Extracting Knowledge from Speech

Eric Brown[†], *Savitha Srinivasan*[‡], *Anni Coden*[†], *Dulce Ponceleon*[‡], *James Cooper*[†], *Arnon Amir*[‡], *Jan Pieper*[‡]

[†]IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{ewb, anni, jwcnmr}@us.ibm.com

[‡]IBM Almaden Research Center
San Jose, CA 95120
{savitha, dulce, arnon, jhpieper}@almaden.ibm.com

Abstract

One of the more difficult challenges in knowledge management is capturing and exploiting tacit knowledge. Unlike explicit knowledge captured in structured or semi-structured information, such as relational data and text documents, tacit knowledge is rarely available in a form that lends itself to indexing and mining for knowledge reuse. More often, tacit knowledge is the experience and expertise built up by an individual over time, and it materializes in the words and actions of that individual. To exploit this knowledge, a knowledge management system must be able to capture and analyze discussions, presentations, and a variety of daily actions. These activities appear as audio and video, which the knowledge management system must record, process, and convert into useful sources of knowledge. This paper describes some of the problems and solutions associated with capturing and exploiting audio and video data as a knowledge source, with particular emphasis on speech. We present a family of related applications and systems being developed at IBM Research that analyze audio and video in a variety of knowledge environments (e.g., video archives, seminars, meetings, personal dictation, etc.) and explore future directions in this challenging area of knowledge management.

1 Introduction

The knowledge management problem is quite complex, but at its core the problem is largely about capturing the expertise, experience, and collective "know how" of an organization and making that information readily available and easily accessible. Technological solutions to this problem typically focus on representations of knowledge that are most easily manipulated by a computer. These representations commonly include structured data, such as relational databases, and various kinds of unstructured data, such as text documents.

While there is certainly a great deal of information available in these forms, restricting knowledge management solutions to just these forms is perilous for at least two reasons. First, there is considerable effort involved in creating this kind of data. Documents, in particular, can be tedious and time consuming to create, making the final product incomplete and out of date. Second, the information captured in these data forms is often dictated by formal procedures, such that any knowledge that falls outside the scope of these procedures will not be captured. In particular, tacit knowledge (the knowledge individuals use almost subconsciously to conduct their daily activities) is rarely made explicit in documents and databases. Fortunately, the knowledge that escapes these more conventional data sources often materializes in other forms, namely the words and actions of the individuals who posses the knowledge. The challenge is capturing and exploiting the knowledge expressed in these other, unconventional knowledge sources.

One promising approach to meeting this challenge is the use of automatic speech recognition to convert recorded speech into a text transcript, followed by the application of text analysis tools on the transcript. This allows us to apply a variety of text mining techniques on speech data and creates a

number of opportunities for capturing and exploiting tacit knowledge. We are exploring these opportunities in a variety of contexts, ranging from automated audio and video indexing for management of multimedia digital libraries, to real-time analysis of speech to provide on-line support for meetings, call centers, and applications that combine spoken discourse with information management needs.

This paper describes the problems, technologies, solutions, and systems that address the information capture, retrieval, and analysis issues from unconventional knowledge sources such as audio and video. The focus here is on systems that can automatically extract information from such knowledge sources, rather than approaches that depend on manual annotations or decisions. We begin with a primer on automatic speech recognition technology in Section 2. Section 3 describes the key research projects at IBM that apply automatic speech recognition to the problem of indexing, searching, and managing audio and video. Section 4 presents some more exploratory projects at IBM investigating an area loosely called *speech mining*. Finally, in Section 5 we offer some concluding remarks.

2 Speech Recognition Concepts

Speech recognition applications traditionally fall into one of three categories: dictation or document creation systems, transactional or data-entry systems (e.g., automated voice response systems), and audio indexing systems. Each type of application addresses different requirements, and comprises different design criteria to overcome the challenges imposed by the constraints of the technology. Advances in technology are making significant progress towards 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 one's 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) [Rabiner89], 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 where as a large vocabulary system may consist of tens of thousands of words. Large vocabulary speech recognition systems typically use a sub-word approach where phonetic sub-word 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 language model assists the speech engine in recognizing speech by biasing the output towards 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.

Speech recognition systems output the most probable decoding of the acoustic signal as the recognition output, but keep multiple hypothesis that are considered during the recognition. The multiple hypotheses at each 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 output 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 the application.

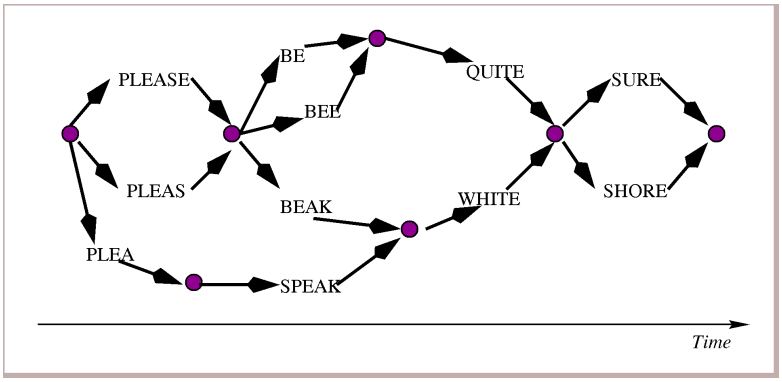

Figure 1 Hypothetical recognition of phrase "Please be quite sure" in a Word Lattice Representation

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. It has been shown that WER can vary between 8-15% on standard speech evaluation data and 70-85% on "real-world" data such as a one hour video documentary with commercials [Dharanipragda98, Hauptmann95, Johnson98]. Singhal has shown that word deletion and word substitution errors occur more frequently than word insertion errors [Singhal98].

The performance of current speech recognition systems is adequate for the traditional applications described above. To address the issue of extracting knowledge from speech, however, we must apply speech recognition in more challenging environments. In the next two sections we describe several on-going research projects at IBM that are exploring more ambitious applications of speech recognition technology. These projects range from being relatively robust to efforts that are more exploratory. In general, the indexing applications are more mature and reliable enough to try out in the field, while the mining applications are more exploratory.

3 Indexing Applications at IBM Research

Advanced indexing algorithms for multimedia retrieval are being pursued at the IBM Almaden Research Center for a broad class of indexing applications. These include indexing multimedia archives such as video assets owned by corporations, and more recently, on live streaming sources of media on the web. This enabling technology opens up a whole new area of applications in the realm of business intelligence that can tap into live media sources on the web as a source of information.

3.1 CueVideo

CueVideo is a research project at the IBM Almaden Research Center that consists of an automatic multimedia indexing system and a client-server video retrieval and browsing system. The CueVideo approach to multimedia retrieval is *"Search the speech, browse the video"*. The video and audio are two parallel media streams of information related by a common time line. Therefore, the system takes advantage of both; it uses the audio stream for search and the video stream for quick visual browsing in a complimentary manner to provide the desired video search functionality. The video indexing system automatically detects shot boundaries, generates a shots table, and extracts representative key-frames as JPEG files from each of the shots. These results are used to generate the following browsable summaries of the input video [Amir00] using the architecture shown in Figure 2:

- Storyboard: a set of one or more pages, each consists of a two dimensional array of key-frames, sorted in chronological order
- Animation: a quick slide show, where each of the key-frames is shown for a fixed short period (e.g., 0.6 seconds)
- Moving Storyboard (MSB): the animated key-frames, fully synchronized with the original audio track. Each key-frame is shown for the entire duration of the associated shot.
- Fast Moving Storyboard (Fast MSB): also referred to as Fast Time Scale Modification (Fast TSM) is an MSB with accelerated audio (typically 1.5 times faster than original). Similarly, a Slow MSB is also generated.

The audio processing starts with speech recognition (using the IBM Speech Recognition system with Broadcast News models [Dharanipragada98]) 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 TSM audio in desired speedup rates for the fast and slow MSBs. A phonetic transcription of the input audio is generated and overlapping triphone and quadphone sequences are selected as subword index terms [Srinivasan00]. The choice of three or four as the length of the phone sequences is motivated by the average length of syllables in the English language. These sequences are derived by successively concatenating the appropriate number of phones from the phonetic transcription. This phoneme sequence representation is augmented with additional phone sequences derived from the observed phones in the transcription and the *phone confusion matrix*. For a given phone set, the phone confusion matrix is defined as the set of phones that may be mistakenly recognized by a recognition system for each phone and its associated confusion probability. A Bayesian probabilistic model is used as the decision function for retrieval to estimate the probability that the phonetic representation of a speech segment satisfies a query based on term weighting.

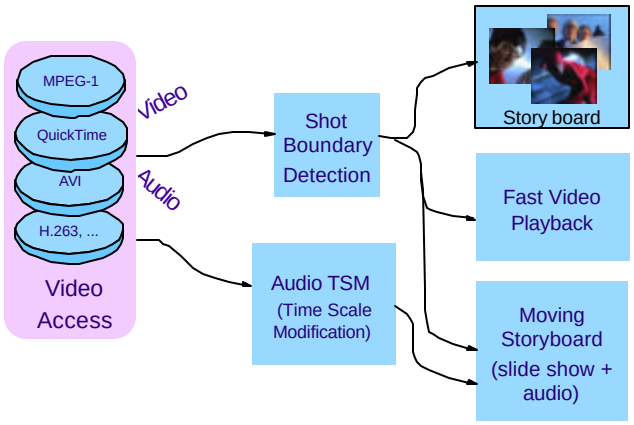

Figure 2 CueVideo Indexing Architecture

The following 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 out-of-vocabulary query terms, phonetic retrieval results in an average recall of 0.88 and an average precision of 0.69 with 35% WER. For higher WERs, retrieval performance is lower. For in-vocabulary query terms, phonetic retrieval results in an average recall improvement of 17% with an average precision loss of 17% even for high WERs over word-based retrieval.

The three-tier architecture of the client-server retrieval system consists of the following components:

- Client: standard Web browser with a streaming media plug-in (e.g., the Real Player G2).
- Web Server: serves the HTML pages, storyboards, and connects to the application interface (via cgi).
- Media streaming server: streams the video and its views and summaries (e.g., the RealMedia streaming server).

- Speech Retrieval Server: the CueVideo search engine, which performs the speech search and returns ranked results.

3.2 CueRadio Live

We are extending the CueVideo architecture to tackle a new problem on the World Wide Web, namely how to sort through the growing number of streaming media sources. A recent study of the web estimates the number of live broadcast channels on the web to be on the order of 5000 [Ashour01]. The increasing number of channels creates a requirement for indexing techniques that can operate on live broadcast channels for searching and filtering applications. The CueRadio Live system is a research prototype that indexes multiple live streaming media channels and provides both a query interface and an agent interface to the content being broadcast. Figure 3 below shows the system architecture. Users connect to the live media indexer using standard browser interfaces. The backend of the system consists of live media indexing workstations. A user query results in the servlet issuing a request to the live media retrieval engine, which in turn is constantly updated by he indexing machines. At this point we have a simplistic keyword search attached to the retrieval engine; adding sophisticated speech retrieval algorithms is a logical next step.

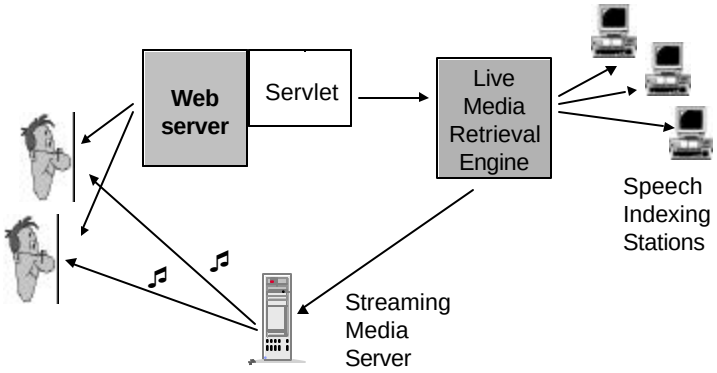

Figure 3 CueRadio architecture.

4 Speech Mining Applications at IBM Research

While indexing applications are 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, any time, and it is an instantiation of the tacit knowledge possessed by the discourse participants. A recent research focus at the IBM T.J. Watson Research Center has been the problem of 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.

Towards that end, the researchers at Watson have built a generic framework for analyzing speech, called WASABI (Watson Automatic Stream Analysis for Broadcast Information) [Coden01]. 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 is relevant to the current discourse. The overall architecture is shown in Figure 4.

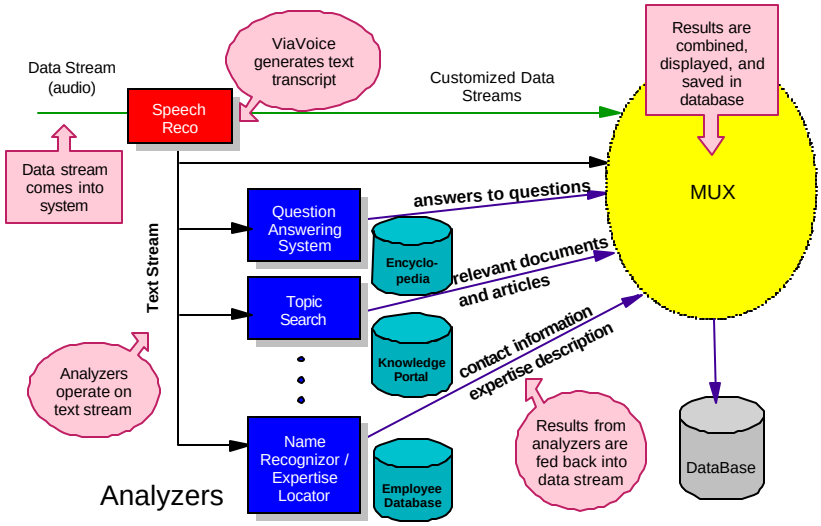

Figure 4 WASABI architecture.

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.

4.1 MeetingMiner

The first application area addressed by the WASABI framework is meetings, which 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. More formal meetings may have actual minutes created by a designated scribe, while less formal meetings 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. The system 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, which include a named entity recognizer, a topic tracker, and a question identifier tied to a question answering system [Prager00].

The MeetingMiner system is still in the early stages of development and testing. A significant hurdle facing the system is the ability of the speech recognition engine to perform well 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 non-intrusive to the meeting participants as possible. Speaker independent speech recognition should improve as more effort is made to support speech recognition in these challenging multi-speaker environments. In the mean time, 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. [Waibel98] from 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 post-processing steps on the speech recognition transcripts to generate summarizes. 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 [Kubala00], 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.

4.2 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 [Coden01]. A high definition television (HDTV) channel has over 1.5 Mbits/sec 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 one appealing 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) 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. The audio/video source is fed to a collection of real time feature extractors that extract text and possibly other visual features. The system sends these features to the event analyzers, which classify the features into topics, extract named entities (e.g., names of people, places, dates, etc.), and combine these events into a data structure called the *Knowledge Chain*, which assembles all of the events on a timeline.

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 the program to be paused while the user explores the collateral information in more detail.

4.3 SAMSA

The SAMSA (Speech Analysis Mining and Summary Application) project [Cooper01] is an experiment in the actual mining of outbound telephone sales calls using text mining techniques. It does not use the WASABI framework directly, but shares some common components. We recorded telephone calls from financial consultants on digital recording tape over a period of several weeks, and then processed them using the IBM Research version of the speech recognition engine optimized for telephone calls and

speaker independent recognition [Dharanipragada97]. From about 11,000 call units, including empty calls, we selected 529 calls of substantive length for speech recognition processing.

The TALENT text mining processes have previously been described [Cooper97]. Briefly, the text mining system recognizes multi-word names [Ravin97], locations, organizations [Justeson95], and detects relations between terms based on proximity and by common English language patterns such as appositives and parentheticals. In addition, TALENT can construct a Context Thesaurus, which allows free text indexing of sentences surrounding each major term. This is similar to and inspired by the Phrase Finder system [Xu96].

These calls presented a fairly difficult series of problems because of the wide variety of client voices to be recognized as well as strong regional accents among many of the financial consultants. The word error rate was easily 60-70% after this speech recognition processing. However, we 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 scores, which 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 multi-word phrases across sentence boundaries. For the same reason, the low-certainty words were not removed but replaced with "z" words to prevent incorrect multiword discovery.

Once these documents had been post processed 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.

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 did contain the query terms. Since the call documents had no titles, the titles were constructed using the consultant's 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 system was developed using DHTML that displayed only the salient multi-word terms recognized in a readable size font. The remainder were rendered as small as possible, as shown in Figure 5.

In addition, this display system was arranged so that each highlighted term was hyperlinked to a JavaScript call that would call the browser's audio player to begin playback of the call 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.

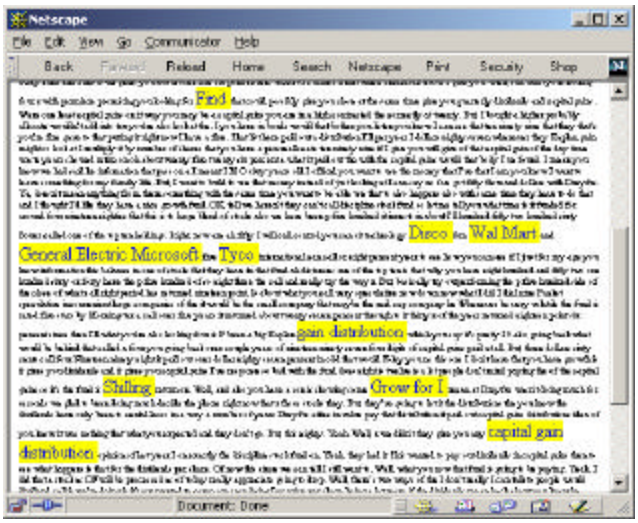

Figure 5 A call and playback display

Our conclusions so far from these experiments have been that while it is not possible to make accurate transcripts of such telephone calls, it is quite possible to recognize a large number of salient terms and index them for searching. This 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.

5 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 semi-structured 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 posses. Fortunately, we can easily capture words and actions with audio and video recordings. The challenge is processing these recordings and extracting the knowledge in a form amenable to further analysis and reuse.

A promising approach to accomplishing this task is the use of automatic speech recognition to convert recorded audio into text transcripts. Speech recognition has a long research history and has been applied in three main application areas, including transactional and data entry systems, dictation systems for creating memos and documents, and indexing systems for searching audio and video. The first two application areas involve direct interaction by the user with the system and have been relatively successful. The third application area is a passive process that can be applied to any recorded speech. It too has been successful in indexing applications and comes closest to meeting the requirements of a knowledge management application. These requirements, however, go well beyond indexing and impose

significant demands on the quality of the transcript, the speed of the recognition engine, and the complexity of the analyses applied to the resulting transcript.

In this paper we have explored these issues in depth and described a number of promising technologies and applications at IBM Research aimed at solving this aspect of the knowledge management problem. Much work remains in this important problem area, but the preliminary results look encouraging.

References

[Amir00] Amir, A., Ponceleon, D., Blanchard, B., Petkovic, D., Srinivasan, S. and Cohen, G. Using Audio Time Scale Modification for Video Browsing. In Proceedings of *HICSS-33*, Hawaii, Jan. 2000.

[Ashour01] Ashour, A., Dom, B., Golden, J., Srinivasan, S. and Bulterman, D. Who s SMILing on the Web? In Poster Proceedings of WWW-1, Hog Kong, May 2001.

[Coden01] Coden, A. and Brown, E. Speech Transcript Analysis for Automatic Search. In *Proceedings of HICSS-34*, Maui, Hawaii, 2001.

[Cooper97] Cooper, J. W. and Byrd, R. J. "Lexical Navigation: Visually Prompted Query Expansion and Refinement." *Proceedings of DIGLIB97*, Philadelphia, PA, July, 1997.

[Cooper01] Cooper, J.W., Viswanathan, M., and Kazi, Z. "Samsa: A Speech Analysis, Mining and Summary Application for Outbound Telephone Calls," In *Proceedings of HICSS-34*, Maui, Hawaii, 2001.

[Dharanipragada98] Dharanipragada, S., Franz, M. and Roukos, S. Audio-Indexing For Broadcast News. In *Proceedings of Seventh Text Retrieval Conference, TREC-7, (NIST Special Publication)* 1998.

[Dharanipragada97] Dharanipragada, S. and S. Roukos, Experimental Results in Audio Indexing in *Proceedings of the DARPA Speech Recognition Workshop*, 1997.

[Hauptmann95] Hauptmann, A.G. Speech Recognition in the Informedia Digital Video Library: Uses and Limitations. In *Proceedings of ICTAI-95 7th IEEE International Conference on Tools with AI*, Washington, DC.

[Johnson98] Johnson, S.E., Jourlin, P., Moore, G.L., Jones, K.S. and Woodland, P.C. Spoken Document Retrieval for TREC-7 at Cambridge University. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7), (NIST Special Publication)* 1998.

[Justeson95] Justeson, John S. and Slava Katz. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text, in Natural Language Engineering, 1, pp 9-27, 1995.

[Kubala00] Kubala, F., Colbath, S., Liu, D., Srivastava, A., and Makhoul, J., "Integrated Technologies for Indexing Spoken Language," *Comm. of the ACM*, 43(2), pp. 48-56, Feb., 2000.

[Prager00] Prager, J., Brown, E., Coden, A., and Radev, D., "Question-Answering by Predictive Annotation," *Proc. of the 23rd Inter. ACM SIGIR Conf. On Res. And Develop. In Information Retrieval*, pp. 184-191, Athens, Greece, 2000.

[Rabiner89] Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE 77(2):257-286.

[Ravin97] Ravin, Y., Wacholder, N., and Choi, M., "Disambiguation of Names in Text," *Proc. of the Fifth ACL Conf. on Applied Natural Language Processing*, pp. 202-208, Washington D.C., 1997.

[Singhal98] Singhal, A., Coi, J., Hindle, D., Lewis, D. and Pereira, F. AT&T at TREC-7. In *Proceedings of the Seventh Text Retrieval Conference TREC-7, (NIST Special Publication)* 1998.

[Srinivasan00] Srinivasan, S. and Petkovic, D., Phonetic Confusion Matrix Based Spoken Document Retrieval. In Proceedings of SIGIR-2000, Greece, July 2000.

[Waibel98] Waibel, A., Bett, M., Finke, M., and Stiefelhagen, R., "Meeting Browser: Tracking and Summarizing Meetings," *Proc. of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[Xu96] Xu, J. and Croft, W. B., "Query Expansion Using Local and Global Document Analysis," *Proceedings of the 19th Annual ACM-SIGIR Conference*, 1996.