, and the M.Sc. and D.Sc. degrees in computer science from the Technion, Israel Institute of Technology, in 1992 and 1997, respectively. In 1997 he joined the IBM Almaden Research Center, where he is currently a research staff member in the USER group. His research interests include computer vision, video indexing and retrieval, eye gaze tracking, graph clustering, and spatial data structures. He is the author of more than 30 technical papers in journals and refereed conferences and has filed more than 10 patents. He was awarded the 1997–1998 Rothschild fellowship and is a member of the IEEE and of the ACM.