Generating Metadata from Acoustic and Speech Data in Live Broadcasting

Masanori Sano, Hideki Sumiyoshi, Masahiro Shibata and Nobuyuki Yagi

NHK Science and Technical Research Laboratories

ABSTRACT

This paper describes a method to generate metadata for TV programs in real-time by utilizing acoustic and speech data in live broadcasting. Various styles of watching TV programs can be provided by using metadata related to the content of the program. The acoustic data to be processed in our case is crowd noise in a football (soccer) stadium, and the speech data is an announcer's voice. The crowd noise is closely related to not only spectators' emotions but also their attention and expectations. In other words, a part in which the crowd noise rises corresponds to an important event in the game. Because the crowd noise conveys no further information about what happened in the scene, the announcer's voice, after speech-to-text conversion, is processed to extract further meaning. By combining these two processes of identifying and extracting, content-based segment metadata is generated automatically. This method was applied to generating metadata for six professional football games, by which its effectiveness was verified.

1. INTRODUCTION

The growing number of TV channels and programs is the reason behind the trend in which viewers watch only highlights or the most interesting parts of programs before flipping channels. A project that reflects this trend is the "Home-server-based service" being developed by the TV-Anytime Forum [1]. This service is intended to promote new styles of watching programs by exploiting segment metadata sent with the content. Although the importance of segment metadata has increased, the technology for generating and handling it has not kept up. So far, NHK (Nippon Hoso Kyokai or the Japan Broadcasting Corporation) has developed a framework for program production with metadata, where the information (the program storyboard) related to the TV program is collected and turned into metadata while the program is being produced [2][3]. Unfortunately, this framework can't be applied to a live broadcasts of sports events in which no scenario of events can be written. Sports programs have relatively stable high audience ratings and other programs such as weekly sports digests and news often use clips from the live coverage. Thus there is a significant demand for segment metadata to be included in "live" contents.

Many researches on generating metadata, in other words indexing, for TV programs have been done. Some of them extract the events by analyzing video [4]. Others analyze audio data to extract highlights from the program [5]. Subtitling and closed captioning have also been used in some reports [6]. Recently, a combination of such analyses has shown even higher effectiveness [7][8].

Our approach is ideally suited to broadcasters because it uses data collected during program production to generate reliable metadata. Specifically for this report, we generated segment metadata for football (soccer) programs in real-time. There are basically two steps to generate segment metadata. The first step is extracting the specific scenes, in other words, segmentation. The second step is labeling the extracted scene, for example, annotating who does what. We use specific audio data for each step. One type is crowd noise in the stadium, and the other is the announcers' voices. The change in crowd noise is a reasonable index to extract scenes, because it is closely related to spectators' emotions, attention, and expectations. When the crowd noise rises, something has definitely happened, and we use this rise to trigger scene extraction. After the scene has been extracted, we abstract its semantic meaning by analyzing the announcers' comments. By combining scene extraction by noise level change and semantic analysis, segment metadata is automatically generated. We used our method to generate metadata for six football games and confirmed its correctness.

Section 2 outlines our method and the audio data to be processed. Crowd noise and announcer's comments are analyzed in Sections 3 and 4, respectively. Section 5 describes the test set and the evaluation and discusses the results. Section 6 includes conclusions and future work.

2. PROPOSED APPROACH

The algorithm is illustrated in Figure 1. Two kinds of audio data are used to generate metadata. One is crowd noise in the stadium and the other is the announcer's voice. The left side in Figure 1 indicates the steps of crowd noise analysis using a base microphone that is set up at central high position of the stadium to collect the whole crowd

noise for program production. This process extracts important scenes without analyzing data to find out what happened. Simply put, any part in which the crowd noise rises is extracted and identified as an important scene for segment metadata, because the crowd noise reflects spectators' emotions, attention, and expectations and thus is closely related to important events in a game.

The right side of the figure shows the steps in analyzing the audio data to identify the speech of the announcer. The aim of this analysis is to abstract information on who does what for annotating the extracted scenes. The signal is collected from the announcer's headset and converted into sentences. These sentences are parsed with morphological and syntactic analysis with two dictionaries. One is a list of players' names, for answering 'who', and the other is a list of football terms, i.e. names of plays such as shot, foul, etc., for answering 'what'. The information on who does what is extracted together with a time code.

The extracted important scenes and the information on who does what are then integrated. To put it concretely, the data of important scenes consists of two times that indicate the starting point and ending point of the scene. On the other hand, the information on who does what also has times indicating when the words were spoken. The integration is based on these times.

3. CROWD NOISE ANALYSIS

There are various strategies to analyze acoustic data, e.g., frequency analysis, using features calculated from short-time energy [9]. We would like to be able to generate segment metadata in real-time for future home-server services. For this purpose, only the short-time energy is used to reduce the amount of calculations. The short-time energy (referred to simply as energy) is defined as (1),

$$E = 10 \times \log_{10} \frac{\sum_{n=1}^{N} \left(\frac{Pn}{MAX} \right)^2}{N} \qquad (1)$$

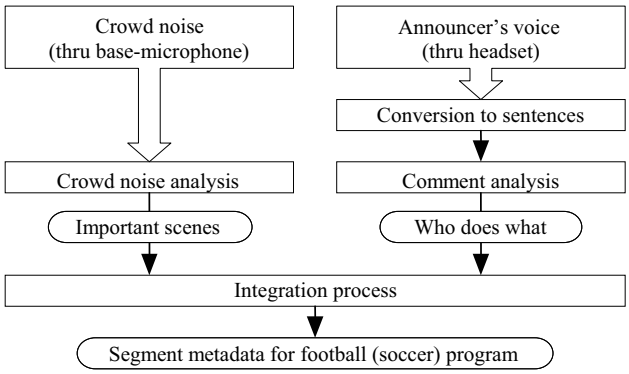

Figure 1: Overview of proposed method

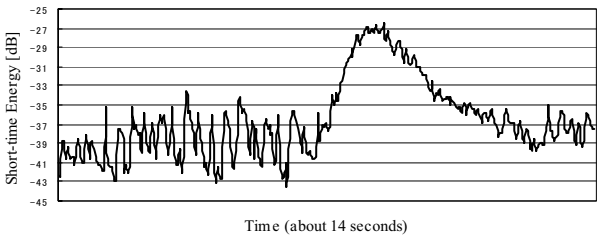

Figure 2: Typical pattern of crowd noise

where N is the number of time points that determine the window size, MAX is the maximum amplitude of the audio signal, and Pn is the amplitude of the audio signal at the sample index n. The crowd noise was digitized at 11 kHz and 16 bits; therefore MAX was 32768. As for the window size, because it is said that periods of less than 50 msec can be treated as steady states for acoustic analysis, we chose 46 msec (N =512) to simplify program coding. Figure 2 shows a typical pattern of short-time energy that includes loud crowd noise. The whole period of this figure is about 14 seconds. The rising noise corresponds to the hump. Statistically speaking, this peak comes around 0.7 second after an event has happened in the game.

To extract such data in real-time, we used a dynamic threshold that outputs the start point and end point each time the crowd noise rises and falls while the game is going. Because various fixed thresholds as current methods need prerequisite information to decide the threshold and retrieve undesirable parts due to the characteristic of fixing. Figure 3 illustrates the change in energy. The following conditions are checked every 46 msec to detect the start and end points.

Condition 1: There is an interval in which a certain level ***E0***+***E1*** is continuously exceeded for at least a period ***D2*** between ***Ps*** and ***Pe***, where ***E0*** is a base line that corresponds to the average energy during ***D0***. ***Ps*** is at the center of ***D0***.

Condition 2: There is no energy sample below the average ***E2*** in ***D0***, where ***E2*** indicates the average energy between the very beginning of the football program and the latest sample point ***Pe***.

If both conditions 1 and 2 are satisfied, ***Ps*** is kept

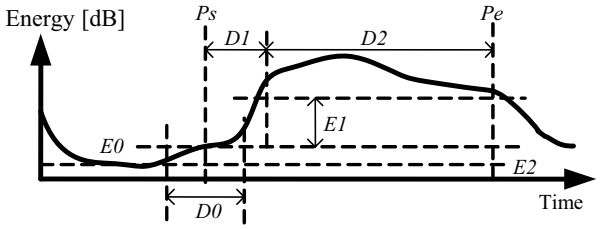

Figure 3: Algorithm of dynamic threshold

as the starting point. To detect the end point, we check condition 1 at the new time position until the energy is less than ***E0***+***E1***. When condition 1 is no longer satisfied, the current ***Pe*** is kept as the end point. Condition 2 is introduced to prevent detection of the part where a large amount of cheers breaks out suddenly, after relative silence. The duration of the part extracted by this algorithm is from 1 to 5 seconds. We added the 10 seconds before and the 5 seconds after this scene based on our experience of when events drawing the audiences' attention actually occur.

4. ANALYSIS OF ANNOUNCER'S COMMENTARY

The announcer's commentary is used to extract the information on who does what and keywords. This is based on the results of analyzing the speech in the running commentary of announcers during TV football programs. There are, broadly speaking, two kinds of comment in these programs. One is 'game descriptions' in which the commentary is synchronized with the flow of the video data, mainly names of players involved in plays and actions such as shot, kick, etc. The other is 'game commentary' not synchronized with the video, such as explanations of the previous play or the summary of the game. We focused on game description because the players' names and actions are directly related to the video data.

To abstract the information on who does what, we use morphological and syntactic analysis with two dictionaries. One dictionary is a list of players' names. The other is a list of football events. We listed all similar terms separately, for example, not only "shot" but also "loop shot," "banana shot," etc. In total, 93 events are included in the dictionary. After extracting the player names and events, syntactic analysis is performed to check to see if there is a correlation (dependency structure) between the player and event. If there is the correlation, the information on who does what is abstracted. Having this relation is much different from the situation in which the player and the event simply exist independently. With it, we can directly search for and retrieve specific scenes such as "Beckham's free kick." Besides this information, the players and events that have no correlation each other are also outputted as individual keywords. Therefore, the outputs from this analysis are information on who does what, the player's name and event with the time code.

5. EXPERIMENTAL RESULTS

Figure 3 shows the four parameters we initially set. Thirty-nine scenes from two football games were identified by human operators as important scenes. A statistical analysis of these parts gave the following parameters: ***E1***=2 [dB], ***D0***=750 [msec], ***D1***=350 [msec], ***D2***=1000 [msec]. To test the effectiveness and robustness of our method, six football games including the two games used for deciding parameters were tested. These games had audience ranged from 19,000 to 44,000 and were commented by four announcers. The football games were of the Japan professional football league (J-league) held from July to November, 2003.

5.1. Scene Extraction Result

The extracted scenes were evaluated two ways. One was a comparison with the human-selected important scenes. The other was a comparison with the broadcasted sports digests of these games.

For the first evaluation, we selected important scenes from the six games (these scenes were termed the right answers). They included, for example, goal scenes, shots close to the goal and players leaving the game after receiving a red card from a referee. The results are shown in Table 1. The total recall ratio reached almost 95%. This means almost all the scenes that were identified as being important were extracted by the algorithm. On the other hand, total precision was only around 37%. The extracted scenes that were not "right answers" were mostly scenes showing fouls, yellow cards, ordinary shots, good offensive or defensive plays or scenes with organized cheers. Although these may not be suitable for sports digests, they are mostly important for tracking the game flow and deserve to be indexed. From the view point of extracting these important scenes, the precision comes out 86.7%. Only the scenes with cheers being led to a song can be deemed wrong answers. This means the algorithm produces quite suitable extractions for the segment metadata.

Table 1. Algorithm's results matching human-selected scenes

Game	#Ra	#Ex	#RaE	RC	PCd	PCi
1	19	62	19	100	30.6	82.3
2	26	66	26	100	39.4	89.4
3	17	42	15	88.2	35.7	83.3
4	19	49	18	94.7	36.7	81.6
5	20	46	18	90.0	39.1	91.3
6	11	21	10	90.9	47.6	100
Total	112	286	106	94.6	37.1	86.7

#Ra: Right Answer, #Ex: Extracted Scene, #RaE: Right Answer in Extracted Scene, RC(recall) = #RaE / #Ra, PC(precision) = #RaE / #Ex, PCd: PC for digest, PCi: PC for important scene

For the second evaluation, the broadcasted digest programs summarizing J-League games were used as a reference. The results are shown in Table 2. #Ra is the number of scenes that were included in the digest programs. #RaE is the number of extracted scenes that corresponds to the same scenes in the digest programs.

The recall ratio also reached about 85%. The digest scenes not extracted by this algorithm showed facial expressions of the coaches, players' exchanges, specific players in action, etc. These tended to relate to the topic at hand or the program editor's concept. From these results, we verified that our algorithm extracts most of the important scenes in digest programs.

Table 2. Algorithm's inclusion of scenes in digest program

Game	#Ra.	#RaE	Recall
1	4	4	100
2	5	5	100
3	11	9	81.8
4	9	8	88.9
5	8	6	75.0
6	10	8	80.0
Total	47	40	85.1

5.2. Semantic Extraction Result

Manually transcribed text was used for this experiment because the accuracy of speech recognition is currently around 80% [10]. To evaluate the correctness of the extracted information, we checked if the information (who, what and who does what) actually appeared in the extracted scene. The results are shown in Table 3. 86.2% of extracted player names referred to players in the scenes and 75.2% of extracted events actually happened in the scenes. Some errors occurred because the announcer's commentary didn't only follow the game, but was made up of more general commentary. The correctness of the extracted "who does what" information reached about 86%. The correctness of all information was around 83%. Accordingly, we confirmed that announcers tend to speak game descriptions when the crowd noise rises in the game. This means our approach is valid for labeling extracted scenes.

Table 3. Correctness of extracted information

Game	Who	What	Who-what	Total
1	85.8	81.0	93.1	83.6
2	81.4	56.9	66.7	74.9
3	91.8	77.3	87.5	88.0
4	84.3	75.4	71.4	82.8
5	85.3	80.0	92.3	82.5
6	97.8	81.3	83.3	92.4
Average	86.2	75.2	86.0	82.7

6. CONCLUSION

We have developed and tested a method to generate segment metadata for live sports broadcasts of football (soccer). As inputs, we used two kinds of audio data present in live broadcasts. One is crowd noise in the stadium, and the other is the announcers' commentary. Crowd noise is so closely related to the spectators' emotions and attention that its volume can be used as a barometer of importance of the individual events in a game. The algorithm exploiting a dynamic threshold to extract the parts in which something interesting happened follows the game's time line. As for the announcer's commentary, we paid attention to its functionality of annotating the scenes. Morphological and syntactic analyses are used to abstract the information on who does what. The extracted scenes and the extracted semantic meaning are combined based on the time code.

We verified our method by using it to analyze a diverse collection of six football games. The results were that 86.7% of the extracted scenes are appropriate parts for segment metadata. The recall ratio of comparing the human-selected important scenes was 94.6%, and that of comparing the scenes included in the digest program was 85.1%. In addition, 82.7% of the extracted semantic information was correct. We confirmed that our method can output valid segment metadata in real-time.

Future work includes extracting the deeper semantic meaning of the scene. It might be applying domain knowledge to complement the extracted information that is fragmented semantically, for example, the flow of the game.

REFERENCES

[1] TV-Anytime Forum: http://www.tv-anytime.org/

[2] R. Nikaido, Y. Fujita, K. Tatsuguchi, Y. Tokoro and M. Yoshida : "Total Management System for Creative Program Production", NAB. On Broadcast Engineering Conference, 2000

[3] H. Sumiyoshi, Y. Mochizuki, S. Suzuki, Y. Ito, Y. Orihara, N. Yagi, M. Nakamura and S. Shimoda : "Network-based Cooperative TV Program Production System", IEEE Transaction On Broadcasting, Vol.42, No.3, pp.229-236, Sep. 1996.

[4] Y. Gong, L.T. Sin and C.H. Chuan : "Automatic Parsing of TV Soccer Programs" IEEE Conf on Multimedia Computing and Systems, pp.167-174, 1995.

[5] Y. Rui, A. Gupta and A. Acero : "Automatically Extracting Highlights for TV Baseball Programs," Proc. ACM Multimedia, pp.105-115, Oct. 2000.

[6] A. Merlino, D. Morey and M. Maybury : "Broadcast News Navigation using Story Segments," Proc. ACM Multimedia, pp.381-391, 1997.

[7] Informedia: http://www.informedia.cs.cmu.edu/index.html

[8] N. Babaguchi, S. Sasamori, T. Kitahashi and R. Jain : "Detecting Events from Continuous Media by Intermodal Collaboration and Knowledge Use," Proc. IEEE ICMCS, Vol.1, pp.782-786, June 1999.

[9] T. Zhang and C.C.J. Kuo : "Heuristic Approach for Generic Audio Data Segmentation and Annotation," ACM Multimedia, pp.67-76, 1999, Orland, FL.

[10] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe, and T.S. Kobayakawa : "Speech Recognition for Subtitling Japanese Live Broadcasts", ICA-2004 (The 18th International Congress on Acoustics), I165-I168, April 5, 2004, Kyoto, Japan