Automatic metadata generation and video editing based on speech and image recognition for medical education contents

Satoshi Tamura[1], Koji Hashimoto[1], Jiong Zhu[1], Satoru Hayamizu[1],
Hirotsugu Asai[3], Hideki Tanahashi[3], Makoto Kanagawa[2]

[1] Gifu University [2] Sanyo Electric Co.,Ltd.

[3] Gifu Prefectural Research Institute of Manufacturing Information Technology

E-mail: {tamura@info, koji@hym.info, sho@hym.info, hayamizu@cc}.gifu-u.ac.jp,
{asai,tana}@gifu-irtc.go.jp, makoto.kanagawa@sanyo.co.jp

10.21437/Interspeech.2006-618

Abstract

This paper reports a metadata generation system as well as an automatic video edit system. The metadata are information described about the other data. In the audio metadata generation system, speech recognition using general language model (LM) and specialized LM is performed to input speech in order to obtain segment (event group) and audio metadata (event information) respectively. In the video edit system, visual metadata obtained by image recognition and audio metadata are combined into audio-visual metadata. Subsequently, multiple videos are edited to one video using the audio-visual metadata. Experiments were conducted to evaluate event detection of the systems using medical education contents, ACLS and BLS. The audio metadata system achieved about a 78% event detection correctness. In the edit system, an 87% event correctness was obtained by audio-visual metadata, and the survey proved that the edited video is appropriate and useful.
Index Terms: metadata, speech recognition, audio-visual integration, automatic video edit.

1. Introduction

Recently, a lot of multimedia contents such as movies, speech and musics, have come to be distributed on the Internet. As this vast sea of information grows larger and larger, the "metadata", which are information about these contents, to find and utilize the contents more effectively, become essential. The demand to give metadata is large, however, most metadata are given manually with high cost. It is now necessary to develop a technology that automatically provides accurate and efficient metadata.

There are some researches regarding metadata generation and its related works in many fields and areas: the metadata generation system for football TV program [1], the highlight scene extraction method for baseball broadcast video [2], and the retrieval system of e-learning (online university lecture) contents using speech recognition [3]. The football metadata generation system [1] is aimed at broadcast archive retrieval. The system uses crowd noise in a stadium to detect goal and shot scenes, and announcer's voice to identify semantics of the scenes by speech recognition. The baseball metadata generation system [2] detects highlight scenes using many kinds of features obtained from image sequences and hidden-Markov-model(HMM)-based methods. These researches use a single channel resource, on the other hand, there are few researches using multi resources such as audio and visual signals. It is obvious that the correctness of metadata can be improved by using multimodal information. Furthermore, except for traditional information retrieval, few researches utilize created metadata for advanced systems and purposes. Thus it is a important issue how to use the obtained metadata effectively.

Focusing on medical education such as emergency medical training, this paper proposes an automatic metadata generation system, as well as a video edit system using created metadata. In our system, speech recognition and image pattern recognition methods are used to obtain metadata including event names and time periods. Using the metadata, an original content is edited in order to provide more effective education.

This paper is organized as follows: in Section 2, the metadata generation system using speech processing is introduced and evaluated. The video editing system including metadata generation is constructed in Section 3. Experiments and survey of the edit system are also described in Section 3. Finally, Section 4 concludes this paper.

2. Audio Metadata Generation System

2.1. ACLS

The advanced cardiovascular life support (ACLS) is widely used in hospitals for education of emergency treatment. The ACLS is a series of life saving methods recommended by the American heart association (AHA), as well as is one of life-saving treatment means based on international guidelines. In this chapter, the ACLS is used in order to evaluate the metadata generation system.

2.2. System architecture

Figure 1 shows the summary of our metadata generation system:

1. Extract an audio channel (a WAV file) from movie data (an AVI file)
2. Find speech data from the audio channel
3. Get defibrillation (DF) time periods by detecting DF alarms
4. Recognize the speech using general language model
5. Determine the kinds of segment according to the result of speech recognition and DF time periods
6. Apply the proper language model for each segment to secondarily perform speech recognition
7. Determine the right event based on the recognition result
8. Create metadata to result data (an XML file)

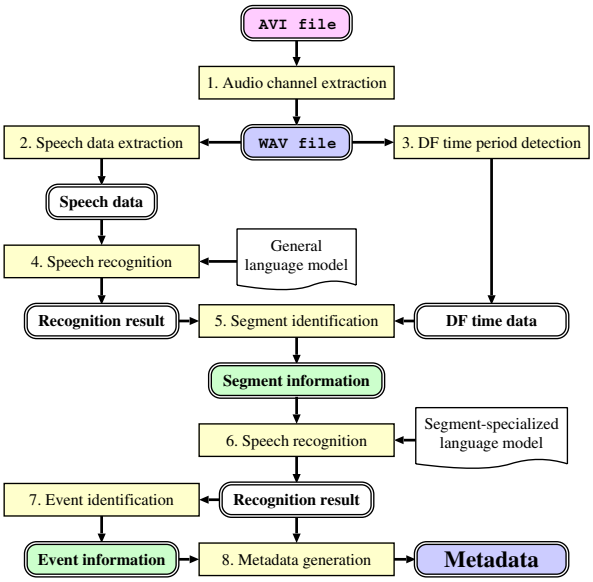

Figure 1: Principle of proposed metadata generation system using speech signal processing.

Table 1: Recognition results [%] of ACLS scenarios.

	Corr.	Acc.	Del.	Sub.	Ins.
Not segmented	49.58	31.73	11.68	38.73	17.85
Segmented	56.17	38.54	10.83	33.01	17.63

In the ACLS there are 24 events, such as defibrillation, cardiac massage, drug administration, etc. Every events are categorized into 4 segments: primary ACLS (PA, 10 events), secondary ACLS (SA, 9 events), DF confirmation (DC), and testing (TS, 4 events). The DC segment follows the PA or the SA.

2.3. Experiment of speech recognition

We conducted an experiment to evaluate effect of the segmentation regarding speech recognition accuracy. We used 368 Japanese utterances spoken by two females and two males. In speech recognition, we used the decoder "Julius" [4] as well as speaker- and gender-independent triphone HMMs as acoustic model. For evaluation, the following measures were used:

$$\text{Corr.} = \frac{C}{C+D+S} \quad , \quad \text{Acc.} = \frac{C-I}{C+D+S} \tag{1}$$

$$\text{Del.} = \frac{D}{C+D+S} \quad , \quad \text{Sub.} = \frac{S}{C+D+S} \qquad \text{Ins.} = \frac{I}{C+D+S} \tag{2}$$

where C is the number of words which were recognized correctly, D is the number of deleted words, S is the number of substituted words, and I is the number of inserted words.

Table 1 indicates the average of recognition results with performing segmentation and without segmentation. This result shows that the recognition accuracy can be improved by performing segmentation and selection of the appropriate language model. This improvement is mainly due to the reduction of substitution errors; the correct words were recognized by using the right language model in which the word distribution for each scene was well modelled. Each recognition correctness rate of four speakers after segmentation was 41-70%. To further improve recognition results, adaptation of acoustic model for each speaker is our future work.

Table 2: Event detection rates [%] of ACLS scenarios.

	Corr.	Acc.	Del.	Sub.	Ins.
Not segmented	73.39	59.30	19.37	7.24	16.05
Segmented	78.28	66.54	14.87	6.85	13.70

2.4. Experiment of event detection

An another experiment was conducted to evaluate the proposed system regarding event detection, or metadata generation. We used 10 ACLS scenarios demonstrated by three females and seven males. There were 511 events in all scenarios. In this experiment, we evaluated event detection rates as well as the speech recognition experiment: after every segment in each scenario was identified, metadata consisting of event name and its beginning and ending time were created. We use the same equations (1) and (2) to estimate the performance.

Table 2 shows the average of event detection rates. Both the correctness and the accuracy with segmentation were higher than those without segmentation (roughly 18% error rate reductions). It is observed that event deletion error and event insertion error greatly decreased, and contributed this improvement.

2.5. Discussion

Shown in Table 1 and Table 2, by performing segmentation and applying right language model to each segment, significant improvements were achieved. According to Table 1 and Table 2, there was a significant decrease of word substitution error in speech recognition, whereas in event detection the event deletion error rate and the event insertion error rate were greatly reduced. The reasons for these error reduction are described as follows: in speech recognition, if a non-keyword term were wrongly recognized as a keyword, it would be a word substitution error, whereas it would be an event insertion error instead in event detection. On the other hand, if a keyword were wrongly recognized as a non-keyword term, it would be an event deletion error in event detection.

3. Video Edit System

3.1. System outline

Figure 2 illustrates the summary of the proposed video edit system, which firstly creates metadata from speech and image recognition, secondly edits multiple video data using those metadata, and finally makes one video file.

Our system consists of two subsystems: a metadata subsystem and a video edit subsystem. The metadata generation subsystem further comprises three modules: an image recognition module, a speech recognition module, and an audio-visual integration module. The former two modules generate metadata from speech and image recognition results respectively. These metadata are integrated to create final metadata in the audio-visual integration module. In the edit subsystem, video edit is performed by selecting the appropriate camera view using metadata information.

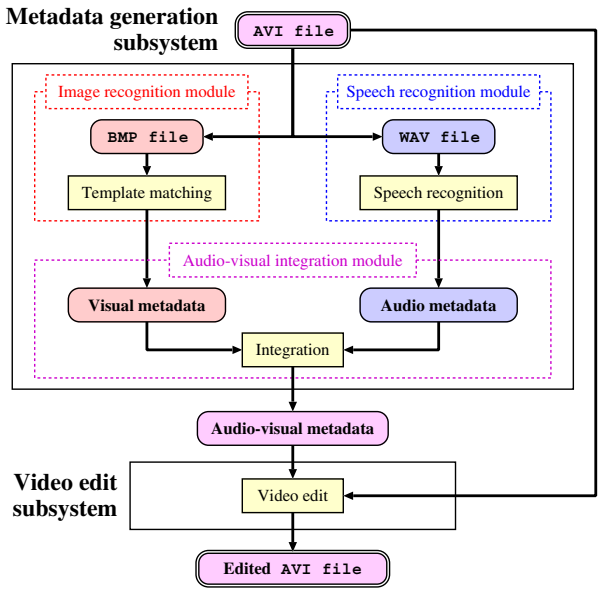

Figure 2: Principle of proposed video edit system.

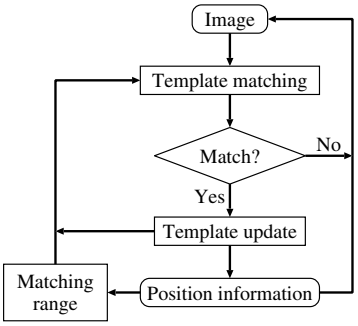

Figure 3: Flowchart of template matching method.

3.2. Metadata generation subsystem

3.2.1. Image recognition module

Head and hand areas in an image are detected by template matching using the RGB Euclidean distance shown in (3):

$$\sum_k \sqrt{\{(R_k^P - R_k^I)^2 + (G_k^P - G_k^I)^2 + (B_k^P - B_k^I)^2\}} \quad (3)$$

where R_k^P is a red value of the k-th pixel on the template image, G_k^I is a green value of the k-th pixel on the input image region. At the beginning, initial head and hand template images are prepared. Figure 3 illustrates the template matching method where the template is updated. This method has the advantage that accommodate the rotation of the matching targets as well as a certain level of zooming. In addition, the computational effort can be greatly reduced by limiting search ranges.

From the relative locations of heads and hands obtained by image recognition, the kind of event can be determined. Event information (visual metadata) including time and event name are then created.

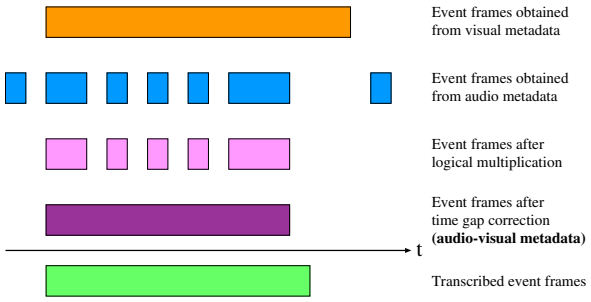

Figure 4: Event frames obtained from audio, visual, audio-visual metadata and transcription respectively.

```
0m28s· · · Start: Event [1] confirmation of consciousness
0m37s· · · End:  Event [1] confirmation of consciousness
0m46s· · · Start: Event [6] artificial respiration
0m53s· · · End:  Event [6] artificial respiration
```

Figure 5: The example of audio-visual metadata.

3.2.2. Speech recognition module

The speech recognition module is almost the same introduced in Section 2. Time information of recognized keywords are also obtained from recognition results. As a result, audio metadata including time, keyword, and event name are generated.

3.2.3. Audio-visual integration module

Audio-visual integration methods are divided into two groups: early integration and late integration [5]. We adopted a late-integration-based method since the late integration can compensate an error obtained from one resource using correct information obtained from the another.

Visual metadata have time continuity, on the other hand, the ending time of event is often incorrect since it is difficult to determine it by using visual information only. Keyword detection made by speech recognition has little time continuity, however, time information of audio metadata are more accurate than those of visual metadata. This audio-visual integration module combines audio and visual metadata by applying logical multiplication and time gap filling (see Figure 4). The example of obtained audio-visual metadata is shown in Figure 5.

3.3. Video edit subsystem

Multiple camera perspectives are assumed in our system. Beforehand, for each event, it is determined manually and statically which camera view should be used. In this subsystem, timing of camera switching is obtained according to audio-visual metadata information, and is stored to an XTL file. Video edit is subsequently conducted using the XTL file and the application: DirectShow Edit Service "xtlTest" [6].

3.4. Experiment

Video contents of basic life support (BLS) were used in this experiment. The BLS is one of the cardiopulmonary resuscitations for medical nonprofessionals, whereas the ACLS is for specialists such as rescue life guards and doctors. There were six events in the BLS: confirmation of consciousness (CC), artificial respiration

Table 3: Event detection correctness [%] for audio(A), visual(V) and audio-visual(A-V) metadata.

	CC	AR	CM	AI	CF	DF
A metadata	63.3	—	—	52.6	31.2	37.5
V metadata	36.0	70.0	65.0	—	—	—
A-V metadata	90.0	91.0	96.0	64.5	100.0	57.1

Figure 6: An image of an unedited (4-camera) BLS movie.

(AR), cardiac massage (CM), [1]AED installation (AI), confirmation of safety for AED (CF), and defibrillation (DF).

We recorded BLS training movie using four cameras and a lapel microphone. The three fixed cameras were allocated around a trainer. The trainer wore a head-set camera and the microphone. A 5-minutes movie including four video channels and speech data spoken by one Japanese trainer was used. Using speech data and image sequences obtained from the movie, audio-visual metadata were generated. The language model selection described in Section 2 was not applied to BLS since the scenarios of ACLS and BLS are different. Video edit was subsequently conducted and an edited video file was created.

Table 3 shows the event detection correctness rates for every time frame (roughly 4500 frames in total) compared with the manual transcription. For audio and visual results, metadata obtained in 3.2.1 and 3.2.2 were respectively used. Some events could not be determined using audio or visual metadata only; for example, using audio metadata, we can only distinguish CC, AI, CF, DF, and the other event. By using audio and visual information complementarily, we become able to distinguish all events. And the average correctness of all events using audio-visual metadata was approximately 87%. These results mean that significant improvements were achieved in all events by using audio-visual metadata produced by the proposed system.

3.5. Survey for video edit system

We also carried out a customer survey to BLS trainers. Before the survey, they watched an unedited movie and an edited movie created by the proposed system. The unedited movie consisted of four camera images is shown in Figure 6. After comparing these movies, they answered some questions. The result of the survey answered by 16 respondents is shown in Table 4. This result indicates that the proposed system can provide well-edited, appropriate video for BLS training.

[1]Automated External Defibrillator

Table 4: Distributions of answer to the questions described below:
(1) Was the automatic edit video better than the 4-camera video as a medical education content ?
(2) Was automatic switch editing appropriate ?

(1) Answer	# resp.
No	2
Not very much	1
So-so	2
A little	10
Yes	1

(2) Answer	# resp.
No	0
Not very much	2
So-so	5
A little	8
Yes	1

4. Conclusion

This paper proposes automatic metadata generation and video edit systems for the medical contents ACLS and BLS. In the metadata generation system, speech recognition is used to identify the segments (event group) and subsequently the events. The system achieved a 78% event detection correctness due to proper language model selection. The video edit system edits multiple videos to a new movie using created metadata. An 87% event correctness was obtained, and from the survey, it is turned out that the edited movie was appropriate and useful.

Our future work includes; (1) improvement of speech recognition accuracy for spontaneous speech in noisy conditions in order to achieve better event detection performance, (2) development of more effective audio-visual integration algorithm, (3) application of proposed system to other task by adapting segment and event identification method, and (4) reduction of computational effort in the recognition modules in order to construct real-time system.

5. Acknowledgement

The authors would like to thank the members of ACLS-Gifu and MEDC for their assistance about contents. This research was carried out as a part of the knowledge cluster initiative “Robotics Advanced Medical Cluster” commissioned by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

6. References

[1] M.Sano et.al., “Generating metadata from acoustic and speech data in live broadcasting,” Proc.ICASSP2005, MSP-P2.4.

[2] H.B.Nguyen et.al., “Robust highlight extraction using multi-stream hidden Markov models for baseball video,” Proc.ICIP2005, pp.III:173-176.

[3] K.Hashimoto et.al., “Retrieval of e-learning media contents using speech recognition,” Proc.VSMM2004, pp.1043-1049.

[4] T.Kawahara et.al., “Free software toolkit for Japanese large vocabulary continuous speech recognition,” Proc. ICSLP2000, pp.IV:476-479.

[5] A.Verma et.al., “Late integration in audio-visual continuous speech recognition,” Proc. ASRU’99.

[6] M.D.Pesce, “Programming DirectShow for Digital Video and Television,” Microsoft press, Redmond, U.S.A. (2003).