Annotation and Segmentation for Multimedia Indexing and Retrieval

Lynn Wilcox and John Boreczky
FX Palo Alto Laboratory
[wilcox, johnb]@pal.xerox.com

Abstract

In this paper we describe a method for indexing and retrieval of multimedia data based on annotation and segmentation. Our goal is the retrieval of segments of audio and video suitable for inclusion in multimedia documents. Annotation refers to the association of text data with particular time locations of the media. Segmentation is the partitioning of continuous media into homogenous regions. Retrieval is performed over segments of the media using the annotations associated with the segments. We present two scenarios that describe how these techniques might be applied. In the first, we describe how excerpts from a video-taped usage study of a new device are located for inclusion in a report on the utility of the device. In the second, we show how sound bites from a recorded meeting are obtained for use in authoring a summary of the meeting.

1. Introduction

The inclusion of multimedia data in documents can be very effective. For example, while a text summary of a meeting can describe the decisions reached during a meeting, including sound bites from critical portions of the meeting conveys the emotional tone at the time those decisions were made. Similarly, while a text report can be generated describing the results of a study on the usability of a new device, including clips from video-tapes of user testing can describe the results more succinctly, since the user's actions are apparent in the video. Many multimedia authoring tools exist which allow inclusion of such media in documents. However, it is difficult to find the appropriate segments of the media that best illustrate the point. For example, it is difficult to locate the discussion of a particular topic in an audio recording of a two hour meeting, since it requires the user to search linearly through the recorded audio. Similarly, finding a particular segment of a video-taped usage study requires a sequential search through the video recording.

In this paper, we describe techniques for indexing and retrieval of time-based multimedia data. There are two components to our indexing scheme, namely annotation and segmentation. Annotations, or text associated with particular time locations of the media, can be obtained in several ways. One common type of annotation is closed caption text that is provided with some television broadcasts [11]. Here, we consider two types of annotations that can be acquired as a by-product of normal activity. The first is derived from notes taken while viewing the media. Text notes taken while simultaneously recording or viewing the media are linked with the media via time-stamps [12]. Handwritten annotations can be similarly obtained [19][22]. The second type of annotation is useful in situations such as usage testing where it is common to produce a text transcription of the video-taped testing sessions. Speech recognition techniques [21] are used to automatically align the transcription with the spoken text contained in the media, thus obtaining the text links to the media required for annotations.

Annotations, particularly those derived from note-taking, do not always provide complete segments of the media for presentation. For example, while a time-stamped note can indicate the approximate location of a particular topic, it does not always give the correct starting or ending time of the conversation. Thus, segmentation of the media is needed to present coherent portions to the user. Segmentation of media can be performed in several ways. For situations where a transcription has been time-aligned with the media, text-based segmentation could be used [10]. In this paper, we consider two alternate types of segmentation which operate on the media itself, namely segmentation based on audio properties and segmentation based on video properties. Both segmentation algorithms use hidden Markov models. Audio segmentation partitions the audio into intervals containing a single speaker or sound type [23][13]. Video segmentation partitions video by identifying shots, boundaries between shots, and camera motion within a shot [1].

Given a segmentation of the media and annotations associated with these segments, topic based retrieval is performed to determine the most relevant segments. The user provides a set of keywords that define the topic of

1060-3425/98 $10.00 (C) 1998 IEEE

interest. A score for each media segment is computed based on the frequencies of occurrence of the keywords in annotations associated with the segment [16]. Using these scores, an ordered list of the most relevant segments is returned to the user.

We present two scenarios that illustrate how automatic indexing can be used to facilitate authoring of multimedia documents. The first concerns the use of video-taped user testing of a new device in generating a report on its usability. In order to find segments of the video suitable for inclusion in the report, the video is first segmented and aligned with its text transcription. Search is performed over the transcription, and retrieved segments are selected for use in the report. The second example concerns the creation of a multimedia document summarizing a meeting. The meeting is digitally recorded and segmented by speaker. Then, text notes taken by meeting attendees are searched to locate discussions on topics of interest. Sound bites suitable for inclusion in the meeting summary are found using this speaker and topic information.

2. Segmentation of Continuous Media

Segmentation is the partitioning of continuous media into homogenous regions. In this paper, we concentrate on segmentation based on properties of the audio and video. This results in a partition of the media that is natural for listening or viewing. For audio-based segmentation, regions can be defined by different speakers, as in the segmentation of recorded meetings [23]. For video-based segmentation, regions are typically defined by shots, where a shot is an unbroken sequence of video frames from a single camera [1].

Our basic framework for both audio and video segmentation is provided by hidden Markov models [14]. In this framework, the audio or video is represented by sequences of feature vectors, each of which gives a short-term characterization of the signal. A hidden Markov model (HMM) includes (i) a set of M states $S_1,\ldots,S_M$, (ii) transition probabilities among the states, and (iii) state-dependent output probabilities that specify the conditional probability of observing a feature given the state. The model parameters of an HMM, namely the transition probabilities and the output probabilities, are determined from training data using the Baum-Welch algorithm [14]. This algorithm iteratively adjusts the model parameters to maximize the probability of the training data sequence. The Viterbi algorithm [14] is used to segment the media. This algorithm determines the most likely sequence of states for the observed sequence of features.

2.1. Video Segmentation

Figure 1 shows a typical hidden Markov model used for video segmentation. There are HMMs corresponding to the shots of a video, camera motion such as pans and zooms, and the transitions between shots such as cuts, fades, and dissolves. The HMM in Figure 1 is for a shot transition. The arcs show the allowed sequences of states, and are labeled with the transition probabilities. The transition probabilities model the durations of the video elements.

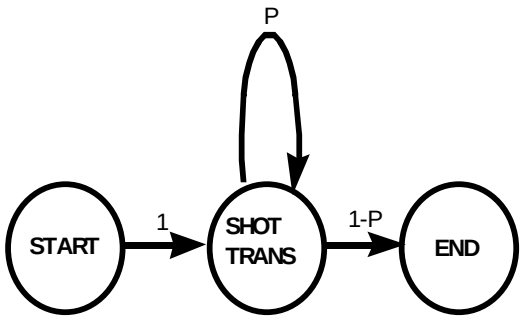

Figure 1: Hidden Markov model for a shot transition.

The features used for video segmentation are a gray-level histogram difference [24] and a motion estimator [17]. The histogram feature is useful in detecting shot transitions. It is computed as the difference between 64-bin histograms of luminance values in adjacent frames. The motion feature is computed from the coherence of the motion vectors of nine evenly distributed blocks, and indicates camera motion such as panning or zooming. The motion feature and the histogram feature are modeled with state-dependent Gaussian mixture distributions. The Baum-Welch algorithm is used to train the HMMs with approximately 5 minutes of video labeled according to shot, shot transition, and camera motion.

Figure 2 shows the video segmentation network. Transition probabilities between the HMMs are uniform. Video segmentation is accomplished using the Viterbi algorithm. The resulting sequence of states consisting of shots, camera motion, and shot transitions defines the different segments of the video.

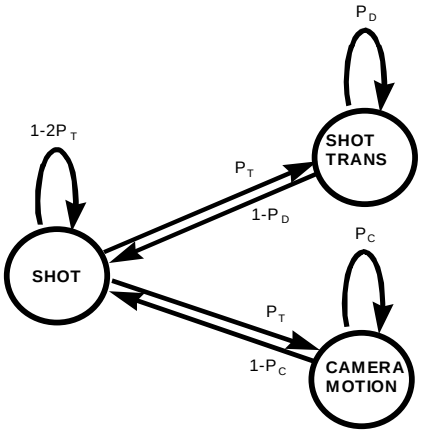

Figure 2: Video segmentation network.

Figure 3 shows an interactive video browser that displays the raw feature data, the results of the segmentation, and a keyframe for each segment to allow the user to verify the correctness of the video segmentation.

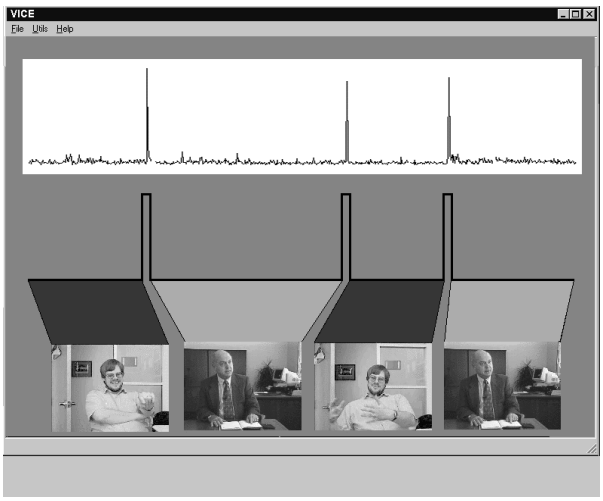

Figure 3: Interactive video segmentation tool interface.

2.2 Audio Segmentation

Audio segmentation is based on changes in speaker or sound type. Figure 4 shows a hidden Markov model used to model a speaker. It is composed of L states, where each state corresponds roughly to the different phones produced by the speaker. In our models, we use 32 states. Each state has a self transition and an exiting transition, whose probabilities model typical durations for the phones. Similar models can be used for non-speech sounds.

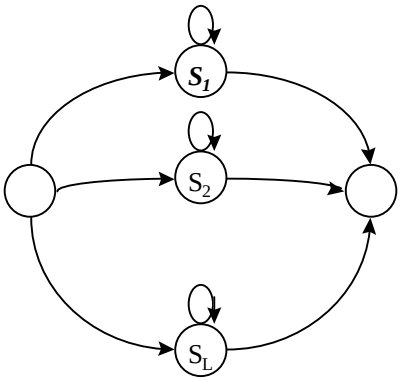

Figure 4: Hidden Markov model for a speaker.

Feature vectors correspond to cepstra [14] computed over short (20 msec) windows of sampled speech. The output distribution for feature vectors in each state is Gaussian, parameterized by a mean vector and diagonal covariance matrix. Training data, typically about one minute of speech, is used to train the speaker models using the Baum-Welch algorithm.

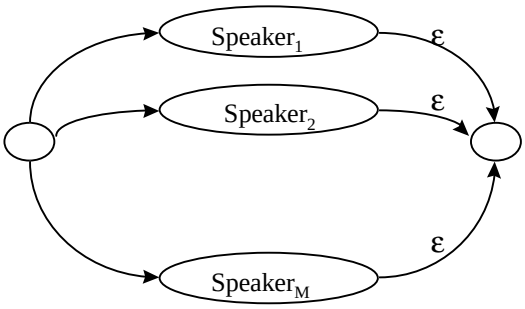

Figure 5: Speaker segmentation network.

Speaker models are combined to form the speaker segmentation network, as shown in Figure 5. The speaker models are combined in parallel. Transition probabilities into any of the speaker models are uniform, so that each speaker is equally likely. The transition probability out of each speaker model is set to ε, where ε is selected empirically to discourage speaker changes based on isolated samples.

Basic segmentation of audio is performed using the Viterbi algorithm [14]. Since each state corresponds to a speaker, segment boundaries can be determined by noting those times when the state changes from one speaker to another.

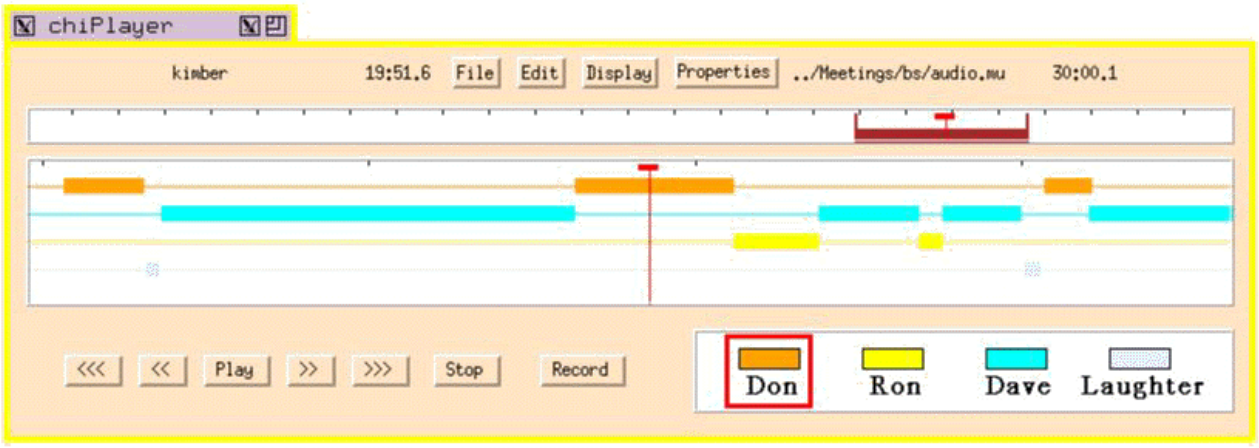

Figure 6: Audio browser showing speaker.

Segmentation results can be improved by using the iterative re-segmentation algorithm described in [23]. The speaker models are trained as above, and segmentation is performed using the Viterbi algorithm. The results of the segmentation are then used to retrain the speaker models, and a new segmentation is obtained. This process is repeated until the segmentation converges. The iterative segmentation algorithm has been shown to improve the accuracy of segmentation, particularly when the initial training data is limited or corrupted by noise or other speakers.

In some cases, it is difficult to obtain even a minute of labeled training data for each speaker. In this case, unsupervised clustering can be used to partition the audio into clusters, each of which corresponds roughly to a single speaker. The user can identify the speaker corresponding to each cluster by listening to the data. This can then be used as the initial training data for the speaker. Since this data tends to be noisy, iterative re-segmentation is used to improve results. Details of the unsupervised clustering algorithm can be found in [23]. Basically, an initial set of clusters is obtained by partitioning the audio into equal length intervals of several seconds each. The pairwise distance between clusters is computed, and the nearest two segments are merged. This process is continued until the distance between the nearest pair of clusters exceeds a pre-defined threshold. The distance between two clusters is derived from a likelihood ratio test [5].

Figure 6 shows an audio browser designed to display an audio timeline showing speaker changes [8]. Each speaker (Don, Ron, and Dave) is displayed on a different bar. Note that laughter has also been identified. This is done by treating laughter as a separate speaker.

3. Techniques for Annotation

By annotation, we mean text that is associated with the continuous media at a specified time. We consider two types of annotations. One is text notes written while the media is being recorded or viewed, along with the times when they were written. The other type of annotation we consider is a text transcript of the audio, which has been time-aligned with the spoken dialog.

3.1. Time-stamped Notes

Time-stamped notes are text notes that are associated with specific times in the media. One example of this is notes typed on a laptop during a meeting. This type of annotation was used in the meeting capture system discussed in [12]. Time stamps can be associated with each typed character or with each line of text. These time stamps are correlated with the recorded media, thus providing annotation. Text notes can be taken while the media is being recorded, or while viewing previously recorded media. Handwritten notes from a pen computer can also be associated with the meeting, as in [22] and [19]. However, since handwriting recognition is slow, text retrieval in these systems is based on manually entered text keywords rather than on the handwritten notes themselves.

3.2. Transcription Alignment

It is common to transcribe certain audio recordings for later study. For example, text transcriptions are created for presidential speeches and congressional sessions [15]. In addition, closed caption text is created for many television broadcasts [11]. In our case, video-taped usage study data is transcribed for analysis. In order for this textual data to satisfy our requirements for annotations, it must be time aligned with the recorded media. This is done using speech recognition techniques.

In speech recognition, alignment of word-level or phone-level transcriptions with audio is used to obtain training data for particular tasks or user environments. In our case, we need an approximate alignment of the text transcription with the audio so that audio can be retrieved based on its text transcription. Alignment is done using a simplified speech recognition system. Simplifications are possible because the recognizer does not have to determine the word-level transcription, but only the alignment of the known transcription with the audio.

We use the Aligner by Entropics [21] to align text transcriptions to recorded audio. The system works well assuming that the recorded audio is relatively noise free and that an accurate transcription is provided. While our recorded audio is fairly noise free, transcriptions are only accurate in terms of word content. Partial words, corrections and noises are not always marked. For example, fillers such as "um" and "er" and non-speech sounds such as coughing and laughter are not transcribed. Thus we use the interactive version of Aligner. Here, the user can interact with the system, selecting segments of audio for transcription and modifying the transcription as required. Non-speech audio is simply skipped.

4. Retrieval

The goal of retrieval is to locate segments of the media for inclusion in multimedia documents. In order to retrieve the desired segments, the user forms a query by specifying a number of keywords. While it would be possible to search the annotations for these keywords and find their corresponding time locations, this would not provide coherent segments of the media for viewing. In the case where annotations are obtained from note-taking, the time at which the words were entered may not correspond to the topic being discussed at the time, since our notes are rarely that well synchronized. For time-aligned data, text and audio are perfectly synchronized, but knowing the times when keywords occur still does not produce meaningful segments of the media. For this reason, we perform keyword search over segmented media.

Prior to retrieval, the media is segmented using the audio or video based segmentation as described above. This provides segments containing either a single speaker or a single camera shot. The audio or video segmentation also implies a segmentation on the text annotations, since annotations are linked to specific points in time. Keyword retrieval is performed over text annotations contained in these segments. Thus it is possible to use not only locations of the keywords, but their frequency of occurrence.

We use the common term frequency and inverse document frequency [16] weighting for keywords. Let tf_i be the frequency of keyword i in a segment, or the term frequency, let n_i be the number of segments which contain keyword i, and let N be the total number of segments. The inverse document frequency for keyword i is defined as $log(N/n_i)$. The weight w_i for keyword i is the product of the term frequency and the inverse document frequency, or

$$w_i = tf_i \, log(N/n_i).$$

Retrieval is performed by computing a score for each segment, where the score is the sum of the weights for keywords in the query. (We assume the keywords are weighted equally.) The result of the query is a list of segments ordered by their scores, where the higher scoring segments should be more relevant to the query. The user can then review the segments in order and select the appropriate ones.

A limitation of the current system is that only those keywords that actually occur in the annotations are useful for retrieval. While a thesaurus could be used to augment the query keywords, we take the simpler approach of suggesting query keywords to the user. A list of keywords, ordered by the term frequency and inverse document frequency weighting described above, is presented to the user. Since keywords with high weights are ones that occur frequently in only certain segments of the media, they are the most promising candidates for retrieval. In the relevance feedback experiments discussed in [9], a similar technique for keyword suggestion was used successfully. We also allow the user to search the list of keywords alphabetically.

In cases where the audio is segmented by speaker, we give the user the option of specifying the speaker in addition to the keywords in the query. This can be useful in cases where the user wishes to retrieve comments from a specific person on the topic of interest.

5. Applications

We give two examples of how segmentation and annotations can be used for the retrieval of audio and video clips for authoring multimedia office documents. In the first, excerpts from video-taped usage testing of a new device are located for inclusion in a report on the utility of the device. In the second, sound bites from a recorded meeting are retrieved for inclusion in a document summarizing the meeting.

5.1. Usage Study Video

In the development of a new computer device, it is common to conduct usage studies to determine how well people interact with the device, and what changes need to be made before marketing the device. These usage studies are typically video-taped for subsequent analysis. In addition, it is common to obtain a text transcription of the video, so that analysis can be performed using text as well as video. The annotation and segmentation techniques described above are used to index the video, so that relevant clips can be retrieved for authoring.

The video is first segmented using the techniques described in Section 2.1. Video-taped usage studies typically contain several segments corresponding to different people using the device. These are easily detected as cuts. In addition, the video will contain camera motion corresponding to the change of focus from the user to the device and visa-versa. This motion is detected and the appropriate video segments are formed. The transcription is then aligned with the video using the interactive mode of the Aligner, as described in Section 3.2.

To retrieve portions of the video on a specific topic, the user specifies a number of keywords. The user can select the keywords from a list ordered by retrieval value, or can simply type in keywords. The system returns an ordered list of video clips, along with the aligned transcriptions. The user can then view the clips, and select the relevant ones for inclusion in a report.

5.2. Recorded Meeting

The recording of meetings is becoming increasingly common. In addition, systems exist which allow users to take notes electronically, so that their notes are synchronized with the meeting [12][22]. Our annotation and segmentation techniques can be used to index the meeting, so that relevant portions can be included in a multimedia meeting summary.

The recording of the meeting is first segmented according to speaker using the techniques described in Section 2.2. Unsupervised clustering is used to obtain an initial set of sound clusters. These clusters are then labeled by speaker and used to train a speaker model for each participant. The iterative re-segmentation algorithm is then used to segment the audio according to speaker. We assume that a note-taking device such as that in [12] is used to obtain text notes synchronized with the audio.

In order to retrieve sound bites from the recorded meeting in which a particular topic was being discussed, the user specifies a number of keywords. In addition, the user has the option of specifying the speaker. The system returns an ordered list of audio segments, along with the associated annotations, from which the desired segment can be chosen for inclusion in the meeting summary.

6. Related Work

The Informedia project at CMU [18] aims to establish a digital video library with content search and retrieval. Text content for the video is obtained from closed captions aligned with the audio, or ideally from automatic transcription using a large vocabulary connected speech recognizer. Video retrieval is done via text or voice queries. Video segmentation is used to provide video summaries of the retrieved data.

In the news video browsing system described in [10], video clips for browsing are generated by segmenting closed caption text associated with the video. Text segmentation is performed using a modification of the Text-Tiling algorithm [6], or by using discourse-based clues specific to the news format.

In the news story search system described in [11], information from the closed caption text is used to segment the news into stories. Closed caption text is synchronized with the video at segment boundaries using audio and video cues. Keyword-based retrieval for news stories is performed.

In the broadcast news retrieval developed as part of the Medusa project [3], closed caption (teletext) is divided into segments which are indexed. No alignment of video and text is done, and retrieval results in an entire news broadcast.

The Jabber system [7] provides multiple methods for indexing, browsing, and retrieval of video conferences. A speech recognizer is used to obtain a partial transcription of the meeting, and topics or themes are identified using lexical chaining. In addition, people's interaction patterns (*e.g.* discussion, presentation) are identified and can be used for indexing.

In [15] speaker identification is used to obtain a rough alignment of audio recordings of congressional sessions to their text transcripts. The text transcripts contain speaker information, so that the sequence of speakers in the audio is known. Speaker identification is used to identify points in the audio where the speaker changes. However, this does not provide word-by-word alignment.

In the Mixed Media Access system [4], indexing and retrieval are performed across a variety of media. In particular, speaker indexing and keyword indexing are performed on the audio, while text segmentation is

performed on associated text. Retrieval is based on queries across the media, however no attempt is made to align segments of the various media.

The video mail retrieval project at Cambridge University and Olivetti Research Limited [2] uses word-spotting from a phone lattice to retrieve voice mail messages from a text query. In [20], audio is automatically indexed by generating a phonetic transcription. Audio is retrieved using a text query and fuzzy matching.

7. Summary

This paper has described the use of annotation and segmentation for the indexing and retrieval of multimedia data. Our goal was to simplify the task of finding segments of audio and video data suitable for inclusion in multimedia documents. Segmentation of the media was performed to identify coherent portions of the media, while annotations provided the means for text-based search.

We considered two methods for obtaining text annotations of the media, both derived from natural practices. The first method used notes taken while viewing the media. The second used speech recognition techniques to time-align a transcription with the media.

Two types of segmentation were used, both performed on the media itself. The first was audio segmentation, which identified portions of the media corresponding to a single speaker or sound. The second was video segmentation, which located sequences from a single camera shot.

We described how annotation and segmentation allowed retrieval of relevant segments of the media in response to keyword inputs by a user. We presented two scenarios in which such techniques might be used to facilitate authoring of multimedia documents.

12. References

[1] J. Boreczky and L. Rowe. "Comparison of Video Shot Boundary Detection Techniques," *Storage and Retrieval for Image and Video Databases IV, Proc. SPIE 2670*, SPIE, San Jose, CA, February 1996, pp. 170-179.

[2] Brown, J. Foote, G. Jones, K. Spark Jones, and S. Young. "Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval", *Proceedings of ACM Multimedia 96*, ACM, Boston, MA, November 1996, pp. 307-316.

[3] Brown, J. Foote, G. Jones, K. Spark Jones, K.S. Jones, and S. Young. "Automatic Content-Based Retrieval of Broadcast News", *Proceedings of ACM Multimedia 95*, ACM, San Fransisco, CA, November 1996, pp. 35-43.

[4] F.Chen, M. Hearst, D. Kimber, J. Kupiec, J. Pedersen, and L. Wilcox, "Metadata for Mixed Media Access", *SIGMOD Record, vol 23, no. 4*, ACM, December 1994, pp. 64-71.

[5] H.Gish, M.H. Siu, and R. Rohlicek. "Segmentation of Speakers for Speech Recognition and Speaker Identification", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,vol. 2*, IEEE, Toronto, Canada, May 1991, pp. 873-876.

[6] M.A. Hearst. 'Multi-Paragraph Segmentation of Expository Text", *Proceedings 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994.

[7] R.Kazman and J. Kominek. "Supporting the Retrieval Process in Multimedia Information Systems", *Proceedings of the Thirtieth Annual Hawaii International Conference on Systems Sciences*, IEEE, Wailea, HA, January 1997.

[8] D.G. Kimber, L.D. Wilcox, F.R. Chen, and T.P. Moran. "Speaker Segmentation for Browsing Recorded Audio,. *Proceedings of CHI: Conference Companion*, ACM, Denver, CO, May 1995, pp. 212-213.

[9] J.Koenemann and N.Belkin, "A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness". *Proceedings of CHI 96*, ACM, Vancouver, BC Canada, April 1996, pp. 205-212.

[10] I.Mani, D. House, M. Maybury, M.Green. "Towards Content-Based Browsing of Broadcast News Video", *Intelligent Multimedia Information Retrieval*, M. Maybury, ed., AAAI Press, 1997, pp. 241-258.

[11] R.Mohan. "Text-Based Search of TV News Stories", *Proceedings of SPIE Conference on Multimedia Storage and Archiving Systems* SPIE, Boston, MA, November 1996 pp. 2-13.

[12] T. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, "I'll Get That Off The Audio: A Case Study of Salvaging Multimedia Meeting Records", *Proceedings of CHI 97*, ACM, Atlanta, GA, March 1997, pp. 202-209.

[13] S. Pfeiffer, S. Fischer, W. Effelsberg. "Automatic Audio Content Analysis," *Proceedings of ACM Multimedia 96*, ACM, Boston, MA, November 1996, pp. 21-30.

[14] L.R. Rabiner and B. Juang. *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[15] D.Roy and C. Malamud. "Speaker Identification Based Text To Audio Alignment for and Audio Retrieval System", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 2*, IEEE, Munich, Germany, April 1997), pp. 1099-2002.

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

[17] B. Shahraray. "Scene Change Detection and Content-Based Sampling of Video Sequences," *Digital Video Compression: Algorithms and Technologies, Proc. SPIE 2419*, IEEE, February, 1995, pp. 2-13.

[18] H.D. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens. "Intelligent Access to Digital Video: The Informedia Project", *IEEE Computer*, IEEE, May 1996, pp. 46-54.

[19] K.Weber and A. Poon. "Marquee: A Tool for Real-Time Video Logging", *Proceedings of CHI '94*, ACM, Boston, MA, April 1994), pp. 58-64.

[20] M. Wechsler and P. Schauble. "Metadata for Content Based Retrieval of Speech Recordings", *SIGMOD Record, vol. 23, no. 4* , ACM, December 1994.

[21] C.Wightman and D.Talkin, "The Aligner". *Entropic Research Laboratory* 1994.

[22] L. Wilcox, B. Schilit, N. Sawhney, "Dynomite: A Dynamically Organized Ink and Audio Notebook", *Proceedings of CHI 97* , ACM, Atlanta, GA, March 1997, pp. 186-193.

[23] L.D. Wilcox, F.R. Chen, D. Kimber, and V. Balasubramanian. "Segmentation of Speech Using Speaker Identification", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing vol. S1*, IEEE, Adelaide, Australia, April 1994, pp.161-164.

[24] H.J. Zhang, A. Kankanhalli, and S. Smoliar. "Automatic Partitioning of Full-Motion Video," *Multimedia Systems* vol. 1, no.1, 1993, pp. 10-28.