Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Slideshow system that automatically switches photographs based on a musical acoustic signal

Nao SHOJI (1) and Masanobu MIURA (2)

(1) Graduate School of Science and Technology, Ryukoku University, Shiga, Japan
(2) Faculty of Science and Technology, Ryukoku University, Shiga, Japan

PACS: 43.75.-z Music and musical instruments

ABSTRACT

This study describes a method for automatically switching photographs when producing slideshows. Specifically, the appropriate times for switching photographs were determined by the beat times of music excerpts. Also, an investigation was conducted to confirm the relationship between the time of cutting in cartoon films and the time function of flux in accompanying musical excerpts, and we found a strong correlation between them. The correlation is then used in the slideshows generated by the proposed system. To test the performance of the proposed system, slideshows it produced were evaluated, and we compared the proposed method with other conventional methods in terms of the congruency of music and photograph switching in a slideshow. The results confirmed that the slideshow the proposed system produced gave a stronger feeling of congruency than those other methods produced.

1. INTRODUCTION

A slideshow is an image display in which photographs are continuously switched while music plays. Some slideshows synchronize the photograph switching with the music while others do not. Manually producing a slideshow with synchronized photograph switching and music excerpts requires several time-consuming steps, such as individually setting the time of each photograph. To solve this problem, a slideshow system was developed that automatically switches photographs at the time of onsets that appear clearly in music excerpts [1]. However, it is still unclear whether or not such onsets are the appropriate time for switching photographs. Therefore, we developed a method for estimating the appropriate time to switch photographs from acoustic signals for automatic production. Thus, a slideshow system was constructed on the basis of the proposed method. Firstly , we investigate the appropriate switching time and appropriate switching interval for a slideshow by using actual cartoon films. Secondly, the time to switch photographs is determined on the basis of results of investigating the appropriate photograph switching using acoustic signals. Finally, the appropriateness of a produced slideshow is evaluated by an evaluation experiment.

2. THE APPROPRIATE PHOTOGRAPH SWITCHING

2.1 The congruency of music and video

A slideshow is used as a movie, and a previous study indicated that the subjective evaluation on a video with synchronized music should give better results than that on a video without music, because a video with synchronized music gives people the feeling of congruency [2]. Thus, a subjective evaluation of a slideshow with synchronized music is expected to give better results than that without. A manually produced slideshow is thought to also have congruency between photographs and music excerpts, implying that people implicitly want the congruency between them. Therefore, the time someone switches photographs is not thought to be independently determined from the music excerpt in a slideshow. In sum, when a slideshow is automatically produced, it must have a strong congruency between photographs and music. Thus, the automatic production of a slideshow needs to deal with the contents of a music excerpt, with respect to its beats.

2.2 Appropriate time for photograph switching

It is thought that the time for switching photograph needs to be determined on the basis of the music excerpt content. Regarding to the accents of video and music, a previous study investigated factors to determine the congruency of video and music. The results indicated that a subject feels congruency when audio-visual content has synchronized video and music accents [3], where the accent of video means motion of objects or cutting of a movie. This study claimed that the accents of video and music corresponded to the photograph switching in the slideshow and the beats of music, respectively. Therefore, the slideshow is expected to have a strong congruency when photographs were switched in accordance with the beat in the musical excerpt in the slideshow. Thus, in this study, photographs are switched in accordance with the beat of the music excerpts in a slideshow.

2.3 Appropriate interval for photograph switching

Section 2.2 described how the appropriate time for switching photograph depends on the accent of music. However, the accent of a video does not always appear at a constant rate in an actual video with music. For example, the accent of a video, represented as motion of objects or cutting of a movie, occurs a number of times in an exciting section of an accompanying excerpt. To confirm the relationship between motion and music in a video, actual slideshows should be investigated, but investigatable slideshows are rarely seen, because they are a new visual media and so are few in number. Here, the switching photographs in the slideshow were interpreted as an accent of a video, which is represented as motion of objects or cutting of a movie. Therefore, actual cartoon films could be used as targets of this investigation. Time-series data of cartoon films was

evaluated on a 5-point subjective activity scale for every 5 seconds by 5 subjects. The results showed that the subjective activity scores for motion in cartoon films have a strong relationship with several musical features. Specifically, the correlation coefficient was obtained between the time function of the musical features and the subjective activity score of motion. The features extracted from acoustic waveform such as flux, centroid, rolloff, zerocrossings, and the acoustic level were used in the investigation, where flux is an index of spectral change, centroid is an index of spectral brightness, rolloff is an index of spectral shape, zerocrossings is an index of the amount of noise, and the acoustic level is a power of sound. These values were calculated for every 5 seconds.

Table 1 shows the correlation coefficient r between the musical features and the subjective activity score in each cartoon film. The results confirmed that the p values in some cartoon films were less than 0.1, 0.05, or 0.01, so seven cartoon films among ten have a somewhat strong relationship between the time function flux of the musical excerpts and the subjective activity score for motion. Therefore, the appropriate switching of photographs for a slideshow could be defined as follows:

- The photographs should be switched on the beat of the music excerpt.
- The number of switches is greater when the value of flux is higher than its mean value.

Table 1: Correlation coefficient r between musical features and subjective activity score in each cartoon film.

ID	samples	flux	centroid	rolloff	zerocrossings	acoustic level
0	17	-0.189	0.069	0.032	0.127	-0.173
1	18	0.331	0.132	-0.479 *	0.079	-0.133
2	18	0.326	0.375	-0.106	0.275	0.308
3	19	0.724 **	-0.006	-0.213	-0.142	0.386
4	28	0.462 *	0.308	0.315	0.281	0.122
5	19	0.665 *	0.481 *	0.179	0.502 *	-0.172
6	18	0.879 **	0.242	-0.048	0.351	-0.014
7	18	0.764 **	0.758 **	0.332	0.787 **	-0.493 *
8	24	0.772 **	-0.168	-0.306	0.017	0.410 *
9	19	0.514 *	0.165	-0.130	0.260	-0.233

+: p<.10 *: p<.05 **: p<.01

3. OUTLINE OF THE PROPOSED SYSTEM

Section 2.3 determined the beat to be the appropriate time for switching photographs. Thus, the proposed system firstly extracts the beat times from input acoustic signals. Secondly, the proposed system changes the rate of switching by observing the mean value of the flux, so it introduces a phrase section, which is defined between the two beginning times of the musical phrases. Finally, the photographs are switched on the beat of the music excerpt, and the rate of switching changes as the value of flux is observed.

Figure 1 outlines the proposed system. The input of the proposed system is an acoustic signal, the format of which is monaural, 16 bit, and 44100 Hz.

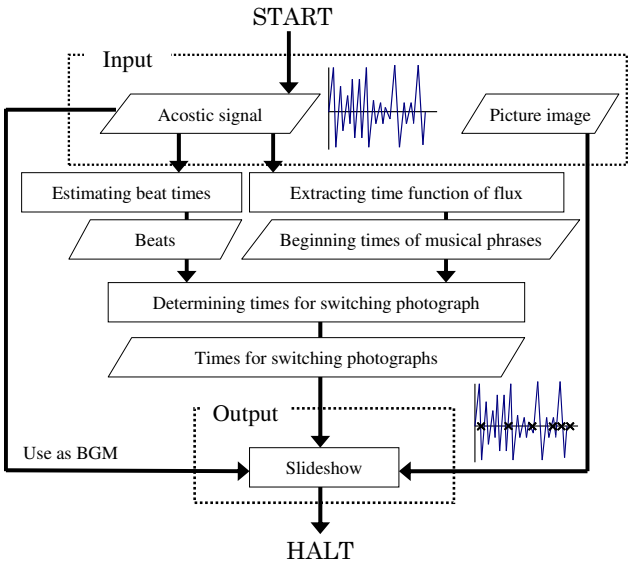

Figure 1: Outline of proposed system

4. ESTIMATING THE BEAT TIMES

4.1 Definition of beat in this study

First, the proposed system extracts “the beats ”and “the beginning times of musical phrases ”. In general, the beat is represented by something periodical, like accents such as high acoustic levels of acoustic waveform. Therefore, it was thought that the beat times in music excerpts could be interpreted as the accent times of a periodically performed sound such as the drums. Also, when the accent does not exist in a specific section on the waveform, the beat is assumed to exist in that section, because the lack of an accent is compensated by the periodic accents in neighbourhoods. The beat in this case is called “the missing beat ”.

In other words, the beats were thought to be “the explicit or implicit periodic accents ”. Thus, the times of beats were determined by the periodic accents and the missing beats by the proposed system.

4.2 Estimation of the beat times

In this study, “the accent time ”means when an occurs. Also, the cycle of appearing accents is called “CAA ”. Thus, the beat times were assumed to be based on both the accent times and the CAA. Additionally, an index of possibility of existence for the beat times is called “beatness ”, where the accent times are peaks of a locally averaged amplitude envelope in an acoustic signal, because an amplitude in an acoustic signal becomes high on the accent times. Additionally, the amplitude envelope in an acoustic signal was regarded as a periodic function of accent times. Thus, by analysing the frequency of the amplitude envelope, the CAA was expected to be obtained.

Figure 2 shows the flowchart of determining the beat times. Firstly, the proposed system calculates the sequence of accents for an inputted acoustic signal to obtain the accent times and the CAA (in Fig.2, (i)). Secondly, the accent times and the CAA are obtained from the calculated sequence of accents (in Fig.2, (ii) and (iii)). Thirdly, the beatness is obtained on the basis of both the accent times and the CAA (in Fig.2, (iv)). Finally, the times of high beatness is outputted as the beat times by the proposed system.

4.3 Calculation of the sequence of accents

Section 4.2 described the calculation sequence of accents to estimate the beats in the proposed system.

Figure 3 shows the flowchart for obtaining the sequence of accents. Firstly, the proposed system decreases the sampling rate of the input acoustic signal under the temporal resolution required to estimate the beats. The proposed method uses a new method that can obtain the sequence of accents by removing the high frequency component of the curve and decrease the sampling rate by calculating local averages of the curve. In this study, the new method is called “smoothing downsampling ”(in Fig.3, (i)). X_t is defined as the amplitude envelope calcu-

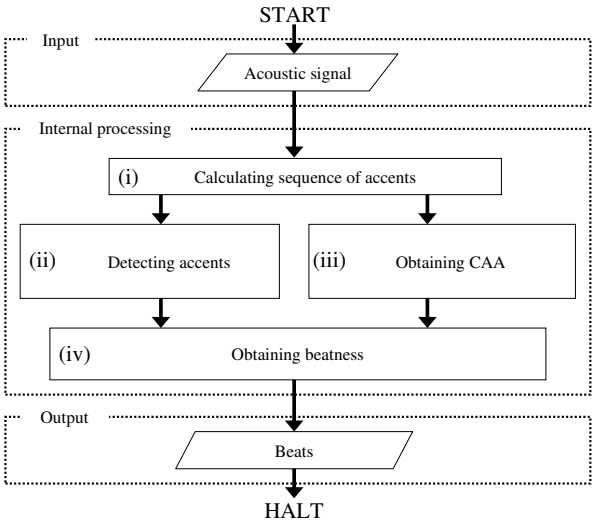

Figure 2: Flowchart of determining the beat times

lated by the smoothing downsampling, shown in Eq.(1), where x_i is the input acoustic signal, $i(i = 0,1,2,..,I)$ is the time of the input acoustic signal, I is the size of input signal. and D is number of samples to be down sampled.

$$X_t = \frac{1}{D}\sum_{j=0}^{D-1}(x_{t+j})^2 \tag{1}$$

Here, D was set as $D = 1100$ (i.e. approximately 250 msec) in accordance with the consideration of the allowable errors in the subjective evaluation of the beat times. Generally, amplitude envelope obtained by smoothing downsampling has many peaks that cannot be thought of as accents, so the proposed system detects more of them than expected. Regarding this problem, a previous study reported that the curve of peaks required to estimate onset was obtained by ignoring negative values of the difference between the time function indicating onsets and the low frequency component [5].

We apply this method to the proposed system to obtain both the amplitude envelope of the accent times and the CAA. Specifically, firstly the proposed system calculated the first amplitude envelope X'_t by a simple moving average (11 samples, i.e. approximately 274 msec) from the time function X_t that was calculated by the smoothing downsampling (in Fig.3, (i)). Secondly, the proposed system calculated the difference between the X'_t and the second amplitude envelope X''_t which is a low frequency component of the first amplitude envelope (in Fig.3, (iii)), and only positive values were kept.(in Fig.3, (iv)). As a result, the sequence of accents X'''_t was obtained (in Fig.3, third amplitude envelope), in which X'''_t has the peak of the accent times. X''_t is a simple moving average of 51 samples (i.e. approximately 1272 msec, in Fig.3, (ii)). Finally, a sequence of accents X'''_t was used to calculate of the accent times and the CAA.

4.4 Calculation of the accent times and the CAA

The accent time is represented as t_p, when the differential value at the time t changes from positive into negative at the peak of the sequence of accents described in section 4.3. Additionally, the CAA is obtained from a low frequency component of the amplitude envelope of X'''_t. Here, an overall power spectrum $Q[f]$ is obtained as the summation of all times for specific frequency bins f, where the frequency of X'''_t for a specific bin was analysed using 4096 samples (i.e. approximately 102 sec).

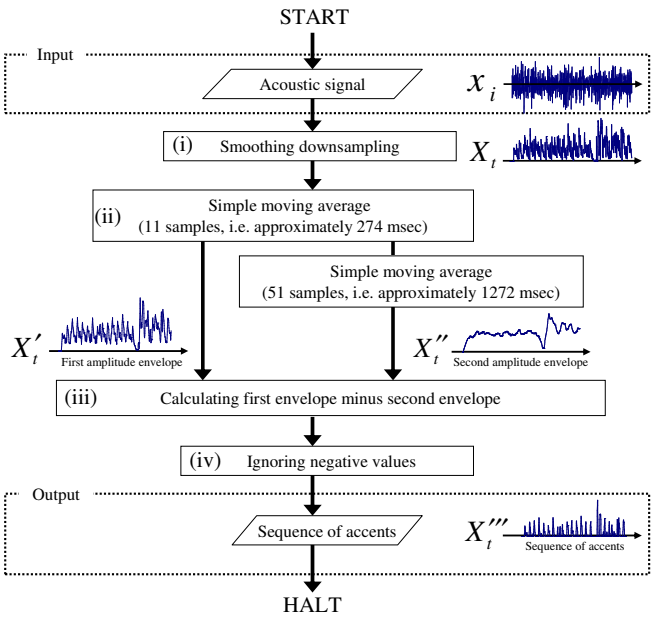

Figure 3: Flowchart of obtaining sequence of accents shown Fig.2 (i)

In sum, an obtainable frequency resolution on the proposed method is 0.0098 Hz (i.e. approximately 0.59 bpm). Finally, a maximum frequency in the range between 0.84 Hz (i.e. approximately 50 bpm) and 2.00 Hz (i.e. approximately 120 bpm) is dealt with as the CAA represented as A.

4.5 Calculation of the beatness

As described in the section 4.1, the periodic accents and missing beats are required to estimate the beats. Therefore, when an accent was located at the time of CAA, the accent time was thought to be correct. On the other hand, when an accent was not located on the time of CAA, it was thought that the missing beat should be located at the time. Both of them were obtained as the beat times. In this study, three criteria were introduced to evaluate the beatness of the accent. Here, The beatness B_{t_p} is defined as a score of beatness at the accent time t_p. Hereinafter, three criteria to obtain scores of the beatness B_{t_p} are described. Also, Fig.4 outlines them.

- Criterion of "Around 10 ": If the number of accents located in the range of $\pm(A/w)$ on the centre of $t_p \pm lA(l = 1,2,..,L$, i.e the number of points $2L$ points) is more than n, the proposed system adds b_1 to B_{t_p}, where n is 5, L is 5, w is 4 and b_1 is 1.
- Criterion of "Pre 10 ": If the accents located in the range of $\pm(A/w)$ on the centre of $t_p - mA(l = 1,2,..,M$, i.e the number of points $2L$ points) is more than n, the proposed system adds b_2 to B_{t_p}, where M is 5 and b_2 is 1.
- Criterion of "Post 10 ": If the accents located in the range of $\pm(A/w)$ on the centre of $t_p + mA(l = 1,2,..,M$, i.e the number of points $2L$ points) is more than n , the proposed system adds b_3 to B_{t_p}, where b_3 is 1.

Additionally, the missing beat should be estimated, as described in the section 4.1. Here, the missing beats are obtained when the accent is not located at the time of CAA.

Figure 5 shows the flowchart for estimating the missing beats, where o is 3. It is thought that the missing beats are located on the time of CAA, when the accent is not located on the time of CAA. Thus, when any accents are not located at t_{p+s} in Fig.5, a missing beat is thought to be actually located at

$t_p + A$. Therefore, the proposed system adds b_4 to B_{t_p+A} (in Fig.5 (ii), and $b_4 = 1$). The proposed system gives 0 to all the scores of beatness B_{t_p} between the accent time t_p and the next beat time t_{p+s}, because the accent at t_{p+s} nearest to $t_p + A$ was used as the next beat. Therefore, if $B_{t_p+A} > 0$ where $t_p < t_{p+s} < t_p + A \pm A/o$ (in Fig.5, (i)), the accent at t_{p+s} nearest to that at the $t_p + A$ was used as the next beat. In other words, it was thought that the interval between t_{p+1} and t_{p+s-1} does not contain any beats. Thus, the value of zero was substituted for $B_{t_{p+1}}, B_{t_{p+2}}, ..,$ and $B_{t_{p+s-1}}$. Finally, the beats that have a B_{t_p} more than zero were outputted as the beats.

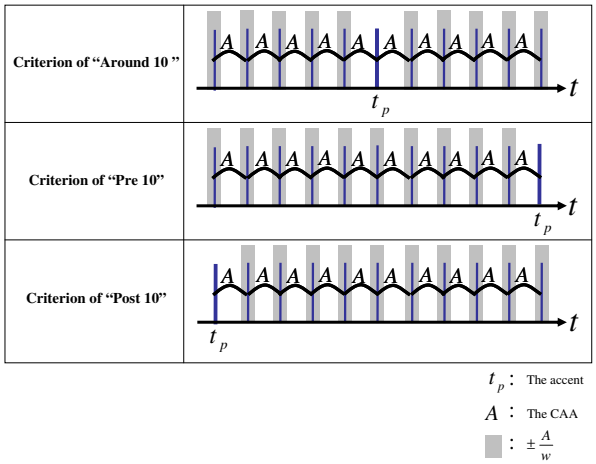

Figure 4: Outline of three criteria for obtaining beatness

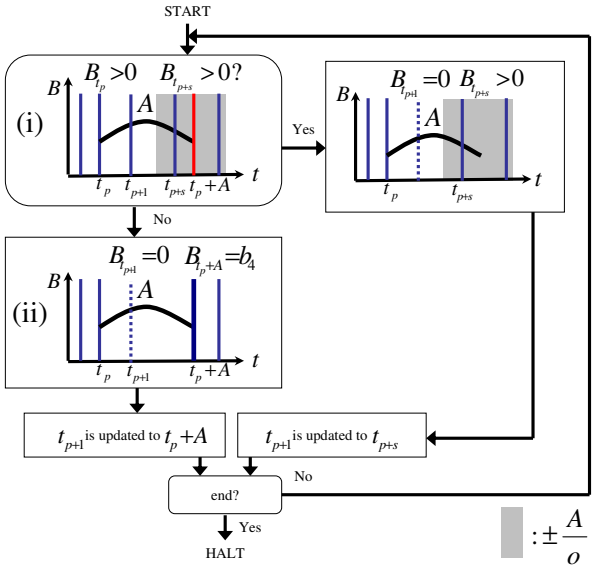

Figure 5: Flowchart for estimating missing beats

4.6 Estimation accuracy of beat

When photographs in a slideshow were switched independently to the beats in the music excerpts, the switching was thought to be inappropriate. Thus, the correctness of beats estimated by the proposed system should be indicated by evaluation experiments. Here, the accuracy of estimating the beats in the proposed system was evaluated by comparing subjective beats with all estimated beat times by means of seeing the photograph switching in a slideshow given by the proposed method. There were six subjects and 17 stimuli. The subjects were asked to evaluate whether or not they felt the switching in the slideshow was appropriate for beats in the music excerpt. Thus, the beats were evaluated using the precision, defined as U, shown in Eq.(2), where P is the number of the beats that correspond with the beats felt by a subject, and E is the number of beats estimated by the proposed system.

$$U = \frac{P}{E} \times 100 \tag{2}$$

Table 2 shows the results of evaluating the beats in terms of the precision. The results confirmed that the beat time was accurately estimated for music in which the accent such as drums was clearly heard.

Table 2: Results of beat evaluation

ID	Precision of each subject [%]					
	subject 1	subject 2	subject 3	subject 4	subject 5	subject 6
0	100.0	95.1	98.4	93.4	90.2	100.0
1	100.0	100.0	97.5	92.5	100.0	100.0
2	95.6	86.7	91.1	91.1	91.1	84.4
3	93.4	91.8	91.8	68.9	100.0	90.2
4	100.0	100.0	93.2	95.5	100.0	100.0
5	98.3	91.7	96.7	86.7	91.7	85.0
6	100.0	100.0	90.0	100.0	100.0	100.0
7	100.0	100.0	96.8	100.0	93.5	100.0
8	96.8	90.3	90.3	93.5	90.3	96.8
9	69.6	69.6	78.3	87.0	47.8	87.0
10	88.9	86.1	88.9	63.9	55.6	88.9
11	83.3	83.3	83.3	83.3	66.7	75.0
12	92.9	78.6	85.7	85.7	75.0	85.7
13	97.8	91.1	93.3	80.0	88.9	88.9
14	93.1	82.8	86.2	86.2	86.2	75.9
15	92.7	90.9	61.8	85.5	92.7	87.3
16	87.2	84.6	92.3	71.8	87.2	84.6
17	82.2	88.9	88.9	93.3	86.7	86.7
Ave.	**92.9**	**89.5**	**89.1**	**86.6**	**85.8**	**89.8**

5. EXTRACTING THE BEGINNING TIMES OF MUSICAL PHRASE

5.1 Method for extracting the beginning times of musical phrase

Section 3 described how the proposed system changed the rate of photograph switching by observing the mean value of the flux in the phrase section. To obtain the phrase section, the beginning times of musical phrases should be obtained by the proposed system. Section 2.3 described how the timing of significant changes in the flux could introduce the beginning times of musical phrases. In this study, the beginning times of musical phrase were obtained by observing large gradient times of the amplitude envelope in the time function of flux.

5.2 Procedure for extracting the beginning times of musical phrase

Figure 6 shows the flowchart for extracting the beginning times of musical phrase. Firstly, a frequency analysis was conducted by STFT (Short-Time Fourier Transform) for acoustic signal x_i to obtain a power spectrum $S_h[f]$ at each time h. Secondly, the absolute value of the spectral difference between the current time h and the previous time h_p at the powers in each frequency was calculated to obtain the value of flux F_h. F_h is defined as the time function of flux. Equation (3) shows I is the size of input signal, $h(h = 0, 1, 2, .., H)$ is the time of flux, H is the size of time function of flux, $S_h[f]$ is the power spectrum on the time h, and $S_{h_p}[f]$ is the power spectrum of the previous time h_p.

$$F_h = \|S_h[f] - S_{h_p}[f]\| \tag{3}$$

Thirdly, the amplitude envelope F_h' was obtained by calculating the simple moving average (in Fig.6 (i)) for 101 samples (i.e. approximately 587 msec), and then repeating it twice. Fourthly, a regression line was obtained using the least-square method from F_h' around 300 samples (i.e. approximately 1744 msec), so that a time function K_h of the gradient in regression line was obtained, which shows local tendencies for 300 samples. Fifthly, the time function K_h' was obtained by calculating full-wave resolution (in Fig.6 (ii)). Finally, the times of the top g values of the gradient was outputted as the beginning of musical phrases (in Fig.6 (iii)), where g is set to 10.

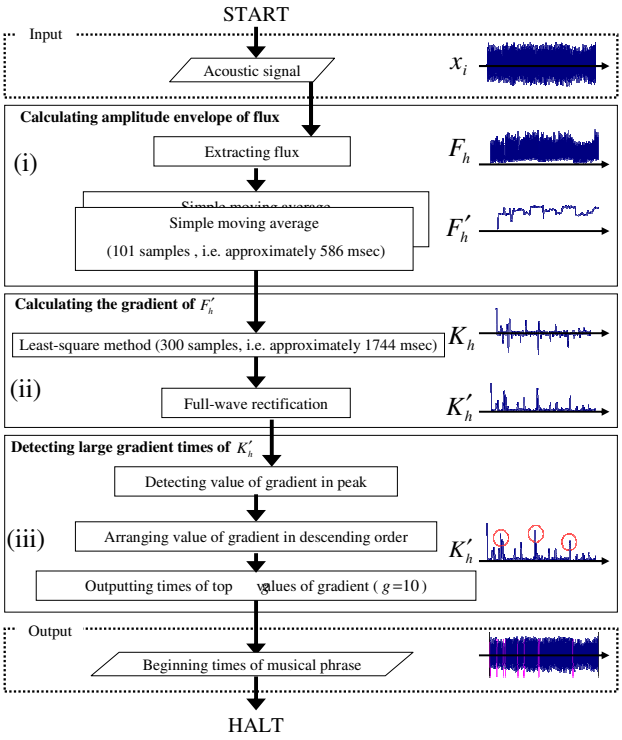

Figure 6: Flowchart for extracting beginning times of musical phrase

6. DETERMINING THE TIMES TO SWITCH PHOTOGRAPHS

Section 3 described how the photographs are switched to the beat of the music excerpt.Additionally, the photograph-switching rate should be changed by observing the value of flux in the phrase section in order to synchronize it with the exciting feeling in the music and slideshow. Therefore, the value of flux in the phrase section was calculated by observing the beginning times of musical phrases and the time function of flux, so the rate of photograph switching was determined by the mean value of flux in each phrase section. Specifically, the number of switched photographs is greater when the value of flux is higher than its mean value. After this, the photograph-switching times are required.

Figure 7 shows the flowchart for determining the times for switching photographs. In Fig.7, the first time to switch photographs in the musical phrase was shifted to the cycle of beat times. Secondly, the photograph-switching rate was determined by observing the mean value of flux in the phrase section. Finally, the times for switching photographs was determined by photograph-switching rate synchronized to the beats of the music excerpt, and these times were outputted by the proposed system.

If a photograph was switched at the beginning of a musical phrase, the switching is inappropriate, because the beginning time of a phrase is not always synchronized to the beats. Therefore, to first determine the times to switch photographs, the first time of switching in the musical phrase were shifted or quantized to the cycle of beat times (in Fig.7 (i)), where the quantization in this study means that the first time to switch photographs is synchronized to the beat times. Secondly, the mean value of the flux in the phrase section was obtained to determine the photograph-switching rate. When the mean value of flux in the phrase section was higher than the mean value of flux in the music excerpts, the number of switched photographs becomes greater at the phrase section, so that the photograph-switching rate was changed. When the number of switched photographs was greater in a phrase section, the section was called a "quick section" and the other sections "slow sections" (in Fig.7 (ii)). Finally, the number of switched photographs was determined in all the quick and the slow sections (in Fig.7 (iii)). The number of switched photographs in the quick section was defined as one every beat. On the other hand, the number of switched photographs in the slow section was defined as one every eight beats. Because the length of the phrase section is not limited, a broken number appears among the slow section when the beat times cannot be divided evenly by eight.

Figure 8 shows an example of processing the broken number, represented as b. If $b/8$ was more than 0.5, b was adapted as the switching interval just behind the beginning of musical phrase. On the other hand, if $b/8$ was less than 0.5, b was added to the switching interval just after the beginning of musical phrase. For example, if the 18 beats exist in a phrase section, a photograph is switched for the first time at the first beat time. The next photograph is switched at the beat time of the number 8 + 2, because b is $2(18 = 8 \times 2 + 2)$, and 2/8 is less than 0.5. On the other hand, if the 22 beats exist in a phrase section, the second photograph is switched to another at the beat time of the number 6, because b is $6(18 = 8 \times 2 + 6)$, and 6/8 is larger than 0.5.

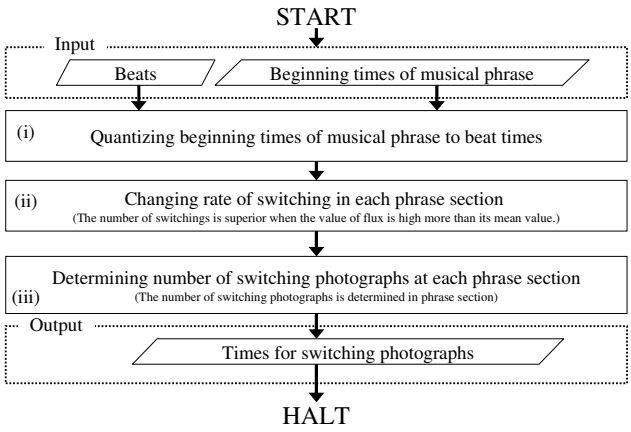

Figure 7: Flowchart for determining times for switching photographs

7. EVALUATION OF SLIDESHOW

The photographs of a slideshow generated by the proposed method were switched to the beat. Additionally, the rate of switching photographs was changed by the mean value of flux. In other words, the slideshow satisfies the definition of the appropriate photograph switching described in chapter 2. However, the slideshow satisfying the criteria has not been confirmed to actually make people feel the motion and music are congruent. Therefore, the feeling of motion and music congruency in the slideshow produced by proposed system was evaluated in an evaluation experiment by eight subjects. In this

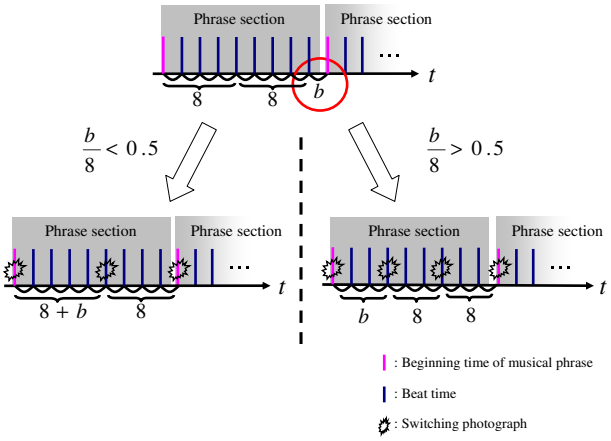

Figure 8: Example of processing broken number *b*

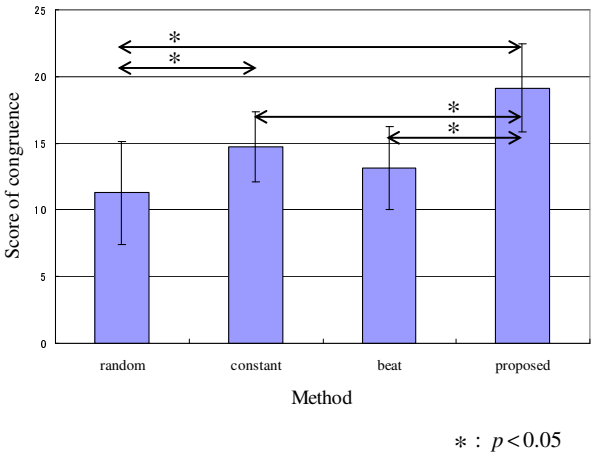

Figure 9: Results of evaluation experiment of slideshow for all music excerpts

evaluation experiment, the slideshows under four photograph-switching methods were ranked in order of the strength of feeling of image and music congruency. Specifically, the method evaluated to have the strongest congruency was given 4 points, and other methods were given 3, 2, or 1 point in descending order of the strength of congruency. The four methods were as follows:

- Method 1: Switching photographs at random time intervals. (random)
- Method 2: Switching photographs at periodic time intervals. (constant)
- Method 3: Switching photographs at beat times. (beat)
- Method 4: Controlling switching time intervals in method 3 in each phrase section of musical phrases (proposed).

Methods 1 and 2 were realized by using the commercially sold photo-frame systems [6]. Method 3 switched photographs at just the beat times estimated by the proposed method. In slideshows produced by Method 3, all the accents of a motion image such as switching a photograph were always matched to the accent of music such as the beat time. Thus, the slideshow produced by Method 3 was expected to give the feeling of congruency. Method 4 was the proposed method of switching photographs. If a slideshow produced by Method 4 gave stronger feelings of congruency than the other methods, it was thought that the proposed system could produce the slideshow with stronger feelings of congruency than other existing methods.

Figure 9 shows the results of evaluation experiments of a slideshow for all music excerpts. In this evaluation, it is assumed that the differences among subjects can be ignored. Factors of the method and music excerpt were analyzed using two-way ANOVA between them. The results confirmed a significant difference in the main effect in the method. On the other hand, they did not confirm significant difference in the main effect in the music. In other words, effects on evaluation were not confirmed due to differences in music but were for differences in method. Here, a significant difference of interaction effect between the method factor and music factor was confirmed, where the probability of a significant difference was less than .05, so tests for significant differences were conducted for each musical excerpt, where the Freedman test was used. The results confirmed a significant difference between the proposed method and other methods in four cases out of six. Thus, the method factor of the four cases was tested by Steel-Dwass test for multiple comparisons. Finally, it was confirmed that the slideshow produced by the proposed system gave stronger feelings of congruency than other methods. Therefore, the proposed system was confirmed to automatically produce a slideshow with stronger feelings of congruency than existing systems.

8. CONCLUSION

In this study, the appropriate switching of photographs was defined on the basis of a previous study. Additionally, the musical features were compared with cartoon films that contain frequent cutting or objects in rapid motion to investigate the musical feature required to determine the photograph-switching rate. Thus, a method for switching photographs was proposed for the system by defining the appropriate switching of photographs and the results of investigation. The proposed method was implemented to the proposed system to automatically produce a slideshow. Firstly, the proposed system extracted the beats and the beginning times of the musical phrase from input acoustic signals. Secondly, the times for switching photographs was automatically determined using the beats and the beginning times of the musical phrase. Finally, photographs were switched at the times for switching photographs. In future works, we will consider how to determine the visualizing effect and handle tiny fluctuations in between beats.

REFERENCES

[1] Xian-Sheng Hua et al., "Content based photography slideshow with incidental music", International Symposium on Circuits and Systems, vol.2, pp.648-651, (2003).
[2] Shin-ichiro Iwamiya and Hanako Ozaki, "Formal congruency between image patterns and pitch patterns", ICMPC8 International Conference on Music Perception and Cognition, Evanston, IL 2004, pp.145-148, (2004).
[3] Yoshimori Sugano and Shin-ichiro Iwamiya, "Effects of synchronization between musical rhythm and visual motion on the congruency of music and motion picture" Journal of Music Perception and Cognition, Vol.5, No.1, 1-10, (1999, in Japanese).
[4] George Tzanetakis at al, "Musical genre classification of audio signals", IEEE Trans. Speech Audio Process, 10, pp.293-302, (2002).
[5] Peter Grosche et al, "A mid-level representation for capturing dominant tempo and pulse information in music recordings", 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp.189-194, (2009).
[6] Samsung Electronics Co., "Digital photoframe SPF-86P", (2008).