METADATA FOR MIXED-MEDIA ACCESS

Francine Chen, Marti Hearst, Julian Kupiec, Jan Pedersen, Lynn Wilcox
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Abstract

In this paper, we discuss *mixed-media access*, an information access paradigm for multimedia data in which the media type of a query may differ from that of the data. The types of media considered in this paper are speech, images of text, and full-length text. Some examples of metadata for mixed-media access are locations of keywords in speech and images, identification of speakers, locations of emphasized regions in speech, and locations of topic boundaries in text. Algorithms for automatically generating this metadata are described, including word spotting, speaker segmentation, emphatic speech detection, and subtopic boundary location. We illustrate queries composed of diverse media types in an example of access to recorded meetings, via speaker and keyword location.

1 Introduction

Modern document databases contain information in a variety of media; audio, video, and image are becoming increasingly common companions to plain text. Access to this rich and variegated information is typically accomplished in standard systems through queries over manually supplied topic keywords or descriptive text. Recent work suggests that fully automatic methods operating directly on the data can offer comparable characterizations. Moreover, analyses tuned to particular media can expose structure that form the basis for more natural access modes.

We are exploring these issues through a paradigm we call *mixed-media access*, which encourages the user to query in the medium of greatest convenience regardless of the media type of the data. This encompasses, for example, spoken access to a textual databases, as well as queries that combine cues across the media types present in a complex document.

Special metadata considerations arise within such a paradigm. For our purposes, mixed-media metadata is defined as derived properties of the media which are useful for information access or retrieval. These properties can be derived either in advance or "on the fly". Our focus in this paper is on automatically derived metadata for speech, scanned text images, and full-length text.

In a purely textual database, metadata for information access typically consists of indices on word tokens. The state-of-the-art in speech and image recognition is such that we cannot reliably create a word-level transcription for an arbitrary speech document [18] or text image [5]. Therefore, in a multimedia database prepared for mixed-media access it is unrealistic to suppose that a full transcription is available in advance as metadata. We can, however, robustly recognize particular keywords, a process known as *word spotting* [26]. Word spotting produces metadata in the form of time indices of keywords in audio [27], or locations of keywords in a text image [3].

In addition, we can enrich this word-level metadata with information that captures some of the context implicit in particular media. For example, in speech data one important aspect is the identity of the speakers. We can automatically detect speaker changes in audio, in a process which we refer to as *speaker segmentation* [28]. This produces metadata in the form of time indices of the audio segments corresponding to the different speakers. When the speakers are known each segment can be annotated with the identity of the speaker, a process known as *speaker identification*. This information helps characterize the data, can be stored as metadata and used for indexing.

Another source of information in speech not present in plain text is prosodic information. Prosodic information includes changes in pitch, amplitude, and timing, and is used by a speaker to signal regions of speech that are important. We can automatically detect regions of emphatic speech [4] and note the time indices of the audio segments in which emphatic speech occurs. These regions are another form of metadata and can be used as a method of indexing into a conversation.

Another source of information for metadata which can be applied to both spoken and written text is that of subtopic change [10]. We can determine when the discussion within a text changes from one subtopic to the next. Subtopic boundary locations can then be used to generate indices that indicate which paragraphs or which regions correspond to each subtopic segment. The next step is identification of the contents of the subtopics within the subtopic boundaries. Note that determining subtopic boundaries versus subtopic content is analogous to determining speaker change versus speaker identification.

In the remainder of this paper, we first provide more detail about the three media types and their corresponding metadata. We then describe how this metadata can be derived automatically, and finally present examples of the use of such metadata in mixed-media access.

2 Characteristics of the Media and the Derived Metadata

Digitized speech and scanned images of text are not easily searched for content. However, they contain information which can be organized to provide easier access. Metadata providing indices to speech includes keyword locations, segmentation of a conversation by speaker, and regions which a speaker highlighted by speaking more emphatically. In text images, keywords and layout may be identified.

Queries may consist of a boolean expression, which requires searching for a small number of keywords or phrases, or for a particular speaker. Research indicates that there is a tradeoff between preset keywords or topics and unlimited vocabulary in information access [17]. If the set of keywords is fixed, metadata based on keyword locations can be pre-computed and stored for later indexing. This is efficient, in that keyword spotting can be done off-line, but restrictive, in that the available query terms are limited to a predefined set of keywords. In general, unrestricted vocabulary searching produces better results than restricted vocabulary searching. It is not currently possible to generate precomputed metadata that supports unrestricted vocabulary searching over image or audio data. However, it is possible to support this search style by spotting for keywords "on the fly".

2.1 Speech

Although speech data can be recorded in analog form, it is commonly sampled at equally spaced time intervals, that is, digitized. For telephone quality speech, the sampling rate is typically 8 kHz, or 8,000 samples per second. For better quality audio, higher sampling rates are needed to reproduce the higher frequencies. For example, the sampling rate used for Compact Disc (CD) audio is 44.1 kHz. The quality of the recorded speech is also influenced by the number of bits used to code a sample. For telephone quality speech, each sample is represented by a single byte, or 8 bits. For a higher dynamic range, more bits per sample are required. For CD quality audio, 16 bits per sample is standard.

Portions of the audio data can be reproduced from the digitized form by specifying starting and ending times for the desired segment. These starting and ending times are converted into sample number based on the sampling rate. However, indexing solely by time interval is restrictive; the development of sophisticated speech analysis techniques allows for attribute-based interval specification, such as locating the portion of an audio stream in which a comment was made by a particular speaker.

Speech can be analyzed in different ways. One is in terms of the presence of specific keywords in the speech [26]. Another is in terms of the identity of the speakers in the audio [28]. A third is in terms of prosodic information [4], which can be used by a speaker to draw attention to a phrase or sentence, or to alter the word meaning. This information can be exploited to obtain metadata for access to audio.

Metadata describing keywords consists of the keyword and its starting and ending time in the audio. Because the process of spoken keyword identification is not perfectly accurate, a confidence score is also a part of the metadata representation.

Metadata describing the identity of the speakers consists of the name of the speaker, and a list of the time intervals during which the speaker talks. Speaker-independent indices are also maintained by taking note of when one speaker finishes talking and silence or another speaker follows. In this case, starting and ending times for each of the different speakers in the audio are recorded, but only symbolic names are attached to the speakers, for example speaker A, speaker B, etc. This allows for retrieval of logical speaker units even when the identity of the speakers are not known. Additionally, silence and non-speech sounds can be identified and stored as metadata.

Metadata describing emphasized regions of speech consists of the time indices of the intervals where the speech was emphatic, as well as a measure of certainty that the particular region did indeed contain speech which the speaker intended to emphasize.

2.2 Text Images

Images of text may be analog, as in microfiche, or digital, as is commonly used by facsimile machines. In addition, documents may be stored as digital text images to preserve the document's original formatting and figure information. Digitized text images can be characterized by the sampling rate in terms of the number of dots, or pixels, per inch and number of bits per pixel. Images are typically scanned at sampling rates ranging from 200 to 500 dpi. The higher sampling rates may be used when documents contain smaller fonts or thin strokes [22]. Currently, most images of text documents are represented as binary images; that is, 1 bit is used to indicate whether a pixel is on or off. However, photographs and figures are better represented using 8 bits per pixel.

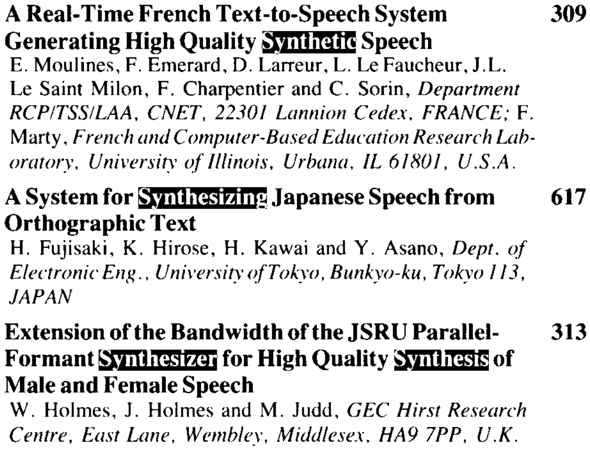
Text image data can be characterized in terms of the layout structure of the document, e.g., columns and paragraphs [12], semantic information contained in the document [8], and by the words in the document. However, a reliable word-level transcription of arbitrary pages containing text is not yet possible. Therefore, rather than use a word-level transcription, we characterize image data by the the location and identity of keywords, which can be stored as metadata. As in the audio example, the representation may also include a score indicating the degree of confidence in the identification of the keyword.

2.3 Full-length Text

Full-length texts are natural language expressions of sufficient length to exhibit topical substructure. For example, a magazine article will be composed of numerous sections each illuminating aspects of the overall topic. Often these sections will be demarked by author provided typographical annotations, perhaps in a markup language such as SGML [23]. However, author provided subtopic markup is neither always available nor always reliable.

Full-length text shares the same basic representation as shorter text forms, such as titles and abstracts: words. Therefore standard mechanisms for text indexing, such as inverted indices [21], can act as metadata. In addition, current work in computational linguistics allows for the assignment of additional information at the word token level, e.g., part-of-speech tags [6] and morphological derivation [7].

Full-length texts can be segmented at topic and subtopic boundaries. Algorithms that detect subtopic structure can partition the text or allow overlap among multi-paragraph units. In both cases, the metadata consists of indices indicating which paragraphs or which regions of tokens correspond to each subtopic segment. Additionally, information that characterizes the content of the subtopics and the main topics can serve as useful metadata [11]. Automated determination of main topic and subtopic content information is an active area of research.


A Real-Time French Text-to-Speech System Generating High Quality Synthetic Speech 309
E. Moulines, F. Emerard, D. Larreur, L. Le Faucheur, J.L. Le Saint Milon, F. Charpentier and C. Sorin, *Department RCP/TSS/LAA, CNET, 22301 Lannion Cedex, FRANCE;* F. Marty, *French and Computer-Based Education Research Laboratory, University of Illinois, Urbana, IL 61801, U.S.A.*

A System for Synthesizing Japanese Speech from Orthographic Text 617
H. Fujisaki, K. Hirose, H. Kawai and Y. Asano, *Dept. of Electronic Eng., University of Tokyo, Bunkyo-ku, Tokyo 113, JAPAN*

Extension of the Bandwidth of the JSRU Parallel-Formant Synthesizer for High Quality Synthesis of Male and Female Speech 313
W. Holmes, J. Holmes and M. Judd, *GEC Hirst Research Centre, East Lane, Wembley, Middlesex, HA9 7PP, U.K.*

Figure 1: Result of spotting for "Synthe*"

3 Strategies for Generating Metadata

This section describes implemented techniques for the automated extraction of the kinds metadata described in Section 2.

3.1 Word-Image Spotting

Word-image spotting refers to the task of detecting and locating user-specified keywords and phrases in images of text. Several systems for spotting whole words in scanned images of text have been developed; these systems have been found to locate keywords more accurately in noisy document images, where optical character recognition systems perform poorly [14] [15]. However, these systems require words enclosed by a correct bounding box as input.

In a word-image spotting system developed at Xerox PARC, keywords and phrases can be partially specified, similar to a simple "grep", but over images of text [3]. This system detects words in a variety of fonts. For each word identified as a keyword by the word-image spotter, the location of the word in the image can be stored as metadata. Figure 1 shows the result of spotting for "Synthe*". Note that the alternate word forms "Synthetic", "Synthesizing", Synthesizer", and "Synthesis" were detected. Because the word-image spotter identifies interword space uniquely from non-keywords, the location of the entire word can be determined during the search for a partially specified key.

The word-image spotter is based on the use of multi-resolution image morphology [1] to identify bounding boxes of text lines, and hidden Markov modeling to identify specific words within a text line. Each text line bounding box is normalized to a standard height and the width of the bounding box is scaled proportionately, producing a gray-scale image. The scaling permits recognition of words in a variety of fonts in a range of sizes.

A prespecified set of keywords in not required. Instead, for each keyword or key phrase specified by the user in a query, a hidden Markov model (HMM) is created "on the fly" from pre-trained character models. Another pre-trained HMM is used to model the data which are not part of a keyword or phrase. The non-keyword model coarsely represents the columns of pixels in a bounding box. A non-keyword model composed of all characters and symbols connected in parallel could be used, but would be much more computationally expensive.

The models are trained on data labeled with the characters appearing in each line of text and with the location of each line of text, but not the location of each character. Baum-Welch training [20] is used to estimate the parameter values of the models. To detect keywords, the keyword models and non-keyword model are connected in parallel to create a spotting network. Keywords within a bounding box are identified using Viterbi decoding [20] on the spotting network. The detected keywords and their locations in a text image can then be used as metadata.

3.2 Audio Word Spotting

Audio word spotting is the ability to locate keywords or phrases in the context of previously recorded speech. It differs from isolated word recognition, in which words to be recognized must be spoken in isolation, and continuous speech recognition, in which each word in a continuous stream must be recognized. Word spotting generates metadata in the form of time indices for the beginning and ending of the keywords. This provides indexing by keywords into long audio files, thus allowing retrieval of specific information without the need to listen to the entire recording.

Certain word spotting systems assume there are a fixed set of keywords to be spotted in continuous speech from many different talkers. An example is the operator assisted telephone call task in [30], where spotting for only five keywords is required. Such systems are based on whole word models, and require training data for each of the keywords from a large database of speakers. They are thus appropriate in tasks for which a small number of fixed keywords suffice. Other speaker-independent keyword spotting systems are based on large vocabulary continuous speech recognition. For example, the system proposed by SRI [25] uses the *Decipher*^{TM} large-vocabulary speech recognition system to transcribe the speech, and any keywords that occur in the transcription are hypothesized. A drawback of this approach is that certain keywords, for example proper names, are unlikely to be included in the vocabulary of the recognizer.

In contrast to the above speaker-independent word spotting systems is the interactive system developed at Xerox PARC [27]. The system is speaker-dependent, so that the audio is restricted to speech from a single talker. When word spotting is to be performed, the talker simply speaks the keyword or phrase to be located. Alternatively, a keyword can be manually excised from a recording. There are no linguistic assumptions, so that the word spotting system is multilingual. In addition, spotting can be performed for non-speech sounds such as music or laughter.

The PARC word spotting system uses an HMM to model arbitrary, user-defined keywords in the context of continuous speech [26]. Training the HMM consists of two stages: an initial, static stage in which statistics for a given talker are learned and a model for the non-keyword speech is obtained, and a second, dynamic stage in which the keyword model is trained as the system is in use. Data for the static training stage consists of an arbitrary segment of the talker's speech. The dynamic training stage is novel in that it requires only a single repetition of a keyword; thus, there is no distinction between keyword training and word spotting.

The search technique for locating instances of a keyword in continuous speech is a "forward-backward" search which uses peaks in the *a posteriori*, or forward [20], probability of the keyword end state to detect potential keyword endpoints. State probabilities are then recursively computed backwards to find a peak in the keyword start state. In this way, a score for the keyword is obtained in addition to the starting and ending times, which helps to prevent false alarms. This search is efficient, in that backtracking is only required when a keyword is hypothesized.

3.3 Speaker Segmentation

In speaker segmentation, the audio is partitioned into intervals, with each interval containing speech from a single speaker. The metadata derived from this consists of starting and ending times for each speaker, as well as the identity of the speaker. Pauses, or silence intervals, as well as non-speech sounds such as a musical theme, can also be identified for use in indexing. A speaker index provides the capability to access portions of the audio corresponding to a particular speaker of interest, or to browse the audio by skipping to subsequent speakers.

The basic framework for segmentation of the audio is an HMM network consisting of a sub-network for each speaker and interconnections between speaker sub-networks [28]. Speaker segmentation is performed

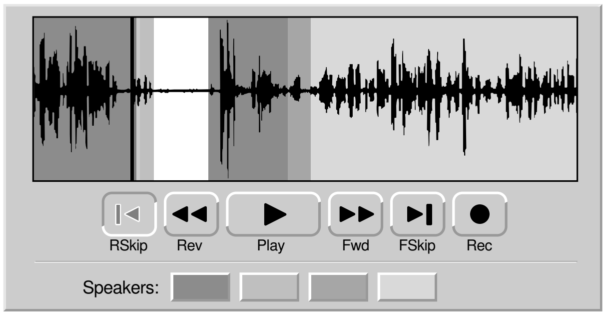


Figure 2: Audio browser with speaker segmentation

using the Viterbi algorithm [20] to find the most likely sequence of states, and noting those times when the optimal state sequence changes between speaker sub-networks. The speaker sub-networks used here are multi-state HMMs with Gaussian output distributions. In addition to modeling speakers, sub-networks are also used to model silence and non-speech sounds such as a musical theme.

In applications where the speakers are known *a priori*, and where it is possible to obtain sample data from their speech, segmentation of the audio into regions corresponding to the known speakers can be performed in real time, as the speech is being recorded. This is done by pre-training the speaker sub-networks using the sample data, and then using the Viterbi algorithm with continuous traceback for segmentation. Real-time speaker segmentation is useful, for example, in video annotation systems where annotations are made during the recording process [24].

When no prior knowledge of the speakers is available, unsupervised speaker segmentation is possible using a non-real-time, iterative algorithm. Speaker sub-networks are first initialized, and segmentation is achieved by iteratively using the Viterbi algorithm to compute a segmentation, and then retraining the speaker sub-networks based on the computed segmentation. It is necessary for the iterative segmentation algorithm to have good initial estimates for the speaker sub-networks. Thus agglomerative clustering is used to obtain an initial segmentation of the speech. This segmentation is then used in Baum-Welch training [20] of the speaker sub-networks.

Figure 2 shows how speaker segmentation can be used in an audio browser. In addition to the usual play, fast forward and reverse options, there are skip buttons to skip forward to the next speaker, or backwards to the previous speaker. Speaker buttons provide the capability to play audio corresponding to the individual speakers.

3.4 Emphatic Speech Detection

By modifying the pitch, volume, and timing, that is, the prosodics of speech, a talker can convey syntactic and semantic information, in addition to the spoken words. Prosodics can be used to alter the meaning of words, to signal whether a sentence is a statement or question, or to indicate a phrase boundary. Butzberger *et al.* [2] used prosodic information to classify isolated words as a statement, question, command, calling, or continuation. Wightman *et al.* [29] combined the use of prosodic information and word recognition information to identify intonational features in speech. Based on prosodic information, metadata can be created identifying when a question was asked, and identifying phrase boundaries for use as endpoints for presentation.

Prosody is also used in natural, conversational speech to give more emphasis to some words and phrases. When making a point, the spoken words are given greater and more frequent emphasis. This prosodic information can be exploited to serve as indices to regions of possible interest.

Emphatic speech has been found to be characterized by prominences in pitch and volume. To estimate pitch, the fundamental frequency (F0) of the glottal source is computed. To locally estimate speaking volume, energy in a short duration of the speech signal is computed. In our work at PARC, emphatic speech is identified by matching the set of prosodic features computed from a speech signal against an HMM network designed to model different prosodic patterns [4]. Prosodic features were selected which contain information to capture emphatic prominences; these features include F0, energy, change in F0, change in energy, and voicing to indicate vocalic regions.

To identify emphasized speech, syllable-based HMM's are created to model different patterns of emphatic speech. Separate models are created for unemphasized speech, which has a relatively flat prosodic pattern, for background noise, and for pauses.

A network modeling variations in emphasis is created by connecting the models of emphasized speech, unemphasized speech, and background noise in parallel. An optional pause is allowed between each of the models. Viterbi decoding [20] is used to find the best path through the network. When the best path passes through an emphatic speech model, the time indices are recorded as an emphatic region.

Regions with a high density of emphatic speech are more likely to contain parts of a conversation which a speaker wished to highlight. The time indices of these regions are stored as metadata indicating regions of possible interest for browsing.

3.5 Subtopic Boundary Location

Both automatically-identified and author-identified structural information is important for locating information in full-text documents. The structure of expository texts can be characterized as a sequence of subtopical discussions that occur in the context of one or a few main topic discussions. Subtopic structure is sometimes marked by the author in technical texts in the form of headings and subheadings. When author-identified structure is available, indices corresponding to SGML markup can be easily generated, therefore this discussion focuses only on automatically-generated structural information.

For the cases in which texts consist of long sequences of paragraphs with very little structural demarcation, we have developed an algorithm, called TextTiling, that partitions these texts into multi-paragraph segments that reflect their subtopic structure [10]. These algorithms detect subtopic boundaries by analyzing the term repetition patterns within the text. The main idea is that terms that describe a subtopic will co-occur locally, and a switch to a new subtopic will be signaled by the ending of co-occurrence of one set of terms and the beginning of the co-occurrence of a different set of terms. In texts in which this assumption is valid, the central problem is determining where one set of terms ends and the next begins. Figure 3 shows the results of TextTiling a 77-sentence popular science article. The larger peaks in the graph represent a relatively large amount of lexical coherence [9] among the words within the sentences within the peak. The valleys represent a break in lexical cohesion between adjacent text blocks. The algorithm's success is determined by the extent to which these simple cues actually reflect the subtopic structure of the text.

The core algorithm has three main parts: tokenization, similarity determination, and boundary identification. Tokenization refers to the division of the input text into individual lexical units. The text is also grouped into 20-word adjacent token-sequences, ignoring sentence boundaries in order to avoid length normalization concerns. A record of the locations of paragraph boundaries is maintained.

After tokenization, adjacent pairs of blocks of token-sequences are compared for overall lexical similarity. Token-sequences are grouped together into a block to be compared against an adjacent group of token-sequences. Actual paragraphs are not used because their lengths can be highly irregular, leading to unbalanced comparisons.

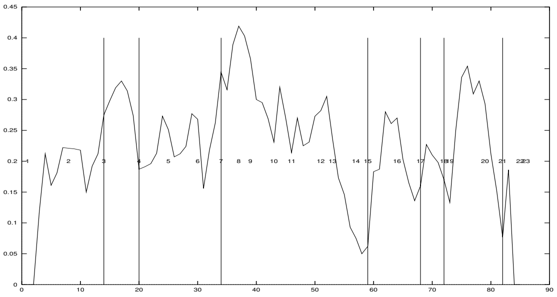


Figure 3: Result of TextTiling. The x-axis represents sentence numbers, the y-axis represents a measure of similarity between adjacent text blocks, and the vertical lines indicate actual topic boundaries as determined by human judges. Internal numbers indicate the locations of paragraph boundaries within the text.

Similarity values for adjacent blocks are computed for every token-sequence gap number. Boundaries are determined by changes in the sequence of similarity scores. The token-sequence gap numbers are ordered according to depth scores, that is, the sum of the heights of the plot on either side of the token-sequence gap. Segment boundaries are assigned to the token-sequence gaps with the largest corresponding depth scores, adjusted as necessary to correspond to true paragraph breaks. The cutoff for boundary assignment is a function of the average and standard deviations of the depth scores for the text under analysis. Currently a boundary is drawn only if the depth score exceeds a statistically motivated threshold. This method scales with the size of the document and is sensitive to the patterns of similarity scores that the algorithm produces.

4 Use of Metadata in Mixed-Media Access

Word-oriented information is present in full-length text, text image, and speech data, hence the user may expect to approach such data with a degree of uniformity. In particular, this may include accessing the media via queries in the same media type, for example, keyword spotting in speech using a spoken keyword. It also includes situations in which the media type of the query is different from that of the data, for example a spoken query to a text database. The metadata used must be flexible enough to accommodate each useful combination. This section discusses two examples of mixed-media access.

4.1 Text Access via Spoken Queries

In some circumstances mixed-media access can be facilitated by directly exploiting intrinsic properties of the data to be accessed. As an example we consider the use of speech to access information in free-text databases.

Speech recognition systems conventionally use language models [13] to aid in choosing the most likely word sequence corresponding to an utterance (e.g. "president kennedy" versus "precedent kennerty"). This data is often constructed from statistical analysis of a large amount of text about the application domain.

This requirement can be obviated by using a text database to simultaneously filter speech recognition errors and find documents relevant to a query [16]. Our system exploits the fact that the intended words of a spoken query tend to co-occur in text documents in close proximity whereas word combinations that are the result of recognition errors are usually not semantically correlated and thus do not appear together. A phonetic index associates a phonetic baseform for each word in the database. The index thus serves as a metadata representation. Each spoken word may match many candidate words. Standard boolean search with proximity constraints is used to find co-occurrences of candidate words. Based on co-occurrence information and the scores assigned by a phonetic recognizer, the most likely candidate words are found, as well as the documents in which they co-occur.

4.2 Recorded Meetings

Another application in which mixed media access is desirable is in a tool which both captures the content of a meeting and later makes its contents accessible to the participants. Meetings are complex events consisting of a duration in time as well as in space, the participation of a number of people, and the use of a number of tools for the creation, storage, rearrangement, and retrieval of information (including whiteboards, laptops, paper notes, projected viewgraphs, etc.).

Retrieval of important portions of the meeting might best be done via the identification of a combination of information types. One example might be "find the point in the meeting during which Karon was speaking emphatically and the term 'ubiquitous computing' was used." This kind of mixed-media retrieval requires coordination between a time-indexed representation of the text [24] as well as the speaker-identification index [28] and emphatic speech index [4].

5 Summary

Multimedia databases are typically accessed through text queries, often referring to manually-assigned keywords. Recently developed methods provide ways to automatically generate metadata for audio, images, and text that enable more natural access modes such as mixed-media access.

Acknowledgements

We would like to thank Jeanette Figueroa for her enthusiasm and administrative support.

References

[1] D.S. Bloomberg, "Multiresolution Morphological Approach to Document Image Analysis," In Proc. of the International Conference on Document Analysis and Recognition, Saint-Malo, France, September 1991.

[2] J.W. Butzberger Jr., M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. "Isolated word intonation recognition using hidden Markov models." In Proc. International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, New Mexico, April 1990.

[3] F.R. Chen, L.D. Wilcox, and D.S. Bloomberg. "Detecting and locating partially specified keywords in scanned images using hidden Markov models." In Proc. International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, October 1993.

[4] F.R. Chen and M.M. Withgott. "The use of emphasis to automatically summarize a spoken discourse." In Proc. International Conference on Acoustics, Speech and Signal Processing, San Francisco, California, March 1982.

[5] S. Chen, S. Subramaniam, R.M. Haralick, and I.T. Phillips. "Performance Evaluation of Two OCR Systems." In Proc. Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, April 1994.

[6] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A Practical Part-of-Speech Tagger," The 3rd Conference on Applied Natural Language Processing, Trento, Italy, 1991.

[7] *Tools for Morphological Analysis*, M. Dalrymple (ed.), Center for the Study of Language and Information, Stanford, California , 1987.

[8] A. Dengal. "The role of document analysis and understanding in multimedia information systems." In Proc. International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, October 1993.

[9] M.A.K. Halliday and R. Hasan, *Cohesion in English*, Longman, London, 1976.

[10] M.A. Hearst. "Multi-paragraph segmentation of expository text." In Proc. 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 1994.

[11] M.A. Hearst. "Using Categories to Provide Context for Full-Text Retrieval Results." In Proc. RIAO 94, Intelligent Multimedia Information Retrieval Systems and Management, Rockefeller, New York, 1994. To appear.

[12] D.J. Ittner and H.S. Baird. "Language-Free Layout Analysis." In Proc. International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, October 1993.

[13] F. Jelinek. "Self-Organized Language Modeling for Speech Recognition". In *Readings in Speech Recognition*, A. Waibel and K.F. Lee, eds. Morgan Kaufmann, San Mateo, California, 1990.

[14] S. Khoubyari and J.J. Hull. "Keyword Location in Noisy Document Images." In Proc. Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, April 1993.

[15] S. Kuo and O.E. Agazzi. "Machine vision for keyword spotting using pseudo 2d hidden markov models." In Proc. International Conference on Acoustics, Speech and Signal Processing. Minneapolis, Minnesota, April 1993.

[16] J. Kupiec, D. Kimber, and V. Balasubramanian. "Speech-based retrieval using semantic cooccurrence filtering." In Proc. ARPA Human Language Technology Workshop, Plainsboro New Jersey, March 1994.

[17] K. Markey, P. Atherton, and C. Newton, "An Analysis of Controlled Vocabulary and Free Text Search Statements in Online Searches", *Online Review* Vol. 4, pp. 225-236, 1982.

[18] R.D. Peacocke and D.H. Graf. "An Introduction to Speech and Speaker Recognition". Computer, Vol. 23, No. 8, August, 1990.

[19] L.R. Rabiner, R.W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall Inc.: Englewood Cliffs, New Jersey, 1978.

[20] L.R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Application". Proc. IEEE, Vol. 77, No. 2, February 1989.

[21] G. Salton, *Automatic text processing : the transformation, analysis, and retrieval of information by computer*, Addison-Wesley, Reading, Massachusetts, 1988.

[22] Jürgen Schürmann, Norbert Bartneck, Thomas Bayer, Jürgen Franke, Eberhard Mandler, and Matthias Oberländer. "Document analysis–from pixels to contents." In Proceedings of the IEEE, Vol. 90, No. 7, July 1992.

[23] International Organization for Standardization. "Information Processing, Text and Office systems, Standard Generalized Markup Language (SGML), International Standard; 8879" 1986.

[24] K. Weber and A. Poon. "Marquee: A Tool for Real-Time Video Logging". Proc. CHI '94, ACM SIGCHI, April 1994.

[25] M. Weintraub. "Keyword-Spotting Using SRI's $Decipher^{TM}$ Large-Vocabulary Speech-Recognition System". Proc. International Conference on Acoustics, Speech and Signal Processing, Minneapolis, Minnesota, April 1993.

[26] L.D. Wilcox and M.A. Bush. "Training and search algorithms for an interactive wordspotting system." Proc. International Conference on Acoustics, Speech and Signal Processing, San Francisco, California, March 1992.

[27] L.D. Wilcox, I. Smith, and M.A. Bush. "Wordspotting for Voice Editing and Audio Indexing." Proc. CHI '92, ACM SIGCHI, Monterey, California, May, 1992.

[28] L.D. Wilcox, F.R. Chen, D. Kimber, and V. Balasubramanian. "Segmentation of speech using speaker identification." Proc. International Conference on Acoustics, Speech and Signal Processing, Adelaide, Australia, April 1994.

[29] C.W. Wightman and M. Ostendorf. "Automatic recognition of intonational features." Proc. International Conference on Acoustics, Speech and Signal Processing, San Francisco, California, March 1992.

[30] J.G. Wilpon, L.R. Rabiner, C.H. Lee, E.R. Goldman. "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 11, November 1990.