

## SceneCabinet/Live!: Realtime Generation of Semantic Metadata Combining Media Analysis and Speech Interface Technologies

*Hidetaka Kuwano<sup>†</sup>, Yuko Kon'ya, Tomokazu Yamada,  
and Katsuhiko Kawazoe*

### Abstract

Reducing the cost of generating metadata would allow more broadcast contents to be transmitted with advanced TV viewing services available through the use of metadata. In this article, we describe SceneCabinet/Live!, a system that generates scene-based semantic metadata for video content in real time by combining media analysis and user interface technologies. The system provides an intuitive operation interface that lets the operator easily generate metadata by only speaking about the content while watching the replayed video. No keyboard input is required for this operation. Scene title, synopsis, and keywords can be obtained using natural language processing based on speech recognition results. Our speech recognition method obtains almost errorless results because it uses specific grammatical rules matched to the genre of video content. In an experiment, for a live baseball program broadcast, we confirmed that semantic metadata concerning scenes of home runs and skilful play could be generated in real time without any delay.

### 1. Introduction

Technology for distributing broadcast content and metadata over broadband IP (Internet protocol) networks has recently been studied with the aim of providing a new style of TV viewing that will create a new market for broadcasting and telecommunication services. Metadata is a key technology for implementing scene-based advanced TV viewing services such as scene navigation, digest viewing, and highlight viewing. Information such as what scenes are contained in the program and the times when the scenes appear must be described as metadata before the program is broadcast. For example, it is necessary to generate metadata describing the content of each individual news topic for a news program and information such as the approximate times of exciting action in a sports program. However, the task of generating such “scene-based semantic metadata” requires a large number of operations when it is done

entirely manually. Thus, reducing the operation cost of generating metadata seems to be the key to distributing more contents that allow advanced viewing services.

We previously developed the SceneCabinet system that semi-automatically generates metadata using video analysis, speech recognition, and natural language processing techniques [1]-[7]. SceneCabinet creates a workflow for generating metadata for programs that have already been produced. To extend the range of metadata-based TV viewing services, we developed new technologies and a workflow for generating scene-based semantic metadata in real time while a live program is being broadcast.

Some TV screenshots of a digest viewing service for a live televised baseball game are shown in **Fig. 1**. Digest scene titles are updated and displayed on the TV screen as the game progresses. To replay digest scenes, the viewer simply selects the scene titles that he or she wants to see. This permits viewers to make the most effective use of their TV viewing time because they can easily replay scenes of interest that they missed during the live broadcast whenever it is convenient for them. For example, viewers can easily

<sup>†</sup> NTT Cyber Solutions Laboratories  
Yokosuka-shi, 239-0947 Japan  
E-mail: kuwano.hidetaka@lab.ntt.co.jp

go back and watch scenes from a baseball game in which their favorite players appear. To implement this kind of new TV viewing service, we must provide a way to generate scene-based semantic metadata in real time while the program is being broadcast.

In the following section, we describe SceneCabinet/Live!, a system that can generate scene-based semantic metadata for video content in real time by combining media analysis and user interface technologies. The system provides an intuitive operating interface that lets the operator easily generate metadata simply by speaking about the content while watching the replayed video. No keyboard input is necessary for this operation. Scene title, synopsis, and keywords can be obtained using natural language processing based on speech recognition results. Our speech recognition method obtains almost errorless results because it uses specific grammatical rules matched to the genre of the video content. We also describe an effective way of reducing the cost of gen-

erating metadata based on the results of an experiment.

## 2. Challenge of generating metadata for live broadcast service

Figure 2 shows the metadata items that are need to support the digest viewing service using a live baseball game broadcast as an example: the start and end times of each individual batting scene in the program and the title, synopsis, and keywords for each of these scenes. Note that these items are also defined in the metadata international standards stipulated by the TV-Anytime Forum [8].

To implement the live program digest viewing service shown in Fig. 1, it is necessary to first generate the metadata in Fig. 2 while the program is in progress and deliver it to the viewer’s terminal immediately. In earlier papers [1], [2], we pointed out that if the work of generating metadata is done manually,

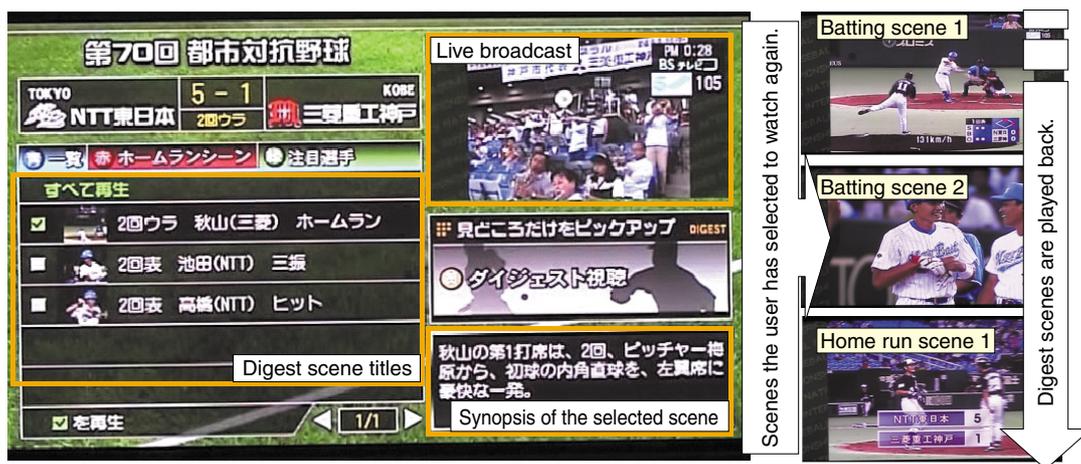


Fig. 1. TV screenshots of metadata-based viewing service.

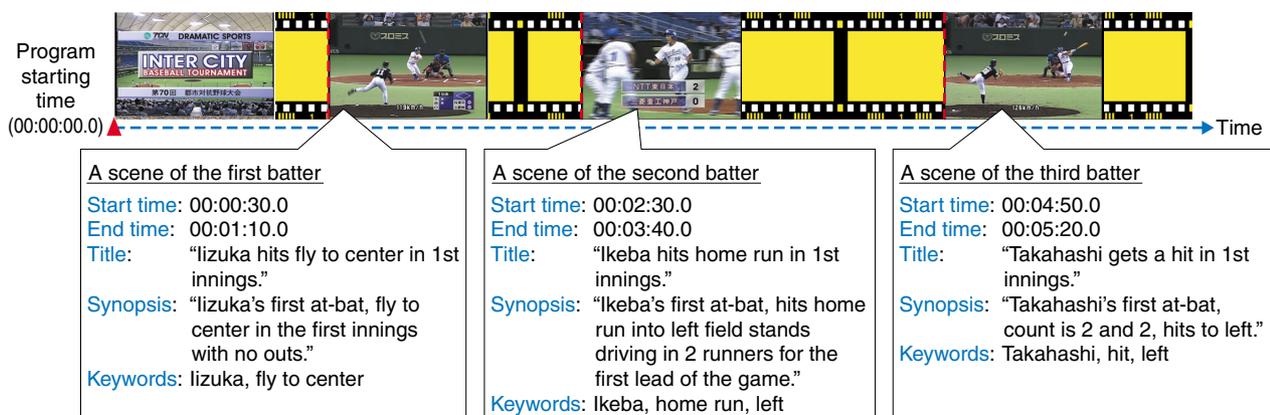


Fig. 2. Example of metadata items needed to support baseball digest viewing service.

it is extremely fatiguing and involves an enormous amount of labor. Although it may be relatively easy and efficient to generate title, synopsis, and keywords for scenes when the program is scripted in advance, an unscripted program such as a live baseball game presents a much more difficult challenge for metadata generation.

For unscripted programs such as live sports broadcasts, metadata such as scene titles can be generated after the end of each scene. Attempting to do this work manually would not only be extremely fatiguing, but would also rule out even the slightest margin of extra time, so the generated metadata would contain many errors. Therefore, it is essential to develop an efficient workflow for generating metadata that can keep up with the live action of the program without exhausting the operator.

### 3. Metadata generation workflow based on speech recognition

Targeting live broadcasts such as baseball games, we defined a workflow that will quickly generate the metadata items listed in Fig. 2 during the game without a significant time delay. We were aware that the most efficient approach would be to take full advantage of media analysis and user interface technologies, reduce manual work to the absolute minimum, and automate the process so that the operator could keep his/her eyes on the game. This led us to define the workflow shown in Fig. 3, which fully exploits the advantages of speech recognition technology. In addition, we developed a realtime metadata generation system “SceneCabinet/Live!” that implements this workflow. An example screenshot of the graphical user interface (GUI) of SceneCabinet/Live! is shown in Fig. 4.

The workflow begins when a home run or other important batting scene ends. The operator orally gives a synopsis of what happened in the scene by speaking in a speech-recognition friendly manner (i.e., enunciating clearly in an environment with no background noise). The synopsis explaining the scene content must be displayed within a limited space on the viewer’s TV screen, so we defined the format taking into account the number of characters. We arranged to have the operators speak according to the fixed format shown in Fig. 5. Speech recognition processing based on a fixed format places restrictive conditions on the recognition results. Applying a fixed format for the scene synopses yields a significantly higher recognition rate than allowing unre-

strained speech, so this is an effective method for generating metadata in a short time. The system works as follows. When the operator speaks into the microphone of the system, the speech recognition results are displayed in the browser shown on the right in Fig. 4. One of the sentences in the speech recognition results browser, corresponding to the batting scene to be given metadata, is selected and copied into the metadata editor shown in the lower left of Fig. 4. Finally, after the work of creating the scene synopsis has finished, the copied sentence can be corrected if necessary.

The scene title and keywords are also input using the metadata editor. The title (which contains the batter’s name, the action occurring in the scene (e.g., a home run or 2nd-base hit), and the innings number) is input using a selection function of the GUI, such as a set of radio buttons. Keywords are generated by pressing the “keyword” button on the metadata editor. This runs a natural language processing routine and keywords are automatically extracted from the title and synopsis already established.

The scene start and end times are set using the speech recognition results browser described earlier, the key image browser shown in the center of Fig. 4, and the playback monitor shown in the upper left of Fig. 4. Representative images on the key image browser are the results of a video recognition process such as scene change detection, superimposed text detection, and camera motion detection. When the

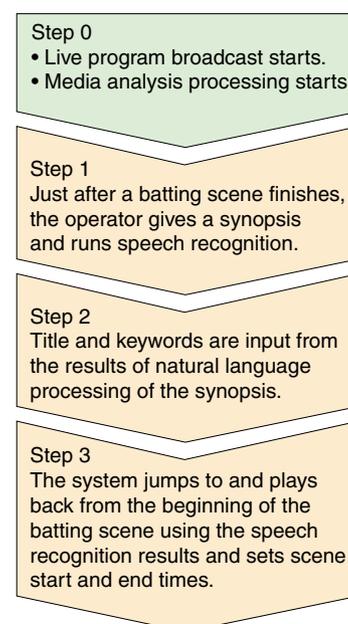


Fig. 3. Metadata generation workflow for live programs.

operator selects the speech recognition results in the speech recognition result browser or representative images in the key image browser, the system displays on the playback monitor the scene that corresponds to the selection. Because the synopsis goes through speech recognition processing right at the end of the scene for which metadata is to be generated, it is easy to jump back to the scene start position on the playback monitor by simply selecting the text in the

speech recognition results browser. Then the operator can set the start and end times of the metadata input scenes using the images in the key image browser, or the fine-control buttons for  $\pm 10$  s or  $\pm 1$  s on the playback monitor.

In the generation of metadata for a live program digest viewing service, the above workflow allows the operator to input titles and other text without touching a keyboard and generate metadata while

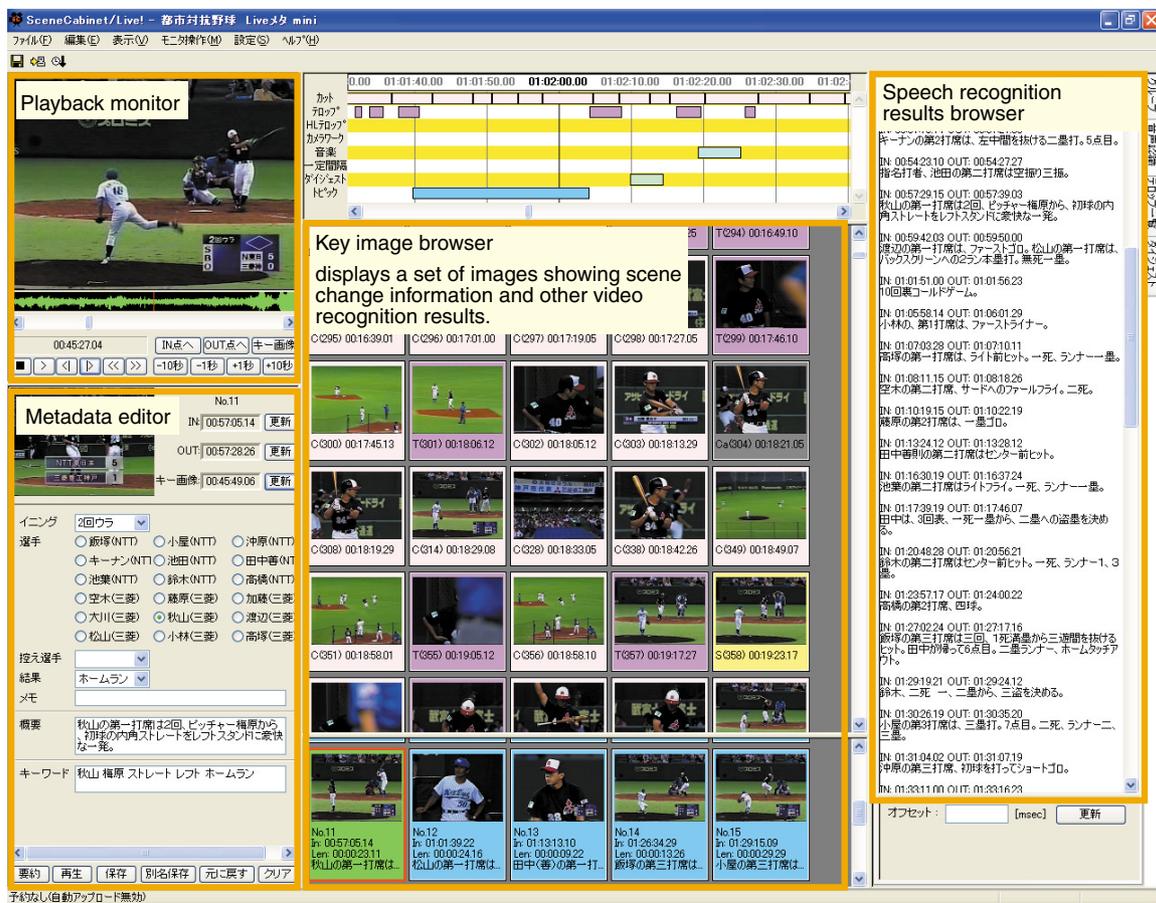


Fig. 4. Screenshot of metadata authoring GUI of SceneCabinet/Live!

**Scene synopsis format**

“(Batter’s name), (At-bat no.), (Innings), (Out count), (Strike-and-ball count), (Pitcher’s name), (Scene action)”

\* Speech rule: State in this order. Not all items are required.

**Scene synopsis examples**

- 1) “Ikeba’s first at-bat, hits home run into left field stands driving in two runners for the first lead of the game.”
- 2) “Takahashi’s first at-bat, count is 2 and 2, gets hit to the hole.”
- 3) “Akiyama’s first at-bat, 2nd innings, first pitch from Umehara, gets an amazing hit to left stands on inside straight.”



Fig. 5. Example of fixed format synopsis for batting scenes.

watching the game simply by speaking synopses about the scenes and navigating an intuitive GUI. The scene start and end times can be set easily using the video analysis processing results.

#### 4. Experimental results and discussion

We conducted a series of experiments to compare the relative efficiency of generating metadata using SceneCabinet/Live! (task model A) versus preparing the same metadata by manual operation of the time-shift replay function while recording a live baseball program with a hard disk recorder (task model B). Six operators tried each task model and all were asked to generate metadata for all the batting scenes during one baseball game broadcast. The detailed rules for generating metadata items for each batting scene were as follows.

- Start time: Time when the pitcher begins to throw the ball to the batter.
- End time: Time when the results of the action (e.g., hit or strikeout) can be distinguished.
- Title: Short sentence containing the batter's name, description of the action in the scene, and innings number.
- Synopsis: Sentence explaining the scene including some of the following items, which must be in this order: batter's name, at-bat number, innings number, out count, strike-and-ball count, pitcher's name, and action occurring in the scene.
- Keywords: Names and important words extracted from the title and synopsis.

We measured how long it took to input the metadata. We defined this time, which we call the task time, as being from the end of the batting scene being broadcast to the completion of the scene end time being input. **Figures 6** and **7** show the results for task models A and B. The horizontal axis of each graph represents the scene number from the beginning of the program, and the vertical axis represents the task time. Figure 6 shows that the task times were almost the same for all operators even if the program progressed, as in task model A. The maximum task time was about ten minutes. On the other hand, in task model B, we noticed a tendency for the task time to increase as the program progressed, as seen in Fig. 7. As a result, we verified that the time taken to perform the task using SceneCabinet/Live! (task model A) was shorter than that using manual operation (task model B). Although a quantitative analysis is difficult because of individual differences in the operators, the average task times for the same scene for models A

and B were comparable. For example, for the 9th scene in task model A, the average task time was 6 minutes and 17 seconds, whereas for the same scene in task model B, it was 17 minutes and 18 seconds. This shows that using SceneCabinet/Live! reduced the task time by about 64% compared with manual operation. Moreover, the task time for operator #5 was especially good: 3 minutes and 10 seconds for the 21st scene in task model A, and 17 minutes and 1 second for the same scene in task model B. In this case, SceneCabinet/Live! reduced the task time by about 81% compared with manual operation. As the operators become more familiar with the task, task model A will become even more efficient compared with task model B.

We also asked the operators to complete a questionnaire to determine how tiring they found the work of generating metadata using the two task models. Those who used only manual procedures felt that the work was “quite tiring” and “really tough”, while the operators who used SceneCabinet/Live! generally responded that the work “wasn't all that hard”. In particular, for creating the synopses, the respondents agreed that speaking about the scene was much easier and more efficient than inputting the information manually. Many of the respondents also commented that the speech recognition results provided good clues that made it easy to find where scenes began when using SceneCabinet/Live!, and this made the task of finding the scene start and end times very easy. All operators also commented that the  $\pm 1$ -s and  $\pm 10$ -s jump buttons on the replay monitor were very effective.

Because comparatively explicit rules for guidance in performing the task content can be specified for baseball games, the operators were able to perform the tasks easily and effectively, even without specialized skills. On the other hand, for tasks involving the selection of important plays rather than individual batting scenes, the operator must have specialized knowledge such as baseball expertise or creative skills such as generating metadata that includes presentation effects in the digest. Even in such cases, however, if the task can be divided into two stages— involving the generation of “basic” metadata for every scene expected to be used in the digest and involving the use of the “basic” metadata by creative personnel to create the final metadata—then the workflow can easily be defined for the former task, and video recognition technology can also be applied effectively. Thus, we can expect that a system such as SceneCabinet/Live! system will lighten the burden on

creative personnel and reduce the overall costs of the task.

We made a prototype system that implements a digest viewing service for a live TV program using the metadata generated by SceneCabinet/Live! and asked general users to complete a questionnaire about the service. We obtained positive opinions that they want to use the service. There was no dissatisfaction with the metadata updating speed. Thus, we confirmed the effectiveness of SceneCabinet/Live!

## 5. Conclusions

This article presented an overview of a digest viewing service and the metadata generation technologies needed to implement such a service, which lets viewers go back and look at missed scenes from live broadcasts at their own convenience. We demonstrated that metadata could be continually generated to keep up with the pace of a live baseball game by having the system automatically generate a synopsis of

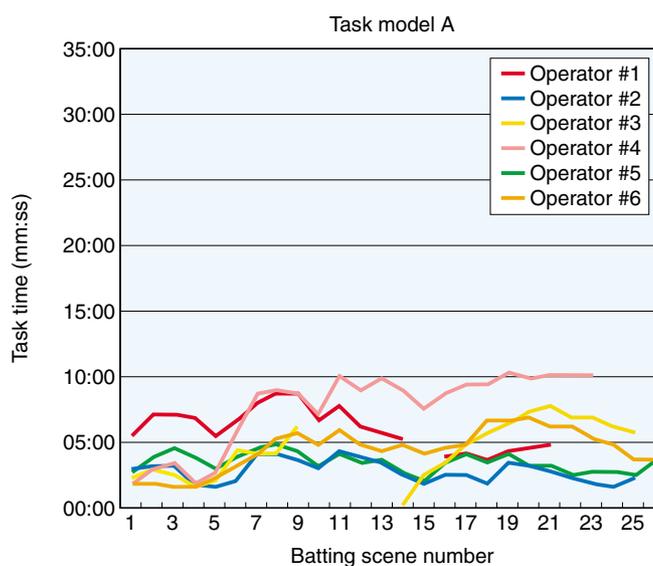


Fig. 6. Task time for task model A.

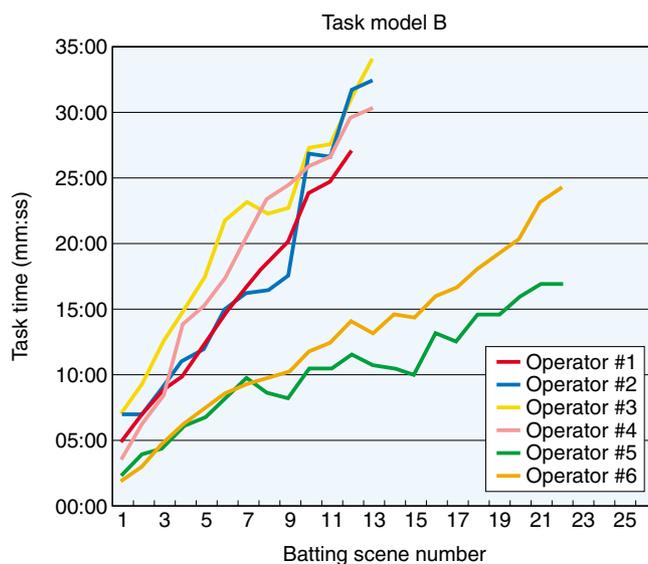


Fig. 7. Task time for task model B.

each individual batting scene “on-the-fly” by applying speech recognition to fixed-format phrases.

In generating metadata for broadcast and communication convergence services, we wish to emphasize the difference between the metadata generation workflow for prerecorded/scripted programs and the workflow described here for generating metadata on-the-fly for live programs. Not only the type of program but also the program genres, such as dramas and variety shows, lead to differences in the metadata generation workflow. Moreover, the metadata generation workflow differs depending on the level of automatic metadata generation technology, such as media analysis and the extent to which program scripts and other reusable information are prepared in advance.

To further improve this approach, we plan to study ways to further automate metadata generation, develop a workflow model that can be used to implement different workflows optimized for different kinds of programs and conditions and find ways to further reduce the total cost of generating metadata including the workflow definition stage.

## References

- [1] H. Kuwano, Y. Matsuo, and K. Kawazoe, “SceneCabinet: Semantic Metadata Extraction System combining Video/Audio Indexing and Natural Language Processing Techniques,” Proceedings of IBC 2004, pp. 458-466, 2004.
- [2] H. Kuwano, Y. Matsuo, and K. Kawazoe, “Reducing the Cost of Metadata Generation by Using Video/Audio Indexing and Natural Language Processing Techniques,” NTT Technical Review, Vol. 2, No. 8, pp. 68-74, 2004.
- [3] Y. Taniguchi, A. Akutsu, and Y. Tonomura, “PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing,” Proceedings of ACM Multimedia 97, pp. 427-436, 1997.
- [4] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, “Video Handling with Music and Speech Detection,” Proceedings of IEEE Multimedia, Vol. 5, No. 5, pp. 17-25, 1998.
- [5] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, “Telop on Demand: Video Structuring and Retrieval based on Text Recognition,” Proceedings of International Conference on Multimedia and Expo 2000, pp. 759-762, 2000.
- [6] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui, “Topic Extraction based on Continuous Speech Recognition in Broadcast-news Speech,” Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 527-534, 1997.
- [7] Y. Hayashi, K. Ohtsuki, K. Bessho, O. Mizuno, Y. Matsuo, S. Matsunaga, M. Hayashi, T. Hasegawa, and N. Ikeda, “Speech-based and Video-supported Indexing of Multimedia Broadcast News,” Proceedings of SIGIR, pp. 441-442, 2003.
- [8] <http://www.tv-anytime.org/>



**Hidetaka Kuwano**

Research Engineer, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in information engineering from Niigata University, Niigata in 1993 and 1995, respectively. In 1995, he joined NTT Human Interface Laboratories, Yokosuka. Since then he has been engaged in R&D of video analysis, video structuring, and video OCR systems. He received the Young Engineer Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2000. He is a member of IEICE of Japan, the Information Processing Society of Japan (IPSI), and the Institute of Image Information and Television Engineers.



**Yuko Kon'ya**

Research Engineer, NTT Cyber Solutions Laboratories.

She received the B.S. degree in mathematics from Ochanomizu University, Tokyo in 1999. She joined NTT Cyber Solutions Laboratories in 2004 and is engaged in R&D of a metadata generation system based on multimedia recognition for broadcasting. She is a member of IEICE.



**Tomokazu Yamada**

Senior Research Engineer, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Science University of Tokyo, Tokyo in 1987 and 1989, respectively. Since joining NTT in 1989, he has been investigating speech recognition. He also engaged in business development concerning contents distribution services. He is a member of the Acoustical Society of Japan and IPSI.



**Katsuhiko Kawazoe**

Senior Research Engineer, Supervisor, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in engineering from Waseda University, Tokyo in 1985 and 1987, respectively. Since joining NTT in 1987, he has mainly been engaged in R&D of radio communication systems, satellite communication systems, and the personal handy-phone system (PHS). His specialty is forward error correction systems. He is currently a co-chairman of the Association of Radio Industries and Businesses Working Group for Broadcasting Systems based on a Home Server. He is a member of IEICE and received the Young Engineer Award from IEICE in 1995.