A Multilingual Video Transcriptor and Annotation-based Video Transcoding

Shigeki Ohira [1], Mitsuhiro Yoneoka [2], and Katashi Nagao [3]

[1] Waseda University
ohira@shirai.info.waseda.ac.jp
[2] Tokyo Institute of Technology
yoneoka@img.cs.titech.ac.jp
[3] IBM Tokyo Research Laboratory
knagao@jp.ibm.com

Abstract. This paper proposes a tool for multimedia annotation and its application. The annotation tool allows users to easily create annotation data including video transcripts, video scene descriptions, and visual/auditory object descriptions. The video transcriptor is capable of multilingual speech identification and recognition. The annotation data enables users to retrieve and transform multimedia content according to their preferences. A video scene description consists of semi-automatically detected key frames of each scene in a video clip and their time codes. A visual object description is created by automatic tracking and interactive naming of people and objects in video frames. An auditory object is also detected semi-automatically. The annotation data is described using XML (Extensible Markup Language). The annotation-based content transformation is called "semantic transcoding" because we deal with semantic features of content. This paper also introduces some examples of annotation-based video transcoding such as video summarization, video-to-document transformation, and video translation.

1 Introduction

The age of digital broadcasting has now arrived. The digitization of broadcasting makes it possible to provide viewers with a wider choice of channels, which leads to the fact that TV program producers must provide indexes/information of their programs, so that viewers can have a better selection.

In the near future, a TV program will be stored mainly in a storage device such as a HDD. Viewers will need some index data with the program for searching interesting scenes speedily, and creating summarized versions the program.

On the other hand, the Internet has an advantage of two-way communication, that is, anyone can put/get information. However, the information has various types or formats and scattered all over the world, we need a search engine to obtain necessary information quickly from the Internet. Since the information include multimedia content, it is difficult to retrieve them using normal search text-based search engines.

As for the digitization of the video/audio signals, standardization has been driven by MPEG (Moving Picture Experts Group), and MPEG-7 (MPEG Phase 7)[1] will be able to add extra information about the content as an index, a note, and so forth. The information is useful for quick search for the preferred content.

However, it takes a high cost to add these indexes and notes by hand. Therefore, automatic method of extraction the data becomes more important. This paper proposes a tool for multimedia annotation and an application based on the annotation.

2 Multilingual Video Transcriptor

2.1 Necessity for semi-automatic multilingual video analysis

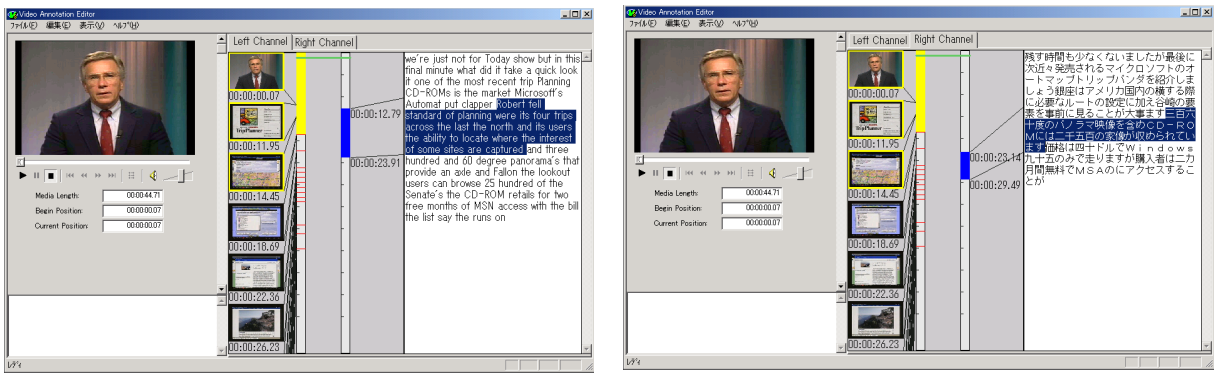
There are many kinds of data types or formats on the Internet. The representative one is text data. Various natural language processing technologies such as retrieval technique, which derives information efficiently from a large amount of data, have been studied and applied. Moreover, in recent years, multimedia data such as video/audio has been dealt with widely, and there is a great demand for information extraction and, above all, analysis of video.

We can obtain the scene information from video data, and the language information from the audio. It is, therefore, important to analyze both audio and video data, and control them using time codes for

an efficient reuse of data. To do all the process by hand is not realistic, because it takes great costs and time. So the semi-automatic system is longed for.

Therefore, we developed a multilingual video transcriptor to extract transcripts, time codes, and scene/object information from multilingual video clips, and semi-automatically provide reusable data structure.

An example of the multilingual video transcriptor is shown in Figure 1.



Left Channel

Right Channel

Figure 1. Multilingual Video Transcriptor

Today because of the widening of the Internet, one can easily get information of the world, and the technology to break the language barrier has become indispensable. The progress of speech recognition technology has made it comparatively easy to transform speech into text. Spoken language identification is needed for processing multilingual speech, because speech recognition technology system knows which language is spoken.

Despite many studies[2–4] on the multilingual speech identification so far, there are few of them with practical applications in sight. The phone speech translation system is one of few representative examples. However, in a usual case, the language information is self-evident (at least, a speaker himself knows), so there is no need to develop a multilingual speech identification system.

On the other hand, speech data in video does not always have definite information about the language. According to the progress of the digital broadcasting and the signal compression technology, it is possible that, in the near future, information about a spoken language will be added.

Most of the data that we can get currently does not have the information, so a large amount of labor is needed to analyze the data. Therefore, the multilingual speech identification has an important role for a speech input of an unknown language.

2.2 Multilingual Speech Identification and Recognition

A multilingual speech identification process is shown in Figure 2. This method determines the input language using a simple discriminate function based on relative scores obtained from parallel processing of speech recognition.

First, the speech recognition engine of each language is connected in parallel to the input speech. It is assumed that each speech recognition engine has the speaker independent model, and each outputted recognition word has the score of the constant range.

When the systems gets speech input, each language type of recognition engine outputs recognition word series and scores. Discriminate unit calculates a discriminate function using outputted scores in every language.

The engine, which has the highest average discriminate value, is judged to be the language of the input speech, and the recognition result is accepted as transcripts.

If there is no distinct difference between discriminate values, that is when all values are not more than a threshold, a judgment is entrusted to the user.

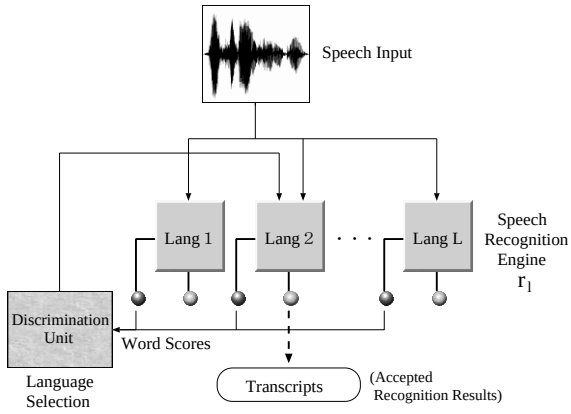


Figure 2. Spoken Language Identification Unit

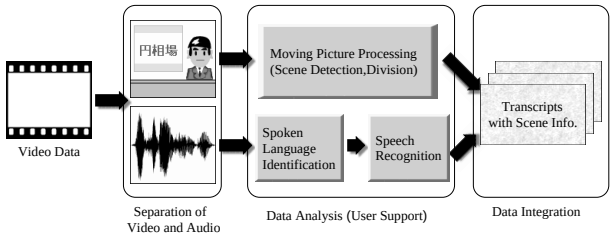


Figure 3. Multilingual Video Data Analysis

The main aim of this method is not language identification itself, but efficient creation of transcripts from large-scale video data. Therefore, it has an advantage that outputted transcripts included word scores are only needed. That is, this method makes it possible to use the existing speech recognition engine, which is tuned in each language, without preparing for a special model for language identification and processing acoustic features. It is also easily to add another language. And, high performance and precision of identification is not required because users annotate for the last time.

Adding to the multilingual speech identification and recognition processing to image processing shown above, users can analyze multilingual video data semi-automatically (shown in Figure 3).

This system drastically reduces the burden of a user who wants video analysis and control, and it will be possible that the conventional natural language processing technique is applied to the transcripts.

Definition of Discriminant Function

The definition of our simple discriminant function is the following. It is assumed that each recognition word obtained from each language type of speech recognition engine r_l ($l : Lang1, ..., LangL$) is shown in Figure 2 has the score of the range to $S_{max}(r_l)$ from $S_{min}(r_l)$.

And $f_I(r_l, i)$, the discriminant function of each engine r_l, when the number of total sentences recognized by each engine is i, is defined with the following formula.

$$f_I(r_l, i) = \log \frac{P_{min}(r_l, i)}{P_{max}(r_l, i) \cdot P_{mean}(r_l, i)}$$

$P_{max}(r_l, i)$: Distance likelihood of cumulative average of highest score and maximum score
$P_{min}(r_l, i)$: Distance likelihood of cumulative average of lowest score and minimum score
$P_{mean}(r_l, i)$: Distance likelihood of cumulative average of average score and maximum score

$$P_{max}(r_l, i) = \frac{S_{max}(r_l) - w_{max}(i)}{S_{max}(r_l) - S_{min}(r_l)}, \quad P_{min}(r_l, i) = \frac{w_{min}(i) - S_{min}(r_l)}{S_{max}(r_l) - S_{min}(r_l)}, \quad P_{mean}(r_l, i) = \frac{S_{max}(r_l) - w_{mean}(i)}{S_{max}(r_l) - S_{min}(r_l)}$$

Now $S_{max}(r_l)$ and $S_{min}(r_l)$ are maximum/minimum score defined in speech recognition engine r_l. $w_{max}(i)$, $w_{min}(i)$, $w_{mean}(i)$ are the average of the highest/lowest/average score in recognized word series at the time i st sentence is recognized, and they are expressed with the following formula.

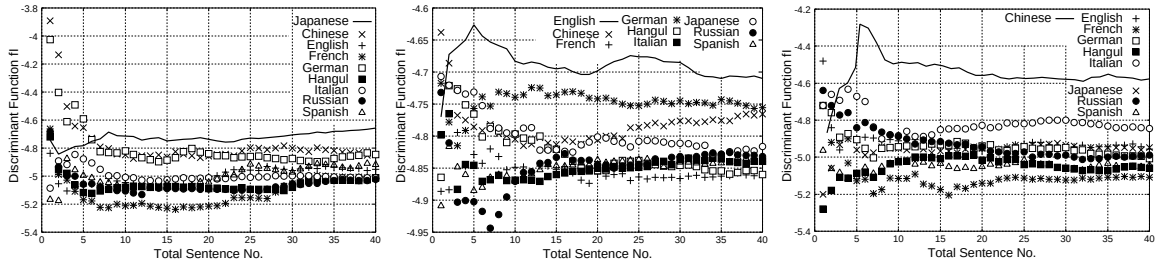
$$w_{max}(i) = \frac{1}{i} \sum_i \max w(i), \quad w_{min}(i) = \frac{1}{i} \sum_i \min w(i), \quad w_{mean}(i) = \frac{1}{i} \sum_i \text{mean } w(i)$$

We use the following formula obtained by expansion of $f_I(r_l, i)$ as a discriminant function. If all recognition engine have the same range of score, you may remove the last constant term.

$$f_I(r_l, i) = \log\{w_{min}(i) - S_{min}(r_l)\} - \log\{S_{max}(r_l) - w_{max}(i)\} - \log\{S_{max}(r_l) - w_{mean}(i)\} + \log\{S_{max}(r_l) - S_{min}(r_l)\}$$

2.3 Experiments of Spoken Language Identification

The result of experiments performed to examine an effectiveness of above discriminant function is the following. We have only Japanese, English, and Chinese speech recognition engine. So, we plotted each language discriminant value on Figure 4, when nine spoken languages are recognized using each speech recognition engine.



Japanese engine English engine Chinese engine

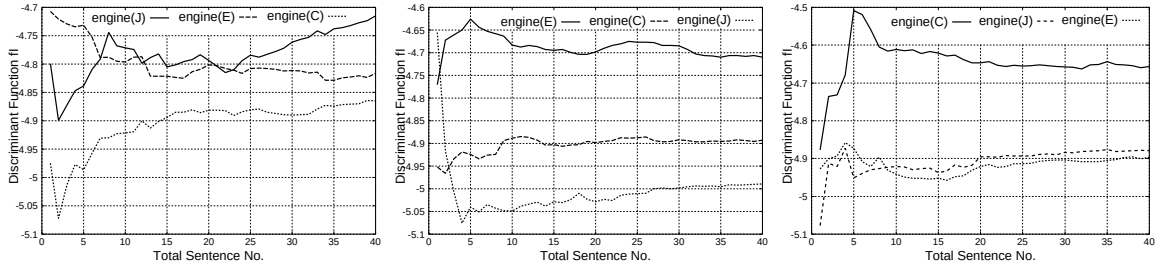
Figure 4. Discriminant values in nine foreign Speech data using Japanese, English and Chinese engine

Therefore, it is showed that an each recognition engine, which is tuned for its language, more prperly discriminate among languages. In next section, we actually apply the discriminant function to these three spoken language.

2.3.1 Trilingual Speech Identification with Colloquial Data

In this experiment, three language speech data (Japanese, English, and Chinese) are used. They are extracted from an audio CD supplied as a foreign language learning tool (40 sentences × 4 sets, male and female speaker two set each). Figure 5 shows the worst results in four test sets.

The solid line in each figure illustrates the result by the same language of the speech recognition engine as input speech.



Japanese speech English speech Chinese speech

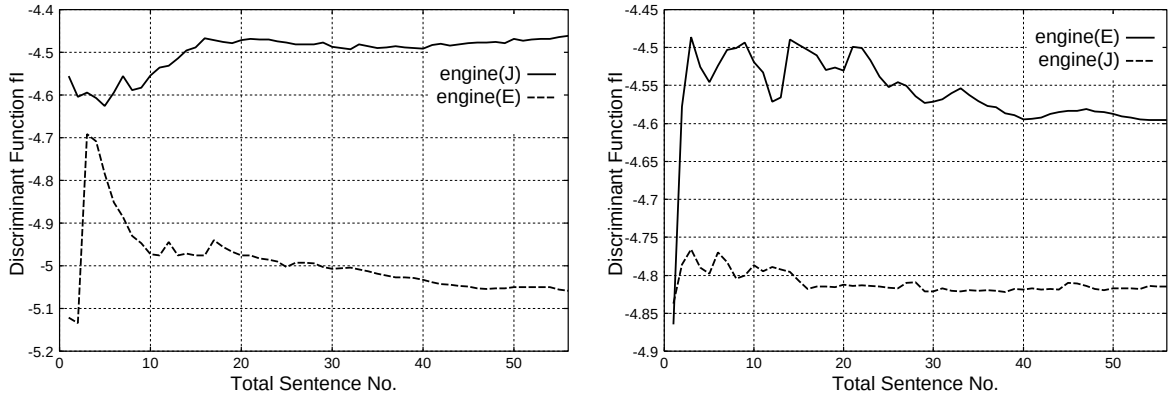
Figure 5. Results of Language Identification with Japanese, English and Chinese speech data

As for Japanese input speech, about twenty sentences are necessary for correct identification with English speech. We think that this cause is in the length of an input sentence. Because sentences used in this experiment are prepared for the practice of the daily conversation, and the length per sentence is very short (the shortest sentence is one word), and our discriminant function which uses the highest/lowest score comes under the great influence of the wrong recognition.

2.3.2 Bilingual Speech Identification with Broadcast News Data

In this experiment, the broadcast news speech data recorded from TV shows are used. This experiment isn't sufficient for evaluation of the validity because the amount of data is small, so the result is shown as a reference data in Figure 6.

In the case of these data, there is much number of words per sentence in comparison with the above data for language-leaning (about more than ten words), therefore the correct identification is obtained in several sentences after recognition started.



Japanese speech English speech

Figure 6. Results of Language Identification with Japanese and English Broadcast News speech data

As mentioned above, it is clear that a recognition word score is one of the rough and ready discirimations in language identification. We think that whether an appropriate recognition engine is used is more exactly determined by calculating n-gram probability and measuring co-occurrence between words against recognized word sequences.

3 Video Annotation

Video data is becoming a prevalent information source. These days, one can store TV programs in a storage device such as HDD and directly record them as digital data with video camera. One can also create and obtain a digitized image easily. In this situation, new services of video distribution are appearing.

Since the size of these collections is growing to huge numbers of hours, retrieval and summarization technique are required to effectively browse video segments in a short time without losing significant content.

However, content-based analysis and processing is more difficult for multimedia content including video and audio than text. Therefore, our aim is to retrieve and summarize the data based on annotations added to them.

The annotation data is described as XML[5] formatted data, and provides multimedia content with a deep semantic structure. In this paper, multimedia content descriptions include scenes and transcripts and objects in a frame, and associate the data with each other. We think that the annotation is a method for understanding semantic content, which makes possible uses of various contents.

The foregoing multilingual video transcriptor makes it possible to efficiently create video annotations described in the following.

3.1 Scene Detection and Object Tracking

As an annotation, except for transcript, by detecting significant changes in the color histogram of successive frames, frame sequences can be separated into key frames. These scene descriptions consist of time codes of the start and the end frames and a scene title.

Additionally, if the user specifies a particular object in a frame by dragging a rectangle, automatic object tracking is executed and time codes and motion trails in a frame (series of coordinates for interpolation of object movement) are checked out.

3.2 Annotation Data

First of all, system separates video data into audio and image. As for the audio data, it is separated into left-right channels. If an average for the difference of the audio signals of both channels is more than a threshold, they are considered to be different data and put on the multilingual speech identification and recognition process.

The obtained transcripts include time codes, recognized word sequences and language information. They are expressed as XML formatted data shown in the Table 1.

Table 1. Transcript Data (.vt file)

```
<?xml version="1.0" encoding="Shift_JIS"?>
<text lang="ja">
<w in="1.264000" out="1.663000">残す</w><w in="1.663000" out="2.072000">時間も</w>
<w in="2.072000" out="2.611000">少なくない</w><w in="2.611000" out="3.180000">ましたが</w>
<w in="3.180000" out="3.778000">最後に</w><w in="3.778000" out="4.856000">次近々発売</w>
<w in="4.856000" out="5.215000">される</w><w in="5.215000" out="5.934000">○○○○○○</w>
<w in="5.934000" out="6.153000">の</w><w in="6.153000" out="6.462000">オート</w>
<w in="6.462000" out="6.802000">マップ</w><w in="6.802000" out="7.191000">トリップ</w>
<w in="7.191000" out="8.039000">パンダを紹介</w><w in="8.039000" out="8.538000">しましょう</w>
...
</text>
```

On the other hand, the dynamic image is described as the data including scene/object descriptions in XML formatted data shown in Table 2, 3.

Scene descriptions have time codes and key frames obtained by the scene detection and the object tracking, and Object descriptions have a name, a related URL, time codes and motion trails in the frame.

Table 2. Scene Information Data (.vs file)

```
<?xml version="1.0" encoding="UTF-8"?>
<scene>
<v in="0.066733" out="11.945279" file="s0.jpg"/>
<v in="11.945279" out="14.447781" file="s1.jpg"/>
<v in="14.447781" out="18.685352" file="s2.jpg"/>
...
</scene>
```

Table 3. Object Information Data (.vo file)

```
<?xml version="1.0" encoding="Shift_JIS"?>
<object>
<vobj begin="1.668335" end="4.671338" name="Davids"
   desc="anchor" img="o0000.jpg" link="http://...">
  <area time="1.668335" top="82" left="34"
        width="156" height="145"/>
  <area ...
</vobj>
...
</object>
```

In the case of video data like a bilingual broadcast, the time code of transcript and the corresponding time code of the scene/object information are often different in channels.

In this case, the annotation data is created for both channels, and video annotation data which integrates each data is described (Shown in Table 4).

Table 4 : Video Annotation Data (.vax file)

```
<?xml version="1.0" encoding="UTF-8"?>
<vax file="D:\demo.mpg">
<text channel="0" src="D:\demo\0.vt"/><text channel="1" src="D:\demo\1.vt"/>
<scene channel="0" src="D:\demo\0.vs"/><scene channel="1" src="D:\demo\1.vs"/>
<object channel="0" src="D:\demo\0.vo"/><object channel="1" src="D:\demo\1.vo"/>
</vax>
```

There are other approaches to the video annotation. For example, MPEG-7[1] is making an effort within the Moving Picture Experts Group (MPEG) of ISO/IEC, which is dealing with multimedia content description.

This content description provides the structure which facilitates retrieval and summarization based on metadata as in our annotation.

Our method will be integrated into tools for authoring MPEG-7 data. Since video annotation includes more complex information processing than so-called video editing, the part which a human is concerned with is complicated. However, in proportion to the improvement in performance of the automatic processing, people will consider that the reusable system of video to annotate rather than edit has more advantage.

4 Video Transcoding

One of the advanced uses of multimedia content based on annotations, is a content processing such as a summarization, translation, and document transformation. We call this semantic transcoding [6, 7].

Semantic transcoding is a transcoding based on external annotations, used for contents adaptation according to user preferences. The transcoders here are implemented as an extension to an HTTP proxy server. Such an HTTP proxy is called a transcoding proxy. We describe the video transcoding from now on.

4.1 Video Summarization

There are some previous work on video summarization such as Infomedia[8] and CueVideo[9]. They create a video summary based on automatically extracted features in video such as scene changes, speech, text and human faces in frames, and closed captions. VideoZoom[10] which IBM Watson Research Center developed makes the resolution of the image of the video change dynamically. This is also a kind of transcoding of video content depending on the restriction of network and device.

They can transcode video data without annotations. However, currently, an accuracy of their summarization is not practical, because of the failure of automatic video analysis. Our approach to the video summarization has sufficient quality, if the data has enough semantic annotations.

Since our annotation data is task-independent and versatile, annotations on video are worth creating, if the video has the possibility to be used in different applications such as automatic editing and information extraction.

We have developed a tool to summarize transcripts with annotations and extract the video scene corresponding to the summary. Since a summarized video transcript contains important information, corresponding to video sequences will produce a collection of significant scenes in the video. An example screen shot of our video player is shown in Figure 7.

Between two time bars under the video screen, the upper one is a scene time bar and lower one is an object time bar. Each dark colored part shows the appearance interval of key frames and objects in the scene corresponding to the summary. Index information about the scene/object is indicated in the right window. The checked indexes in the figure are equivalent to the scene contained in the summary.

4.2 Video Translation

Video translation is to translate transcripts in annotations into different languages as the users want, and to show the results synchronized with the video. The other type of video translation is performed in terms of synchronization of video playing and speech synthesis of the translation results. This translation makes another language version of the original video clip. In the case of a bilingual broadcast, because of the fact that generating transcripts in every channel, the user can get parallel translation data. It is also easy to change the language of a speech to play depending on the user profile.

4.3 Video-to-Document Transformation

Video-to-document transformation is another type of video transcoding. If the client device does not have video playing capability, the user cannot access video contents. In this case, the video transcoder creates and shows a document including important image of scenes and texts related to each scene (shown in Figure 8). Also, the resulting document can be summarized by the text transcoder.

Figure 7. Video Player with Summarization Function

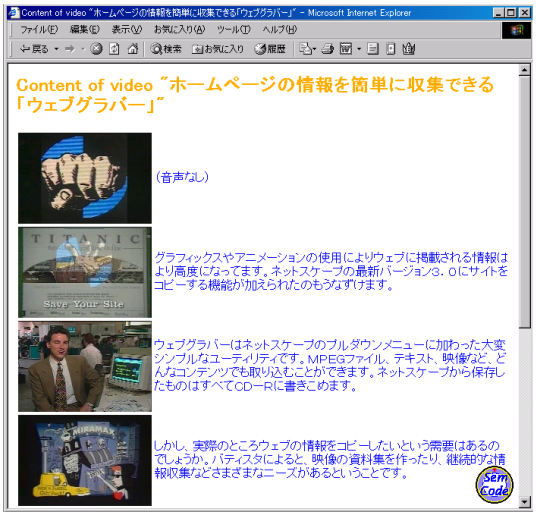


Figure 8. Video Story Document

5 Conclusion and Future Plans

We developed a tool to create multimedia annotation data. The main component of the tool is a multilingual video transcriptor to generate transcripts from multilingual speech of video clips.

The tool also extracts scene and object information semi-automatically, which is described as XML formatted data, and associates the data with the content. Also, we introduced some examples of annotation-based video transcoding as an advanced application of multimedia.

We have implemented video summarization and video-to-document transformation systems. Also, more efficient and content-based retrieval of multimedia content by queries in spoken and written natural language is being pursued. For the last three years, the retrieval of spoken document has also been dealt with in a sub task "SDR (Spoken Document Retrieval) track [1]" at TREC (Text REtrieval Conference)[11].

S. E. Johnson etc.[12] suggested that new challenges such as application of non-lexical information derived directly from the audio and its integration with video data would contribute to the significant improvement of retrieval performance. Therefore, we think that our research has huge impacts and potentials.

We are also planning to modify transcripts using the knowledge on the Web, expand scene/object information and investigate the efficient use of time, place, related URL, and so on.

References

1. Moving Picture Experts Group (MPEG): MPEG-7 Context and Objectives. http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm.
2. T. Arai: Automatic Language Identification Using Sequential Information of Phonemes, Trans IEICE Japan, Vol.E78-D, No.6, pp.705-711, 1995.
3. Hingkeung Kwan and Keikichi Hirose: Recognized phoneme-based n-gram modeling in automatic language identification, Proceedings 4th European Conference on Speech Communication and Technology, Madrid, Vol.2, WEpm2C.5, pp.1367-1370, 1995.
4. Takashi Seino and Seiichi Nakagawa: SPOKEN LANGUAGE IDENTIFICATION USING ERGODIC HMM WITH EMPHASIZED STATE TRANSITION, Proceedings of EUROSPEECH'93.3rd, pp. 133-136, 1993.
5. World Wide Web Consortium: eXtensible Markup Language (XML). http://www.w3.org/TR/PR-xml-971208.
6. Katashi Nagao, Yoshinari Shirai, Kevin Squire: Semantic Annotation and Transcoding: Making Web Content More Accesible, *IEEE Multimedia.* Vol. 8, No. 2, pp. 69–81, 2001.
7. Katashi Nagao, et al.: Semantic Transcoding: Making the World Wide Web more understandable and usable with external annotations, *TRL Research Report,* IBM Tokyo Research Laboratory, 2000.
8. Michael A. Smith and Takeo Kanade: Video skimming for quick browsing based on audio and image characterization. *Technical Report CMU-CS-95-186.* School of Computer Science, Carnegie Mellon University, 1995.
9. A. Amir, S. Srinivasan, D. Ponceleon, and D. Petkovic: CueVideo: Automated indexing of video for searching and browsing. In *Proceedings of SIGIR'99.* 1999.
10. John R. Smith: VideoZoom: Spatio-temporal video browser. *IEEE Trans. Multimedia.* Vol. 1, No. 2, pp. 157–171, 1999.
11. Text REtrieval Conference: http://trec.nist.gov/
12. S. E. Johson, et al.: SPOKEN DOCUMENT RETRIEVAL FOR TREC-9 AT CAMBRIDGE UNIVERSITY, Proc. of Text REtrieval Conference(TREC-9), 2001

[1] This track was finished for a while in TREC-9