Organization and Retrieval of Continuous Media

Yasuo Ariki
Dept. of Electronics and Informatics
Ryukoku University
Seta, Otsu-shi, 520-2194, Japan
ariki@rins.ryukoku.ac.jp

ABSTRACT

Because of the media digitization, a large amount of information such as speech, audio and video data is produced everyday. In order to retrieve data quickly and precisely from these databases, multimedia technologies for organizing and retrieving of speech, audio and video data are strongly required. In this paper, we overview the multimedia technologies such as organization and retrieval of speech, audio and video data, speaker indexing, audio summarization and cross media retrieval existing today. The main purpose of the organization is to produce tables of contents and indices from audio and video data automatically. In order to make these technologies feasible, first, processing units such as words on audio data and shots on video data are extracted. On a second step, they are meaningfully integrated into topics. Furthermore, the units extracted from different types of media are integrated for higher functions.

1. INTRODUCTION

Contents stored in digital libraries or museums can be accessed quickly using their manually constructed table of contents or indices. On the other hand, broadcast news, animation, drama and movies usually have no table of contents or indices, so that the data have to be accessed sequentially. To solve this problem, content based access architecture is required for so called “continuous media”.

In video images, index corresponds to an event occurred in the video image. On the other hand, a table of contents corresponds to a hierarchical structure of topics, like chapters and sections of books. In order to access the continuous media, these indices and tables of contents have to be produced by automatically analyzing video and audio data. This is called “organization of continuous media” which includes mainly functions of indexing and topic segmentation.

In this paper, techniques are described for organizing and retrieving continuous media, together with our developing system, especially for TV news programs. Since TV news programs are broadcast from all over the world through the digitization, TV viewers require to select and watch the most interesting news in a short time. In order to satisfy this requirement, many functions have to be supplied for a TV news database, such as automatic topic segmentation, retrieval of related topics and summarization. These functions can be realized based on the organization of continuous media.

In section 2, an approach to the organization is described and an overview of our system is described in section 3. In section 4 and 5, organization and retrieval functions developed for audio data and video data are described. Integration of different types of media is described in section 6.

2. APPROACH TO THE ORGANIZATION

In order to realize the indexing and topic segmentation for continuous media, indices such as words, objects and shots have to be extracted automatically from audio and video data at first. These techniques are called media analysis. The indices are meaningfully integrated into topics through their combination. This technique is called media integration. By these indexing and topic segmentation, new functions, such as cross media retrieval which can retrieve the content across the different media, will be prepared. This technique is called media application. These three techniques of media analysis, media integration and media application should be incorporated into the organization of multimedia content, especially continuous media.

Fig.1 shows the organization process of continuous media. Signal data is organized into contents through the processes of segmentation & classification, recognition & indexing, association in time and space, and finally construction of topic thesaurus as shown in the left hand side. According to these organization processes, the signal data is converted into concept through pattern, symbol and topic as shown in the right hand side.

At each data presentation, the corresponding retrieval is available. For example, at the signal level, specific images, music or noises can be retrieved based on signal processing. At the pattern level, human faces and speech are retrieved using pattern extraction techniques. An example at the symbol level is a retrieval of president faces and speech based on pattern recognition techniques. At the topic level, president speech concerning about his economic policy can be retrieved or highlight scenes in TV sports program can be retrieved.

Finally at the concept level, the upper or lower topics as well as similar or opposite topics will be retrieved. Here the upper topics mean not only the summaries but also the topics at abstract level such as economy to bankruptcy. In the same way, the lower topics mean not only the details but also the topics at concrete level. These retrieval of opposite, upper and lower topics based on automatic construction of topic thesaurus rather than summaries or detail is uniques approach compared to Informedia-II being developed at CUM. Informedia-II focuses on distillation of information across multiple video paragraphs and visualization over time and space to understand how events evolve and are correlated over time and geographically[1].

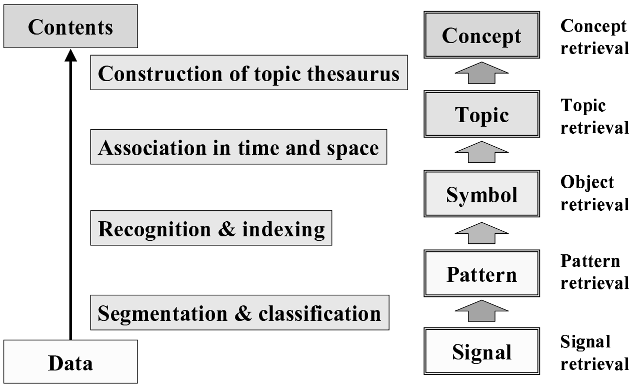

Figure 1: Organization process of continuous media.

3. OVERVIEW OF THE SYSTEM

Fig.2 shows an overview of the system we are developing for a TV news database. It consists of four main modules; (1) organization and retrieval of speech and audio data, (2) organization and retrieval of video data, (3) Media integration, (4) Interactive retrieval which utilizes the multimedia technologies in an interactive fashion.

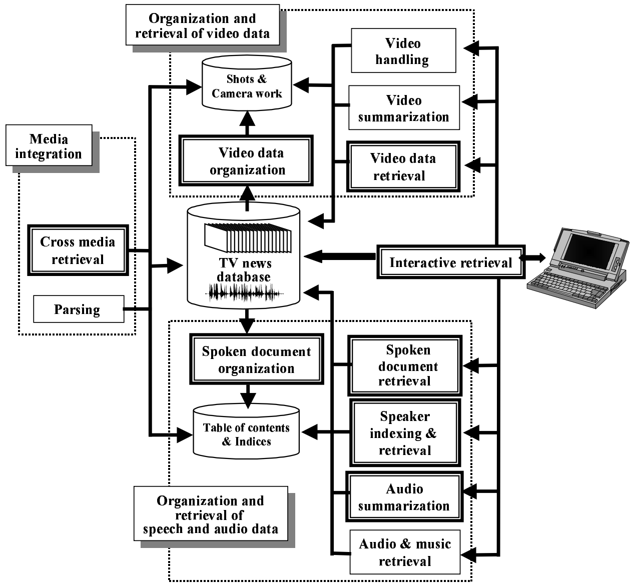

Figure 2: Organization of the system

The organization and retrieval module of speech and audio data, shown in the bottom of Fig.2, supplies the system with a table of contents and indices of TV news through speech transcription and article (topic) segmentation carried out at the spoken document organization module. These TV news articles can be retrieved at the spoken document retrieval module using the indices. Then the retrieved articles are summarized at the audio summarization module. The speaker indexing & retrieval module gives information about who is speaking. The audio & music retrieval module retrieves and classifies audio data into speech and music and searches for the audio signals.

The organization and retrieval module for video data, shown in the top, supplies the system with information about shots and camera works extracted at the video data organization module. These shots and camera works are retrieved at the video data retrieval module. The TV news articles retrieved are summarized at the video summarization module. The video handling module makes it possible to interactively manipulate the video data to extract the most interesting information for users.

The media integration module, shown in the left, supplies the system with a parsing and cross media retrieval functions. The parsing module can segment the TV news into several topics by integrating video and audio data. The typical example of the parsing function is a case where TV news are extracted automatically from TV news program by separating commercials. Audio and video cues are both utilized for TV news extraction. The other example is to detect the topic changes in the debate program. Audio and video cues will be utilized for this topic segmentation. The cross media retrieval module can retrieve news speech articles, which are transcribed and segmented at the spoken document organization module, by using different media such as video captions (speech retrieval by video caption).

The interactive retrieval module, shown in the right, supplies the system with functions to accept interactive queries from users. For example, if a user asks the system the name of a person acting on a TV display, then the system recognizes his face and retrieves his name and profile by consulting a face-name dictionary. The other example is that if a user asks the system unknown words uttered by a TV news announcer, then the system recognizes the unknown words and retrieves the meaning by consulting a word dictionary.

In the following section, more details of these modules are described mainly focusing on the modules which are enclosed with bold rectangles on Fig.2.

4. AUDIO DATA ORGANIZATION

4.1 Spoken Document Organization

Spoken documents are produced through transcription of speech data such as TV news program or documentary program. Spoken document organization is a technique to produce a table of contents and indices from the spoken documents. The table of contents can be produced by finding topics (stories or articles) and their boundaries. This process is usually called topic word extraction or topic segmentation. On the other hand, the indices can be produced by finding keywords included in the spoken documents.

The topic word extraction and topic segmentation are carried out as follows. The flow is shown in Fig.3.

(1) Keywords are selected from news article database.
(2) A word sequence is produced by speech transcription for speech news database.

(3) Keywords are extracted from the produced word sequence.

(4) Topic models are constructed from news article database.

(5) Topic words and topic boundaries are extracted using the keywords and topic models.

The detail of each process is described below.

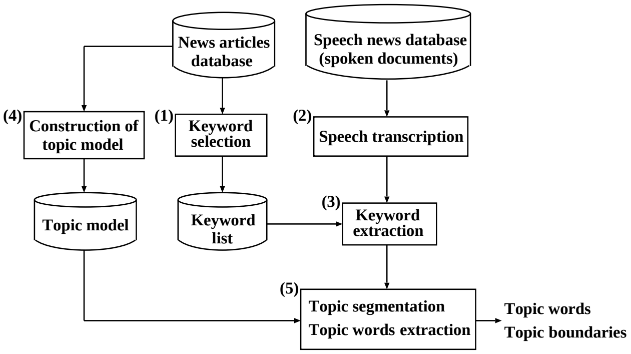

Figure 3: Flow of spoken document organization

4.1.1 Keyword Selection

In the selection of keywords from news corpus, mutual information between the word w_i and the topic t_j is mainly used. Mutual information $i(t_j;w_i)$ is a measure to indicate what amount of information concerning the topic t_j is obtained from the word w_i so that the word w_i with high mutual information strongly contributes to identify the story topics. The equation of mutual information is given as follows;

$$\begin{aligned} i(t_j;w_i) &= i(t_j) - i(t_j|w_i) \\ &= -\log P(t_j) + \log P(t_j|w_i) \\ &= \log \frac{P(t_j,w_i)}{P(t_j)P(w_i)} \end{aligned} \tag{1}$$

Other measures used to select keywords are χ^2 value and *tf-idf*[2] as shown below;

$$\begin{aligned} \chi^2_{i,j} &= \frac{(x_{i,j}-m_{i,j})^2}{m_{i,j}} \\ m_{i,j} &= \frac{\sum_{j=1}^{n} x_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{i,j}} \times \sum_{i=1}^{m} x_{i,j} \end{aligned} \tag{2}$$

m : The number of different words

n : The number of topics

$x_{i,j}$: Frequency of the word w_i in the topic t_j

$m_{i,j}$: Predicted frequency of the word w_i in the topic t_j

$$\textit{tf-idf} = TF(w_i,t_j)\cdot IDF(w_i) \tag{3}$$

$TF(w_i,t_j)$: Frequency of the word w_i in the topic t_j

$IDF(w_i)$: $\log \dfrac{\text{The number of total topics}}{\text{The number of topics including } w_i}$

4.1.2 Speech Transcription

In producing word sequence W from news speech S, speech transcription is mainly performed using acoustic model of phoneme HMMs and bigram or trigram language model[3, 4]. The phoneme HMMs are trained using news speech database or JNAS speech database. Fig.4 shows a flow of this speech transcription process.

In our system, cross-word triphone HMMs were trained using 21,782 sentences (ASJ continuous speech, ASJ pseudo dialogue) spoken by 137 males. MFCC with power and its delta features was employed. The language models were constructed using 45 months newspaper text. The word dictionary contains 20,000 words. The word accuracy obtained by automatically transcribing 55 news stories (articles) included in the speech database was 80.3%. After news speech transcription, keywords are extracted from the obtained word sequence.

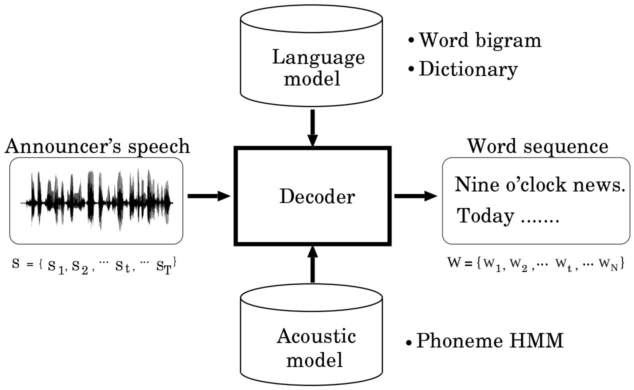

Figure 4: Flow of speech transcription process

4.1.3 Topic Model and Topic Segmentation

A topic model maps a group of keywords $\{w_i\}$ included in a news story to topic words $\{t_j\}$. Several topic words may be produced for a news story. In constructing topic models, a topic vector $v_{topicj} = (w_1, w_2, \cdots, w_N)^T_j$ is generated for each topic word t_j in terms of keywords. For this purpose, a group of news stories with a topic word t_j is manually collected at first. Then frequency of keywords w_i, mutual information $i(t_j;w_i)$, χ^2 value or *tf-idf* is computed as the element of the topic vector.

Given a news story, an article vector $v_{article} = (w_1, w_2, \cdots, w_N)^T$ is generated by counting the number of keywords included in the news story as vector elements. Then inner product is computed between the article vector and the topic vectors as follows.

$$f(t_j) = v^T_{topicj}\cdot v_{article} \tag{4}$$

This inner product $f(t_j)$ is called a topic function. The topic words t_j with value of the topic function $f(t_j)$ higher than some threshold are selected as the topic words for the input news story. This method is called vector space model owing to vector presentation of the news articles. When topic segmentation is required, topic sections are detected where the highest value of the corresponding topic function $f(t_j)$ continues.

If the topic vectors are not available, an analytical window is shifted on the given news story. On each analytical window, the topic vector is generated whose elements are frequency of keywords in the window. Then the inner product is computed between the consecutive windows. If the inner product is greater than some threshold, these two windows are regarded as belonging to the same topic section. This is called unsupervised topic segmentation.

In our experiments, the topic extraction rate, which is defined by the ratio of correctly detected boundaries, was 67.3% for the speech database by using the topic function. On the other hand, in unsupervised topic segmentation, the recall and precision rate of the topic boundaries were 90.7% and 40.2% respectively. The topic segmentation is not so high at present due to its difficulty.

The other topic model is proposed by using HMM whose state presents respective topics[3]. In each state, the word occurrence probability is estimated. Given a time sequence of news stories, the Viterbi algorithm is applied and the corresponding topics are extracted as well as the topic boundaries.

4.2 Spoken Document Retrieval

Spoken document retrieval (SDR) is strongly required to retrieve documents related to input queries from large spoken documents. In order to improve the retrieval capability, TREC (Text REtrieval Conference) is held every year and SDR for TV news speech is competed in the conference [5].

The SDR is carried out as shown in Fig.5 and described as follows;

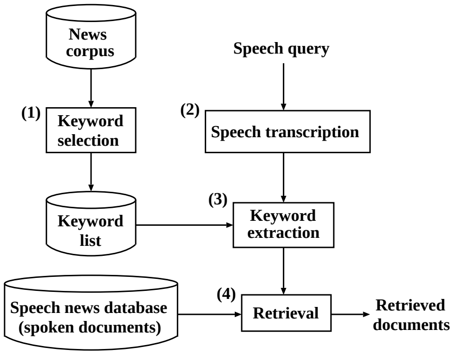

Figure 5: Flow of spoken document retrieval

(1) Keywords are selected from the news article database.

(2) A word sequence is produced by transcribing the speech query and speech news data.

(3) Keywords are extracted from the produced word sequence.

(4) Spoken documents are retrieved using a group of keywords.

Steps from (1) to (3) are the same as those described in section 4.1. The main step of SDR is step (4). The keywords extracted from spoken query are further filtered by using the knowledge of word co-occurrence to remove the unrelated keywords which are produced due to speech recognition error. The spoken document is retrieved using these coherent keywords by full text scanning or vector space model.

In our experiments, recall rate was 58.5% by vector space model using TF-IDF as the keyword extraction method. This value was obtained under the condition that the recall value almost equaled precision value. There is a strong association between word accuracy of speech transcription and the retrieval accuracy of spoken documents so that the improvement of speech transcription is a key to improve the spoken document retrieval accuracy.

The similar study as SDR is cross language information retrieval (CLIR) which retrieves the spoken documents related to an input spoken document, but the language is different between them. In this study, word translation is a key technique which converts coherently a group of words included in an input spoken document in one language into a group of words in the different language[6].

4.3 Speaker Indexing & Retrieval

Speaker indexing is a technique to identify sentences spoken by the same speaker and to give speaker indices to the spoken sentences. Speaker retrieval is performed by using this speaker indices. This task is required, for example, to retrieve a speech of a famous person. The typical example is to “find the president’s speech from the interview”.

In this speaker indexing, speaker recognition or speaker verification techniques are incorporated. When the speaker recognition technique is utilized, the speaker is recognized by comparing the input speech with the speaker models constructed in advance[7]. On the other hand, when the speaker verification is utilized, it is verified whether the same speaker is still speaking or not on the basis of GLR(Generalized Likelihood Ratio) [8] [4]. In this case, the speaker models are not prepared in advance but dynamically constructed in a course of speaker indexing. As basic techniques of speaker recognition or verification, VQ (Vector quantization)[9], GMM (Gaussian mixture model) and subspace method are available.

4.4 Audio Summarization

Audio summarization is a technique to reduce the amount of speech data without losing the semantic meaning. This task is important for quickly browsing long speech data such as TV news or documentary program. The most popular method is to apply a speech transcription technique to input speech and produce the word sequence. From this word sequence, sentences with the important keywords are extracted as a summary.

A completely different method, which does not perform the speech transcription, is being developed by the RWC project[10]. They perform IRIFCDP (incremental reference interval-free continuous dynamic programming) to search the same words uttered in continuos input speech. The IRIFCDP is a kind of continuous DP where the speech template is the same as the input speech so that the template becomes incrementally long as a function of time. The repeatedly spoken words which show local minimum of locally accumulated distance are detected. These repeated words are regarded as important words. Speech is summarized by picking up these repeated words extracted by IRIFCDP. Topic segmentation can also be performed by using the repeated words[10].

5. VIDEO DATA ORGANIZATION

5.1 Video Data Organization

Movies or dramas may be appreciated at home. However, news, sports and variety programs are selectively watched according to our interest. For this purpose, video organization is required which can segment the video data into sections according to the homogeneity of the content.

It is difficult to organize video data semantically because extraction of semantic meaning from video data is difficult. Therefore video organization is performed syntactically in most cases such as cut detection which allows dividing the video data into shots. For semantic organization, speech and video caption should be analyzed together with video data because they have linguistic information. The organized video is used for video browsers in order to search for topics or events.

The techniques to syntactically organize the video data are summarized as follows;

(1) Cut detection: The difference between consecutive frames is computed based on a histogram method and if it is greater than some threshold, it is regarded as a cut point. The shot is extracted as the section ended at the cut points[11].

(2) Extraction of common scene: Common scenes included in long video data give information about the repetition.

(3) Extraction of camera works: Camera work indicates the intention of a cameraman or director. Therefore it is possible to segment the video data in terms of homogeneous camera work within a shot. The camera work can be extracted by a projection method or Affine parameters[12].

5.2 Video Data Retrieval

There are various types of retrieval queries for video data. The most popular query is to search for highlight scenes such as home runs or hits in a baseball game, a particular goal kick or pass in a soccer game or a smash or volley in a volleyball game. These are based on action analysis and recognition of the objects filmed on video data. These recognition results are used as indices in the retrieval.

Other example of a retrieval query takes place when there are video captions superimposed on video data. If the video captions are extracted from TV news video and collected as a video caption list, users can quickly access their interesting news content by pointing one of the captions in the list displayed on the TV.

If particular scenes such as a flight taking off, the Eiffel tower or the Louver in Paris can be retrieved from video database, then we can produce our personal travel movies. The most difficult retrieval is content retrieval such as searching for a scene where a bird just came out of his nest and is flying. For this kind of retrieval, the content description may be required.

The techniques for video retrieval are summarized as follows;

(1) Retrieval of interesting frames, shots and scenes by recognizing objects such as human faces, building or characters.

(2) Retrieval of video captions after object extraction and recognition[13].

(3) Action retrieval after extraction of moving objects and action recognition by dynamic programming or HMM (Hidden Markov Model)[14].

(4) Event retrieval in sport game by object extraction, action recognition and event recognition[15].

(5) Content retrieval based on content description.

6. MEDIA INTEGRATION

6.1 Technologies for Cross Media Retrieval

Cross media retrieval is a technique to retrieve images, video or text documents by speech query and vice versa[16, 17]. In order to realize this retrieval, matching between different media must be performed. In technical paper[16], existence information of speech, female speech and scene cut is extracted automatically from audio and video data at every sampling time and a time sequence of vectors consisting of these three features are produced. In the same way, a similar sequence of vectors are extracted from drama script. Then their two vector sequences are matched by dynamic programming method. After this DP matching, given a part of speech or video, the corresponding drama script can be retrieved or vice versa.

Name-It is a system to associate faces appearing in news programs or dramas with their names automatically by their co-occurrences[18]. Therefore the Name-It system can retrieve human faces by their name query and vice versa.

The RWC project has been researching how to associate the image, speech and text[17]. They present speech as a path of keywords which are extracted by IRIFCDP and recognized by the word recognizer. This speech presentation is associated with a path of word sequence included in a text document. If the text document contains an image, the similar images are searched in the image feature space. In this way, speech, text and images are retrieved in a cross media mode.

6.2 Speech Retrieval by Video Captions

We have also developed a system of cross media retrieval for TV news as shown in Fig.6. At first, the video caption is recognized through video OCR. Then a vector is constructed for a video caption using the recognized nouns. Next, speech transcription of a news story is converted into a vector using extracted important words.

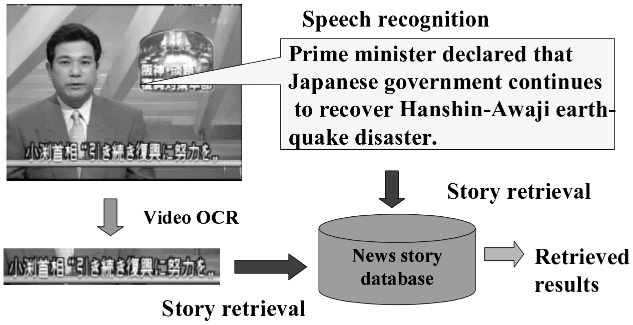
Speech recognition
Prime minister declared that Japanese government continues to recover Hanshin-Awaji earthquake disaster.
小渕首相"引き続き復興に努力を"
Video OCR
小渕首相"引き続き復興に努力を"
Story retrieval
News story database
Story retrieval
Retrieved results

Figure 6: Integration of video OCR and speech transcribing.

The widely used vector space model counts only the number of common words between documents, but it doesn't take into consideration the similarity between non-common words. This causes a decrease of the retrieval performance in a case where the number of words included in video captions is small. To solve this problem, word distance have to be computed based on similarity between words.

Though mutual information and co-occurrence are usually used as the similarity measure, they show the similarity between words in only one dimensional space. On the other hand, word distance is defined as the distance between words in a three-dimensional space of mutual information, TF and IDF. We call this three-dimensional space as a word space.

The method to compute the inner product of two vectors, a vector for video caption and a vector for spoken document, is expanded by incorporating this word distance into vector space model. The experiments to retrieve the spoken documents by the video caption showed 91.7% recall and 64.7% precision. This is extremely higher than the conventional vector space model which resulted in 33.3% recall and 66.7% precision.

6.3 Speaker and Speech Integration

We have developed a system which can automatically index and classify TV news stories into 10 topics based on a speech transcribing technique[4]. After transcribing the spoken news stories, pre-defined keywords are searched and the news stories are classified based on the keywords.

In general, news speech includes reporter speech as well as announcer speech. The announcer speech is clear but the reporter speech is sometimes noisy due to wind or environmental noises so that the transcription accuracy of the reporter speech is lower than that of the announcer speech. Therefore if the speech transcribing process is limited only to the announcer speech, the processing time can be reduced without decreasing the news classification accuracy.

From this viewpoint, our system can automatically divide the TV news speech into speaker sections at first and then index in real time who is speaking. This can be realized by using a technique of speaker verification. However, the speaker verification technique is sensitive to the time lapse. To solve this problem, speaker models are not prepared in advance but are constructed in a course of indexing in self-organization mode.

We verified the effectiveness of our proposed methods by carrying out the experiment of extracting and transcribing only the announcer speech and then classifying the news stories into 10 topics. At present, 75% accuracy of news story classification is obtained. This system is applicable to meeting documents as shown in Fig.7.

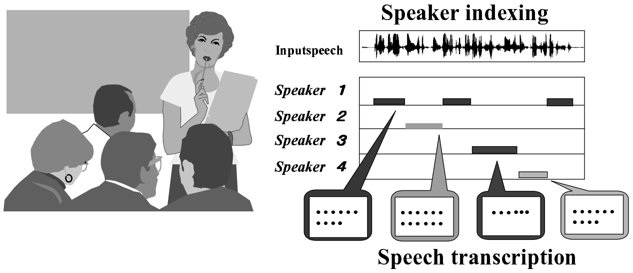

Figure 7: Integration of speaker indexing and speech transcribing.

7. CONCLUSION

In this paper, multimedia techniques for organization and retrieval of continuous media were described together with our developing system from the viewpoint of spoken document organization & retrieval, speaker retrieval, video organization & retrieval and cross media retrieval. In organization and retrieval, techniques to produce automatically a table of contents and indices from multimedia data were described. In order to access spoken documents and video data, this kind of techniques should be developed further and the software should be prepared.

As for media integration, video parser using audio and video is well known, but in this paper, we focused on cross media retrieval, because cross media may be a keyword to information retrieval due to media digitization. Some experimental results showed practical levels, but some are not yet satisfying to be used at present, so that improving the recognition and retrieval technologies will be required further.

8. REFERENCES

[1] http://www.informedia.cs.cmu.edu/dli2/

[2] Dharanipragada, S., Franz, M., McCarley, J.S., Roukos, S. and Ward, T., "Story Segmentation and Topic Detection for Recognized Speech," *Proc. of Eurospeech99*, pp.2435-2438, 1999.

[3] Walls, F., Jin, H., Sista, S. and Schwartz, R., "Topic Detection in Broadcast News," *Proc. of Eurospeech99*, pp.2451-2454, 1999.

[4] Ariki, Y., Ogata, J. and Nishida, M., "News Dictation and Article Classification Using Automatically Extracted Announcer Utterance," *AMCP98*, pp. 78-89, 1998.

[5] Siegler, M.A., "Experiments in Spoken Document Retrieval at CMU," TREC7, 1998.

[6] Gey, F.C., Jiang, H., Chen, A. and Larson, R.R., "Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7," *Proc. of TREC-7*, pp527-540, 1997.

[7] Roy, D. and Malamud, C., "Speaker identification based test to audio alignment for an audio retrieval system," *Proc. of ICASSP97*, pp.1099-1103, 1997.

[8] Delacourt, P., Kryze, D. and Wellekens, J.C, "Detection of Speaker Changes in an Audio Document," *Proc. Eurospeech'99*, pp.1195-1198, 1999.

[9] Linde, Y., Buzo, A. and Gray, R.M., "An algorithm for vector quantizer design," *IEEE Trans. Commun., COM-28*, pp.84-95, 1980.

[10] Kiyama, J., Itoh, Y. and Oka, R., "Automatic Detection of Topic Boundary and Keywords in Arbitrary Speech Using Incremental Reference Interval-free Continuous DP," *Proc. of ICSLP96*, pp. 1946-1949, 1996.

[11] Ariki, Y. and Saito, Y., "Extraction of TV News Articles based on Scene Cut Detection Using DCT Clustering," *Proc. of ICIP96*, pp. III847-III850, 1996.

[12] Smith, M.A. and Kanade, T., "Video skimming and characterization through the combination of image and language understanding technique, " *CMU-CS-97-111*, 1997.

[13] Ariki, Y., Matsuura, K. and Takao, S., "Telop and Flip Frame Detection and Character Extraction from TV News Articles," *Proc. of ICDAR99*, pp. 701-704, 1999.

[14] Muller, S., Eickeler, S., Rigoll, G., "Pseudo 3-D HMMs for Image Sequence Recognition," *Proc. of ICIP99*, 28AP1.11, 1999.

[15] Kurokawa, M., Echigo, T., Tomita,A., Maeda, J., Miyamori, H. and Iisaku, S., "Representation and Retrieval of Video Scene by Using Object Actions and Their Spatio-Temporal Relationships," *Proc. of ICIP99*, 26AO2.1, 1999.

[16] Yaginuma, Y. and Sakauchi, Y., "Content-based Retrieval and Decomposition of TV Drama based on Inter-media Synchronization," *First Int. Conf. on Visual Information Systems*, 1996.

[17] Endo, T., Zhang, J., Nakazawa, M. and Oka, R., "Mutual Spotting Retrieval between Speech and Video Image Using Self-organized Network Databases," *Proc. of AMCP98*, pp. 78-89, 1998.

[18] Satoh, S., Nakamura, Y. and Kanade, T., "Name-It: Naming and Detecting Faces in News Videos," *IEEE Multimedia*, Vol.6, No.1, pp.22-35, 1999.