sfa
EAA European Acoustics Association
eaa
euronoise
Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Detecting Scenes in Lifelog Videos based on Probabilistic Models of Audio data

K. Yamano and K. Itou

Hosei University, 3-7-2 Kajino-cho, 184-8584 Koganei, Japan
n04k1035@cis.k.hosei.ac.jp

Life-log videos must be detected every scene to use them effectively. Scene are detected by colors changing, however, only using color cannot obtain enough accuracy. This paper proposes a detecting method using audio data and using power spectrums and its envelopes as features. Distinction experimentations were carried out with the data recorded in railway stations. The average distinction rates were 39.3% in the pattern distance using average power spectrums, 35.0% in the pattern distance using average power spectrum envelopes, 67.9% in the probabilistic models using seven shots and 86.3% the probabilistic models using three shots. In addition, detection experimentations were carried out using actual data. The average precision was 75.9%, and the average recall was 75.2%.

1 Introduction

Life-log is a personal experience and/or everyday life record captured by wearable devices such as cameras, microphones, and GPS signals [1, 2]. Life-log video data are enormous, heterogeneous, and redundant, thus have little structure such as scene cuts and camera angles. Therefore, life-log indexing has wrestled with many of the problems [3].

Automatic indexing of life-log videos have been done based on color histogram changing. Such methods are effective to detect the scenes in artificial videos such as TV programs. However, in life-log videos, it sometimes happens that unexpected objects come in sight, such as a people/car/bicycle crossing in front of the life-log user. Although short time objects like people or cars can be easily handled as errors, long time impediments like passing/arriving trains or may occur incorrect detection. It is effective to combine other information such as an acceleration In order to index unedited videos, unsupervised clustering method using HMM was proposed [4], however, the method was not evaluated how to use in life-log applications.

In this paper, we propose a scene detecting method using audio data; the method is designed to a specific situation, which restricted by image or other sensor data. The proposed method is examined using “waiting trains ’’ scene as an example, which is one of the difficult scenes to be treated correctly by only using image data.

2 Sound Shot Detecting for Life-Log Video Indexing

2.1 “ Waiting Trains ” Scene Detecting Task

If someone goes to a railway station in order to take a train, he/she is waiting a train on a platform until the train comes, then he/she gets on the train. In our life-log application, this situation is expected to be indexed as two scenes, “ waiting trains ” and “ in train ”. However, various ambient noises can be found in audio data from the “ waiting trains ” scene, such as noises from passing/arriving/departing trains and station attendant ’s announcement. We defined such sound events as a “shot”. By listening in actual data, six major types of shots were found in the “ waiting trains ” scene;

- waiting on a platform only with minor ambient noises (hereinafter called WP)
- passing trains (called PT)
- departing trains (called D)
- arriving trains (called A)

In case of departing and arriving, sounds differ from a user position in the platform. Two typical cases, at the front-end of the platform and the rear-end of the platform, are defined in A and D;

- departing trains at the front-end of the platform (called DF)
- departing trains at the rear-end of the platform (called DR)
- arriving trains at the front-end of the platform (called AF)
- arriving trains at the rear-end of the platform (called AR)

“In train” scene is called TR.

Figure 1 is the example of scenes and shots. “Train waiting ’’ scene includes one or more shot sequence of the above shots except TR and ends with AF or AR.

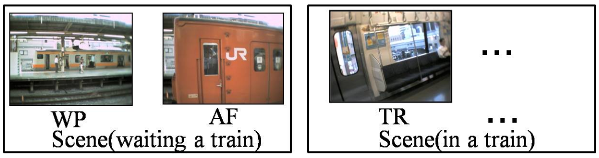

Figure 1: example of scenes and shots

In our purpose, it is required to detect correctly WP and TR, and it does not matter to detect incorrectly among the other five shots. Thus, DF, DR, AF, AR, and PT can be merged into single shot; which is called DA.

In this paper, models and experimentations were carried out with classification using both three shots and seven shots. The results of the experimentations were compared in point of using three or seven shots.

2.2 Feature Extraction

The short time spectrums of the sounds of each shot are analyzed by mel scaled filter bank to extract each shot features. Triangular windows are also used in the filter bank analysis. Because the mel frequency is near the sensory scale of human, the mel scaled filter bank analysis is the proper method for detection of the sounds that human can find difference by listening. Table 1 shows conditions of the short time spectrum and the filter bank analysis.

Table 1: Conditions of the short time spectrum and the filter bank analysis

Short time spectrum	
Length of data	2048 samples(42.7ms)
Shift of time	1024 samples(21.3ms)
Length of FFT	2048 samples
Filter bank analysis	
Length of triangular window	200 on mel frequency
Shift of frequency	100 on mel frequency
Filter degree	38

The power spectrum envelope is obtained by filter bank analysis. Figure 2 shows comparison of the average power spectrum envelope and the average power spectrum.

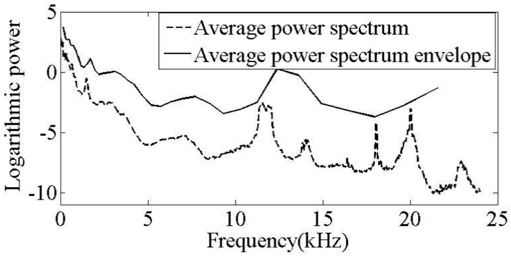

Figure 2: The average power spectrum envelope and the average power spectrum

The power spectrum envelope is integrated features of frequency more than the power spectrum.

2.3 Modeling Sound Shot

The sounds of shots are compared with each shot prototype for detecting the scene. Thus, each shot prototype is extracted beforehand. The comparing is carried out by the filter bank output. Therefore, the prototype is the average power spectrum envelopes obtained from some same shots. Figure 3 and Figure 4 show prototypes of each shot.

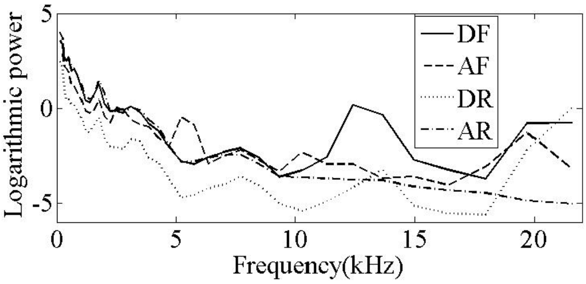

Figure 3: The average power spectrum envelopes of shots (1)

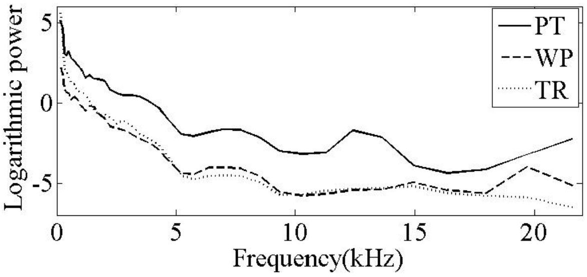

Figure 4: The average power spectrum envelopes of shots (2)

This paper also compares prototypes of the average power spectrums (Figure 5 and Figure 6) with the average power spectrums of the shot inputted.

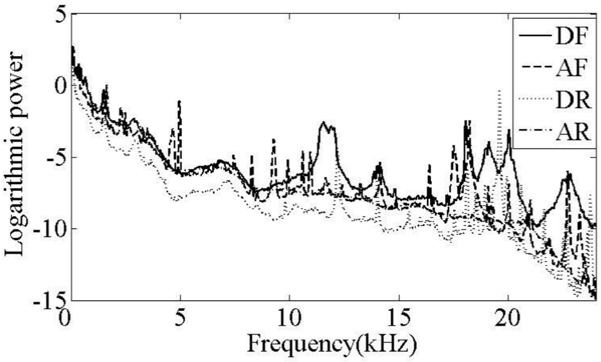

Figure 5: The average power spectrum of shots (1)

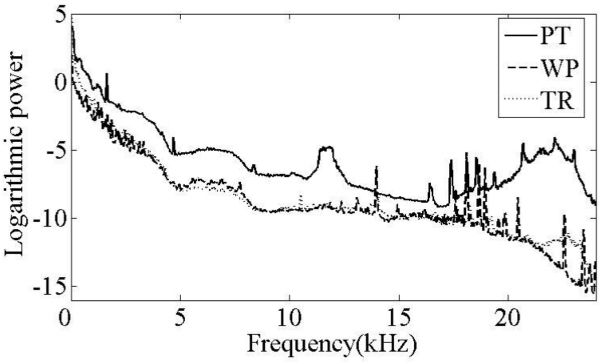

Figure 6: The average power spectrum of shots (2)

The training data used for obtaining the prototypes is extracted from 5 hours audio data recorded by the method mentioned in Chapter 3.

Pattern distance between the shot prototypes and the short time spectrum of the input is defined by Eq. (1). $SC_i(f)$ is the shot prototypes. $x(f)$ is the average filter bank output of the input. The hypothesis with the least distance is picked up.

$$\text{shot} = \underset{i}{\text{argmin}} \left[\int |SC_i(f) - x(f)| df \right] \tag{1}$$

Probabilistic models are estimated by the average and the covariance obtained by logarithmic normal distribution of filter bank outputs of training data; the order of the filter bank is thirty-eight. The hypothesis with the best likelihood is picked up (Eq. (2)). $p(x|SC_i)$ is the likelihood. x is the logarithmic average filter bank outputs of the short time spectrums obtained from an input. SC_i is the probabilistic models of each shot classified by i.

$$\text{shot} = \underset{i}{\text{argmax}} \left[p(x|SC_i)\right] \tag{2}$$

3 Sound Shot Identification

3.1 Data Collection

The audio data for training and evaluation was recorded at two railway stations and in trains between them. The time slot of recording is from 10:00 to 16:00, however most of the data was recorded from 11:00 to 13:00. The data is recorded at both ends of the platform and looking toward the opposite platform. A microphone and a recorder are the binaural microphone (adphox BME-200) and the PCM recorder (EDIROL R-09). Sampling frequency is 48 kHz, quantization bit rate is 24 bits.

The procedures of recording the data are recording in front of the platform at the first station for fifty minutes. Next is getting on a train for the next station to record the sounds in the train. After getting off the train at the next station, the sounds at platform are recorded again in rear of the platform for fifty minutes. And then, the train for the first station is got on. In the same way, the sounds in rear of the platform at the first station and in front of the platform at the next station are recorded after getting off the train for the first station. The total time of recording is eleven hours.

The data recorded like this procedures has large power in low frequency and small power in high frequency. Therefore, the sums in high frequency are too small if the sums are obtained by regular filter density on the normal frequency axis. On the other hand, the methods in this study obtained the sums by mel scaled filter bank. Thus, the sums of narrow bands are obtained in low frequency. Moreover, since the sums of wide bands are obtained in high frequency, the sums aren't too small in high frequency.

3.2 Experimentation

To evaluate the four methods that pattern distance (by the average power spectrums and the envelopes) and probabilistic models (by seven shots and three shots), the experimentations were carried out. The test data for these experimentations is the sounds of each shot outside of the training data. Table 2 shows the average time and the total number of the training data for the prototypes and probabilistic models.

Table 2: Number of training data (center column) and average time (bottom column)

	AF	DF	DR	AR	PT	WP	TR
Number	21	19	25	30	24	74	20
Time(sec)	25	13	13	24	17	53	130

Table 3 shows the average time and the total number of the shots of the test data.

Table 3: Number of test data (center column) and average time (bottom column)

	DF	AF	DR	AR	PT	WP	TR
Number	16	11	10	13	10	36	8
Time(sec)	22	11	12	24	16	42	123

3.3 Results

Distinction rates were calculated from the results of the experimentations (Table 4 and Table 5). The distinction rates were the rate of distinguishing a shot correctly. Table 4 shows the results in case of seven shots classifications, and table 5 shows the results in case of three shots classifications.

Table 4: Distinction rate (seven shots)

	Distinction rate (%)		
	Average power spectrum envelope	Average power spectrum	Probabilistic model
DF	0	6.3	31.3
AF	0	9.1	0
DR	90.0	90.0	80.0
AR	15.4	7.7	100
PT	30.0	30.0	100
WP	22.2	44.4	63.9
TR	87.5	87.5	100

Table 5: Distinction rate (three shots)

	Distinction rate (%) Probabilistic model
WP	72.2
TR	100
DA	86.7

3.4 Discussion

The methods by pattern distance were low rates. These methods distinguish shots by only the average power, although the power of D, A, and PT changes delicately by the speed of trains. Thus, the shot having the similar prototype to other was distinguished incorrectly. Moreover, since the main sounds from trains are wind noises and motor noises, these sounds are changed by speed of trains or revolution of motors. Therefore, using only average powers was not enough because the spectrum peaks change every data.

In probabilistic models by seven shots, DF and AF were low rates. DF was distinguished incorrectly as AR mainly. DF and AR have similar powers, moreover the variance of AR is lower than DF. Thus, AR has higher likelihood than DF if a input of DF has similar power to model of DF. Therefore, DF inclines distinguishing as AR. AR was distinguished incorrectly as DR and AR mainly. AF, DR and AR have similar powers, moreover, AR has higher variances AF and DR. Thus, DR and AR incline having higher likelihoods than AF. Therefore, incorrect distinction occurred.

The method by three shots was not seriously low distinction rates compared to using seven shots. However, DA inclined to be distinguished as TR and WP. Moreover, WP inclined to be distinguished as TR. The test data of DA which distinguished as WP were AF and DR in case of using seven shots. These data had small powers. Moreover, the variance of WP is lower than DA. Thus, DA was distinguished as WP. The test data that DA was distinguished as TR was almost PT in case of using seven shots. By listening these data, these data had smaller volume than other PT data because trains are passing at low speed. Therefore, DA was distinguished as TR. All incorrect results in case of inputting WP were distinguished as TR. By the listening these data, these data contained noises of the train stopping. Meanwhile, the data distinguished correctly didn't contain noises like this. Therefore, these incorrect distinctions were occurred by noises from the train stopping.

Average distinction rates of four methods were 37% in pattern distance by average power spectrums, 27% in pattern distance by average power spectrum envelopes, 68% in probabilistic models by seven shots and 86% in probabilistic models by three shots. Therefore, these experimentations obtained that probabilistic models are batter than pattern distances in this shot distinction.

4 Sound Shot Detection

4.1 Experimentation

To inspect effectiveness of the probabilistic model by three shots for shot detection, the experimentation was carried out. The test data were extracted by hands from the data recorded at platforms. These data contain changing shots like "DA → WP → DA → TR". The number of the test data was 15. The average time of test data was 3 minutes 22 seconds.

The shots are detected by distinguishing the shot every regular time interval. Time interval is 20 seconds to make the interval near the average time of AR, AF, DR, DF, and PT. WP and TR were not considered because time intervals of WP are not regularly, additionally TR is too long, about 2 minutes.

The experimentations were evaluated by precision and recall. The precision was defined the rate how many shots are detected correctly in total number of detecting a shot. For example, when DA is detected ten times, the precision is 60% if six times are correct. The recall was defined the rate how many shots are detected correctly in total number of correct shots in test data. For example, when WP is detected seven times correctly, the recall is 70% if ten times WP exists in the test data. Because the shots are detected every 20 seconds interval, the shots are frequently ranged the boundary of two shots. However, it is important to prevent detecting unnecessary scenes in case of assuming the life-log system in this paper. Thus, ranging until 15 seconds was admitted.

4.2 Results

The example of the results are shown in Figure 7. The example is the data that the shots change "DA → WP → DA → WP → DA". The solid lines are intervals detected correctly. The dotted lines are intervals detected incorrectly. The names of shots written in the graphs are the correct shot of the interval.

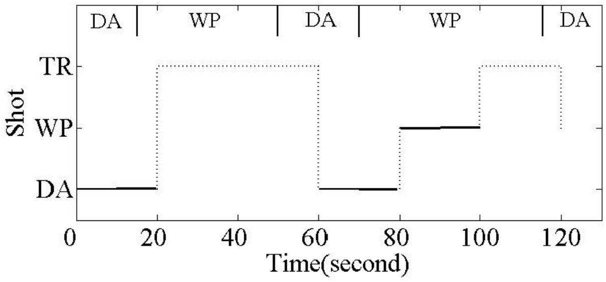

Figure 7: Example of results

The recall and the precision were obtained by above results like follows:

Table 6: Precision and recall

	Precision (%)	Recall (%)
DA	90.0	73.7
WP	89.1	54.7
TE	48.6	97.2

4.3 Discussion

The experimentations were carried out to inspect the availability of the probabilistic models. However, the unnecessary shots were detected. The distinction rates of the shots are low, however another reason is considered. The probabilistic models are estimated from the entirety of shots. On the other hand, because this experimentation applies models to regular intervals extracted from front of time series, the entirety of a shot isn't almost inputted. Thus, the intervals ranging two shots were detected incorrectly. However, the method of detection using these experimentations can't prevent from detecting incorrectly like this. Therefore, the interval distinguished needed to be estimated by a point changing sounds or images.

The shot of low precision means that the shot is detected frequently in the incorrect interval. On the other hand, the shot of low recall means that the shot is often detected incorrectly. The recall of TR was 97%, however The precision is 66%. Thus, about half was detected incorrectly in the shots detected as TR. TR has lowest variance in three, the power of three shots are similar. Therefore, the result was obtained. However, because TR is the longer time shot than other shots, TR can be treated that the shot is detected as TR if some frames are detected as TR successively. In detection experimentations, the power of input is hard to be near the power of training data because the entirety of a shot isn't almost inputted.

5 Conclusion

This paper suggests the methods of distinguishing shots to detect waiting a train scene of life-log videos. One is the method using pattern distance, another is the method using probabilistic model. The distinction experimentations were carried out by these methods. The average distinction rates are 39.3% in the pattern distance using average power spectrums, 35.0% in the pattern distance using average power spectrum envelopes, 67.9% in the probabilistic models using seven shots and 86.3% the probabilistic models using three shots. Moreover, the detection experimentations were carried out by the methods of probabilistic models of three shots. The results were evaluated by recall and precision. The average precision was 75.9%. The average recall was 75.2%. In addition, this experimentation showed unnecessary shots were detected in shot boundaries.

This paper didn 't consider conditions conspiring trains departing or arriving or passing, and trains at an opposite platform. However, these situations must be considered in case of assuming life-log system. Also, the method of shot detection is needed to be considered.

Acknowledgement

We acknowledge and thank Mr.Katsuhiro Takata for his advices and support of our study.

References

[1] Mik Lamming, Mike Flynn, "" Forget-me-not "Intimate Computing in Support of Human Memory", *Proceedings of FRIEND21, '94 International Symposium on Next Generation Human Interface*,(1994),Also available as RXRC TR 94-103, 61 Regent St., Cambridge, UK.

[2] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker and Curtis Wong,"MyLifeBits: Fulfilling the Memex Vision", *Proceedings of the tenth ACM international conference on Multimedia*, 235-238 (2002)

[3] Kiyoharu Aizawa, Tetsuro Hori, Shinya Kawasaki, Takayuki Ishikawa, "Capture and Efficient Retrieval of Life Log", *Proceedings of the pervasive 2004 workshop on memory and sharing experience*, 15-20(2004)

[4] Brian Clarkson, Alex Pentland, "UNSUPERVISED CLUSTERING OF AMBULATORY AUDIO AND VIDEO", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol.6, 3037-3040(1999)