Highlight Scene Extraction in Real Time from Baseball Live Video

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

Masahito Kumano
Dept. of Electronics and Informatics
Ryukoku University
Seta, Otsu, 520-2194, Japan
kumano@rins.ryukoku.ac.jp

Kiyoshi Tsukada
Mainichi Broadcasting System, Inc
17-1 Chayamachi, Kita-ku
Osaka, 530-0013, Japan
tsukada@mbs.co.jp

ABSTRACT

This paper proposes a method to automatically extract highlight scenes from sports (baseball) live video in real time and to allow users to retrieve them. For this purpose, sophisticated speech recognition is employed to convert the speech signal into the text and to extract a group of keywords in real time. Image processing detects, also in real time, the pitcher scenes and extracts pitching sections starting from a pitcher scene and ending at the successive pitcher scene. Highlight scenes are extracted as the pitching sections with the keywords such as home run, two-base hit and three-base hit extracted from speech signals.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio input/output, Video*; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing.*

General Terms

Experimentation.

Keywords

highlight scenes, sports live video, speech recognition, acoustic model adaptation, language model adaptation.

1. INTRODUCTION

Recently a large quantity of multimedia contents are broadcast and accessed through TV and WWW. In order to retrieve exactly what we want to know from the multimedia database, automatic extraction of indices or structuring is required, because it is difficult to give them by manual due to their quantity.

MIR'03, November 7, 2003, Berkeley, California, USA.

So far, many studies have been done on this automatic indexing or structuring to multimedia contents based on video sequence analysis[9], video caption analysis[5], closed caption analysis[8] and audio analysis[12]. It is also reported that the collaborative processing of text, audio and video is effective for their indexing or structuring[11].

Multimedia contents can be classified into two groups; one is text oriented contents such as news, documents and drama which are produced according to the well-organized story text. The other is event-oriented content such as sports live video which is produced according to the player's action. The former contents can be indexed using the text information. However, the latter contents require a quick indexing using information other than text, because users wants to retrieve them just after the contents are broadcast.

The purpose of this study is to automatically extract indices from sports (baseball) live video in real time for highlight scene retrieval just after their occurrences. For this purpose, sophisticated speech recognition and image processing are employed. The speech recognition converts the speech signal into the text in real time using LVCSR(Large Vocabulary Continuous Speech Recognition) system and the keywords are extracted from the converted text. This point discriminates our approach from the previous works mentioned above.

On the other hand, the image processing detects, in real time, the pitcher scenes where the pitcher throws a ball to the catcher, and finally extracts pitching sections each of which starts from a pitcher scene and ends at the successive pitcher scene. Highlight scenes are extracted in real time by integrating the image processing and the speech processing. Namely, they are extracted as the pitching sections with keywords such as home run, two-base hit and three-base hit extracted from speech signals.

We propose, in this paper, a sophisticated speech recognition method where acoustic model and language model, originally constructed using a continuous speech database, are both adapted to the baseball live speech. By this method, speech recognition with high accuracy is available for any baseball games and any announcers. We also propose an efficient image processing where mean values and variances of intensities within certain areas are computed in each frame and pitcher scenes are detected quickly and reliably based on these features.

2. SYSTEM OVERVIEW

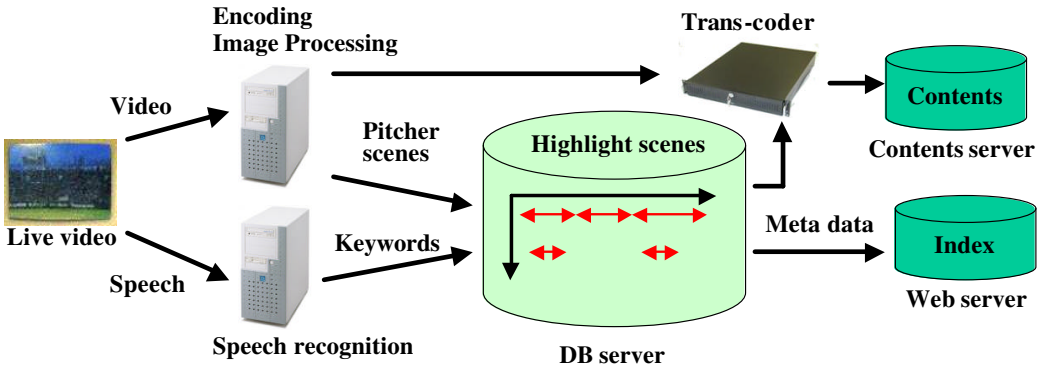


Figure 1: System overview

Fig.1 shows a highlight scene extraction and retrieval system. A sports live video is encoded and processed in real time to extract the pitching sections which start from a pitcher scene and end at the successive pitcher scene, where the pitcher throws a ball to the catcher. The information about pitching sections is stored as meta data on the database server (DB server) in an XML format with time of in-point and out-point of the sections.

Sports live audio is transcribed into the text by speech recognition techniques in real time and keywords are extracted from the transcription. They are also stored on the database server (DB server) as meta data in an XML format with time of in-point and out-point of the keywords. The pitching sections with keywords such as home run, two-base hit and three-base hit are extracted as the highlight scenes. As the sports live audio, we used radio speech in stead of TV speech because the radio speech has much more information about the keywords. Therefore, our live video is composed of TV live video and radio live audio. They are synchronized according to the time stamp.

These meta data are transmitted to the web server as web information. The encoded video is also transcoded and stored on the contents server for web streaming by broad band, a cellular phone and PDA(personal digital assistance). We describe the keyword extraction method from speech data and pitcher scene extraction method from video data in the following sections.

3. KEYWORD EXTRACTION

3.1 Live Speech Feature

Table1 shows a comparison of speaking styles among read speech such as in news and document, lecture speech and live speech. As can be seen, the live speech is noisy, emotional and unclear due to its high speaking rate compared to the other speaking styles. In addition, the live speech is disfluent due to repeat, mistake and grammatical deviation.

In our case, the radio speech was recorded in a relatively quiet booth so that the environmental noise is not so strong. From the above described features of live speech, we constructed the acoustic model by using lecture corpus (CSJ: Corpus of Spontaneous Japanese) including 200 male speakers[6] because there was no baseball speech corpus in the world yet. Then the constructed acoustic model was used as a baseline in speech recognition. The baseline acoustic model is converted to live speech model by adaptation techniques using the live speech data.

The language mode was constructed using baseball text corpus which was originally collected through WWW because there was no baseball text corpus in the world yet. Then the constructed language model was used as a baseline in speech recognition. The baseline language model is converted to live language model by adaptation techniques using the live speech transcription. Hereafter, we describe the employed techniques for the acoustic model adaptation and language model adaptation.

Table 1: Comparison of speaking styles

	Speaking rate	Noise	Emotion
Read speech	7.26 (mora/sec)	Quiet	Weak
Lecture speech	7.31 (mora/sec)	Middle	Middle
Live speech	8.51 (mora/sec)	Noisy	Strong

3.2 Acoustic Model Adaptation

Fig.2 shows the acoustic model adaptation process. The baseline acoustic model (HMM:Hidden Markov Model) was constructed using lecture corpus so that the model is not suitable for the live speech recognition. In order to absorb the difference in both of the speaking style and speakers, the baseline HMM is converted to the adapted HMM by supervised adaptation which utilizes adaptation speech and manually transcribed text data. The adaptation speech was collected from one baseball game (70 minutes) and manually transcribed. The adaptation method is MAP (maximum a posteriori probability) adaptation[3] after MLLR (Multiple linear regression) adaptation[1]. The MLLR adapts the baseline HMM quickly to the target speaking style owing to Affine transformation and the MAP adapts it precisely to the target speaking style based on a posteriori probability.

The adapted HMM is suitable for the baseball live speech, but speech data for evaluation is slightly different from the adaptation speech in speaking style, speaker characteristics and environment noises. In order to absorb this difference, The adapted HMM is further adapted to the speech data for evaluation by an unsupervised technique utilizing input (evaluation) speech and automatically recognized transcrip-

tion. The adaptation method is also MAP after MLLR here. The difference between supervised and unsupervised adaptation lies in the difference of used transcription; manual transcription or automatically recognized transcription. It is clear that the accuracy of supervised adaptation is superior to the unsupervised adaptation.

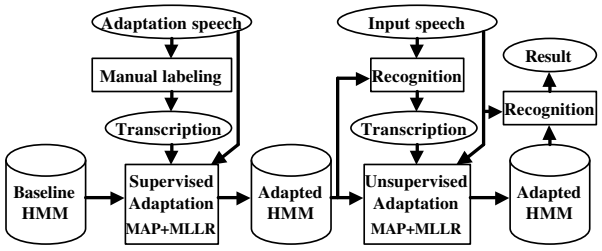


Figure 2: Acoustic model adaptation process

3.3 Language Model Adaptation

Baseball text corpus was originally constructed by collecting baseball text through WWW because there has been no baseball corpus in the world yet. We call this baseball corpus collected from WWW as web text corpus in this paper. The size of the web text corpus is 576,936 words. The collected text are parsed into words through a morphological analysis by ChaSen and bigram / trigram language models as well as the dictionary are constructed using CMU-Cambridge Toolkit.

The baseline language model is constructed using this web corpus. The web text corpus is a collection of the text in a written style so that the baseline language model is not suitable for live (spontaneous) speech recognition. From this viewpoint, the baseline language model is converted to live language model by adaptation techniques using the live speech transcription. We call the corpus used for this adaptation data as adaptation text corpus. The size of the adaptation text corpus is 10,865 words. The transcription of the adaptation text corpus is produced manually for one baseball game (70 minutes). The transcription is parsed into words through a morphological analysis by ChaSen. Then the bigram and trigram language models are constructed using CMU-Cambridge Toolkit as well as the dictionary.

In the language model adaptation, the language model constructed using the web text corpus and the language model constructed using the adaptation text corpus are integrated by the method described in the paper[2]

In the sports live speech, player names and commentator names are usually observed. However, the frequency is not so large to reflect them into the language model. To solve this problem, two classes are employed; one is PLAYER class for a collection of player names and the other is COMMENTATOR class for a collection of commentator names. The language model (bigram or trigram) is constructed using the class names in stead of individual player or commentator names. Namely, in the web text corpus and adaptation text corpus, the player names and commentator names are converted to the corresponding class names PLAYER and COMMENTATOR. In speech recognition, the bigram language probability is used for transition from a word to the class names and the acoustic model probability is computed for all the player names within the class name.

As mentioned earlier, the live speech is fast so that the pronunciation is deviated from the normal one and this causes the speech recognition errors. To solve this problem, the pronunciations of some words in the dictionary are modified manually to absorb the deviation.

3.4 Speech recognition system

We employed a 2-pass decoder as a LVCSR (large vocabulary continuous speech recognition) system as shown in Fig.3[10]. At the 1st-pass, we adopted a lexical tree search using a bigram language model for constructing the word graph. A search method we employed is called "best-word back-off connection" which has been already proposed[4]. This method links the word with the best partial score at each frame to the back-off connection so that it can reduce about half of the processing time without increasing any errors. At the 2nd-pass, the best sentence (word sequence) is searched in the word graph using a trigram language model.

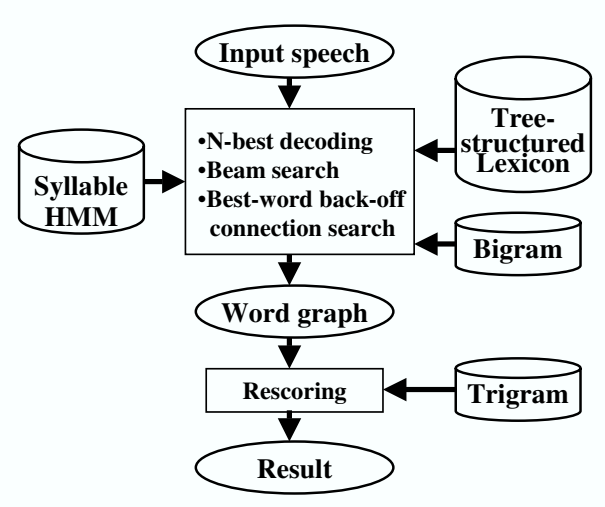


Figure 3: Speech recognition system

3.5 Keyword Extraction

Since the purpose of this study is to extract highlight scenes, the keywords related to highlight have to be prepared in advance. The highlight scene is defined as the scene strongly concerning with the score. From this viewpoint, we prepared the keywords shown in Table2.

Keywords are sometimes observed at irrelevant time before or after the highlight scenes. For example, the announcer remembers the home run and refers to it a little later after it has finished. The difference between the keywords at relevant time and irrelevant time can be observed in the difference of the emotion, especially in the power of the speech.

From this viewpoint, we extract keywords with their time sections and then the power of the keywords is computed within the time section. If the power is bigger than some threshold, then the keyword is confirmed as a true keyword. Otherwise they are rejected as false keywords.

4. PITCHER SCENE EXTRACTION

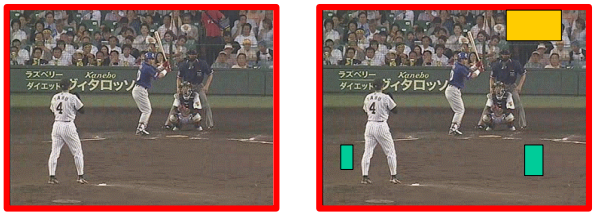
Table 2: Keyword list

Home run, Two-base hit, Three-base hit, Bases loaded, Grand slam, Timely hit, Home steal, Insurance run, The points scored first, Bases-loaded walk

4.1 Feature of Pitcher Scene

Fig.4(a) shows an example of pitcher scenes. The structure is almost fixed, but sometimes slightly shifted in right to left or up to down. The lighting is also almost fixed, but sometimes slightly drifted. Therefore the whole of a pitcher scene is not suitable for a matching template because of the position sifting and light drifting.

In order to solve the position sifting, the observation areas are effective as shown in Fig.4(b). In order to solve the light drifting, the variance is effective as well as the mean value within the observation areas. In the observation area at upper right, the variance of the intensity is large because the audience shows highly textural pattern. On the other hand, in the lower two observation areas, the variance of the intensity is small. The mean value is used to discriminate the audience and ground from the others.



(a) An example of pitcher scenes (b) Observation areas

Figure 4: Pitcher scene detection

According to the above insight, we detected the pitcher scenes based on the mean values and variances within the observation areas. Let A and B denote the upper right and lower observation areas respectively, and μ_A, μ_B, σ_A and σ_B denote the mean value and variance at the observation area A and B respectively. The pitcher scenes are detected according to the following rules. Here θ_1, θ_2, θ_3 and θ_4 are the threshold selected by preliminary experiments. This equation indicates that the audience has high variation and intensity. On the other hand, the ground has low variation and intensity.

$$\sigma_A \geq \theta_1 \quad and \quad \sigma_B \leq \theta_2 \quad and \quad \mu_A \geq \theta_3 \quad and \quad \mu_B \leq \theta_4 \quad (1)$$

In order to achieve the real time extraction of the pitcher scenes, the cut detection technique[7] is applied at first for data reduction, then the detected frames are processed for pitcher scene extraction.

4.2 Highlight Scene Extraction

After the pitcher scenes are detected, the pitching sections are extracted each of which starts from a pitcher scene and ends at the successive pitcher scene. Highlight scenes are extracted as the pitcher sections with the keywords such as home run, two-base hit and three-base hit extracted from speech signals.

5. EXPERIMENTAL RESULTS

5.1 Experimental Condition

We carried out acoustic model and language model adaptation using the adaptation data, and also carried out speech recognition and keyword extraction experiments for the evaluation data both shown in Table3. In the table, the number / number shows the month and day when the game was played.

The baselines for the acoustic model and language model were constructed using CSJ (lecture) speech corpus and web text corpus described in Sec.3.2 and Sec.3.3 respectively. Table4 shows the condition for acoustic analysis (AA) and HMM.

Table 3: Speech data for adaptation and evaluation

Set	Adaptation data	Evaluation data
1	9/8	9/24
2	8/29	10/6

Table 4: Condition for acoustic analysis and HMM

A	Sampling frequency	16KHz
A	Feature parameters	MFCC(39 dim)
A	Frame length	20ms
A	Frame shift	10ms
A	Window type	Hamming
H	Acoustic unit	244 Syllables
H	Mixture Num	32
M	Vowel	5 states with 3 loops
M	Consonant+Vowel	7 states with 5 loops

5.2 Result of Language Model Adaptation

Table5 shows the result of speech recognition for two game sets before and after the language model adaptation. In the table, baseline shows the speech recognition result using the baseline language model constructed by the web text corpus while using the baseline acoustic model constructed by CSJ (lecture) speech corpus. LANG-Adapt shows the speech recognition result after the language model adaptation using the concerning adaptation data.

P.P. shows test set perplexity which shows the language complexity. Corr and ACC show the word correct rate and word accuracy which show the speech recognition performance. The keyword shows the keyword extraction rate before the power discrimination described in Sec.3.5. From the table, it can be seen that by the language model adaptation the P.P. has significantly decreased to about 30%, and Corr and ACC have significantly improved by about 13%. However, the keyword extraction rate was not improved. This indicates that the language model contributes not to improve the missed keywords but contributes to reduce the falsely accepted keywords.

5.3 Result of Acoustic Model Adaptation

Table6 shows the speech recognition result before and after the acoustic model adaptation while the language model was already adapted. In the table, the baseline and Acoustic-Adapt show the result before and after the acoustic model

Table 5: Result of language model adaptation(%)

Set	Processing	P.P.	Corr	Acc	keyword
1	Baseline	258.4	43.8	35.1	82.0
	LANG-Adapt	75.8	58.3	51.9	81.6
2	Baseline	248.7	38.3	27.3	80.0
	LANG-Adapt	69.2	49.2	38.4	76.7

adaptation. From the table, it can be seen that the speech recognition performance was improved by almost 20% and the keyword extraction rate was improved by almost 15%. These improvements are attributed to the effectiveness of the speaker and environmental noise adaptation.

Table 6: Result of acoustic model adaptation(%)

Set		Corr	Acc	keyword
1	Baseline	58.3	51.9	82.0
	Acoustic-Adapt	78.6	74.6	93.9
2	Baseline	49.2	38.4	80.0
	Acoustic-Adapt	72.8	63.8	97.1

5.4 Result of Keyword Extraction

Table7 shows the keyword extraction rate after the power discrimination described in Sec.3.5. For set1, true keywords were 2 "home runs" and they were correctly extracted by speech recognition and power discrimination. However, 2 other "home runs" were falsely extracted due to power discrimination error for the correctly recognized 2 "home runs". For set 2, true keywords were 2 "home runs" and 2 "timely hits". Among them 2 "home runs" were correctly extracted. However, 2 "timely hits" were missed due to speech recognition error for one keyword and power discrimination error for other keyword.

By the power discrimination, the false keywords are significantly reduced and the true keywords are successfully extracted. However, there are still false extraction and missing keywords. They can be explained partly because the acoustic model is still weak and partly because the power discrimination method has its limitation that the keywords occurring at the irrelevant time with strong power are sometimes extracted as true keywords.

Table 7: Result of keyword extraction

Set	Extraction (Correct #/True #)	False #
1	2/2	2
2	2/4	0

5.5 Result of Pitcher Scene Extraction

We also carried out the experiments of the pitcher scene detection and pitching section extraction for four evaluation data (9/8, 9/20, 9/22 and 10/1; month/day). The image size was 320x240 pixels and the observation areas A and B were set to 6x12 and 24x12 pixels respectively.

The result is shown in Table8. The pitcher scene detection rate (recall and precision) showed about 90% at average and is almost sufficient for the highlight scene extraction and retrieval.

The pitching sections are extracted as the sections starting from a pitcher scene and ending at the successive pitcher scene. Highlight scenes are extracted as the pitching sections with the keywords extracted in Sec.5.4. We are now developing more advanced video processing techniques to extract the pitching sections by employing an automatic learning method of observation areas in terms of the size, position and features using the training data.

Table 8: Result of pitcher scene detection

	Definition	9/8	9/20	9/22	10/1
Detected #		158	121	242	152
Correctly detected #	A	152	115	217	125
Falsely detected #	B	6	6	25	27
Missing #	C	0	0	3	26
Recall (%)	A/(A+C)	100	100	98.6	85.3
Precision (%)	A/(A+B)	96.2	95.0	89.7	84.8

6. CONCLUSION

In this paper, we proposed the highlight scene extraction methods for baseball games using sophisticated speech recognition and efficient image processing. In the speech recognition, the language model adaptation achieved 13% improvement at word accuracy in large continuous speech recognition using corpus integration, name classes and pronunciation modification. The acoustic model adaptation achieved 20% improvement at word accuracy and about 15% improvement at keyword extraction rate using the supervised and unsupervised adaptation.

In the image processing, almost 90% recall and precision were achieved for the pitcher scene detection. This processing is performed in real time so that after broadcasting the sports live, even during the broadcasting, the highlights scenes can be retrieved by cellar phone, PDA and internet.

In many researches on sports game videos, often used are closed caption, video caption and scores. However in our research, we did not employed them because our final goal is to develop an automatic speech transcription and scoring system in near future.

In future, we are going to study further adaptation techniques, keyword extraction methods and also an automatic structuring method of baseball games for retrieval of the scenes other than highlight scenes. Emotion analysis is also going to be studied for keyword discrimination using pitch information as well as power information[12].

7. ADDITIONAL AUTHORS

Additional authors: Jun Ogata (National Institute of Advanced Industrial Science and Technology, email: `jun.ogata @aist.go.jp`), Masakiyo Fujimoto (Ryukoku University, email: `masa@arikilab.elec.ryukoku. ac.jp`), Takeru Shigemori (Ryukoku University), Tsuyoshi Kaneko (Ryukoku University), Nobuo Kanzaki (Ryukoku University, email: `nob@arikilab.elec.ryukoku.ac.jp`), Shin Hamaguchi (Mainichi Broadcasting System Inc., email: `shin -h@mbs.co.jp`) and Hajime Kiyose (Mainichi Broadcasting System Inc., email: `h-kiyose@mbs.co.jp`).

8. REFERENCES

[1] C.L.Leggetter and P.C.Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[2] A. Ito, M. Kohda, and M. Ostendorf. A new metric for stochastic language model evaluation. In *Proceedings of Eurospeech99*, pages 1591–1594. ISCA, 1999.

[3] J.L.Gauvain and C. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

[4] J.Ogata and Y.Ariki. An efficient lexical tree search for large vocabulary continuous speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, pages 967–970, 2000.

[5] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki. Indexing of baseball telecast for content - based video retrieval. In *Proceedings of International Conference on Image Processing*, pages CD–ROM. IEEE, October 1998.

[6] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous speech corpus of japanese. In *Proceedings of LREC2000*, pages 947–952, 2000.

[7] M.Kumano and Y.Ariki. Automatic useful shot extraction for a video editing support system. In *Proceedings of MVA*, pages 310–313, 2002.

[8] N.Babaguchi. Towards abstracting sports video by highlights. In *Proceedings of International Conference on Multimedia and Expo*, pages 1519–1522. IEEE, 2000.

[9] P.Chang, M.Han, and Y.Gong. Extract highlights from baseball game video with hidden markov models. In *Proceedings of International Conference on Image Processing*, pages 609–612. IEEE, 2002.

[10] S.Ortmanns, H.Ney, and X.Aubert. A word graph algorithm for large vocabulary continuous speech recognition. volume 11, pages 43–72, 1997.

[11] Y.Chang, W.Zeng, I.Kamel, and R.Alonso. Integrated image and speech analysis for content-based video indexing. In *Proceedings of International Conference on Multimedia Computing and Systems*, pages 306–313. IEEE, 1996.

[12] Y.Rui, A.Gupta, and A.Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of ACM International Conference on Multimedia*, pages 105–115. ACM, 2000.