Browsing Audio Life-log Data Using Acoustic and Location Information

Kiichiro Yamano
Graduate School of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: kiichiro.yamano.rr@gs-cis.hosei.ac.jp

Katunobu Itou
Faculty of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: itou@hosei.ac.jp

***Abstract*—The use of the log of personal life experiences recorded on cameras, microphones, GPS devices, etc., is studied. A record of a person' s personal life is called as a life-log. Since the amount of data stored in a life-log system is vast and since the data may also include redundant data, methods for the retrieval and summarization of the data are required for the effective use of the life-log data. In this paper, audio life-log recorded by wearable microphones is described. The purpose of this study is classifying audio life-log according to places, speakers, and time. However, the places stored in an audio life-log are obtained by GPS devices; information about rooms in buildings cannot be obtained. In this study, experiments were carried out on audio life-log. The audio life-log was divided into segments and clustered by spectrum envelopes according to rooms. The experiments show two situations in which the location information are captured and not captured. The results of the experiments showed that the location information helped in room clustering. Audio life-log browsing on a map using GPS is also suggested.**

Keywords - lifelog; audio; GPS; clustering; browsing

I. Introduction

Many studies on the use of the log of personal life experiences recorded by devices such as cameras, microphones, and GPS loggers have been carried out [1]. Such records are called as life-logs. Life-logs are considered to play an important role in the development of multi-modal personal memorandum and in the development of an automatic diary. They are also considered to be used as dynamic personal marketing tools and personal recommendation systems that share multiple persons' life-logs. However, it is difficult to use life-logs since a vast amount of data is involved and also because the data may be redundant. Therefore, for the effective use of life-logs, it is necessary to develop methods for their retrieval and summarization. Many methods have been proposed recently.

In this paper, audio and location life-logs are addressed. The life-log was recorded by a wearable microphone and a GPS logger. Audio life-logs provide considerable information from various acoustic signals. For example, speech provides information on conversations and speakers, and other sounds such as background noise provide information on locations, activities, and contexts (noisy or quiet place). However, the data have many redundant parts that do not contain any sound or contain sounds that cannot be identified. Therefore, it is difficult to search desired parts without indexing, eliminating redundant parts, or clustering. Moreover, it is difficult to browse audio data because these data have intervals and are of various lengths. In an earlier study, audio events that correspond to locations such as library, street, and campus are extracted and they are displayed in a personal calendar [7].

In this study, we focused on clustering and browsing of multi-modal audio life-logs. Audio events that correspond to locations are extracted automatically from logs by audio data, and they are clustered by both acoustic information and GPS information. They are browsed with a timeline with the help of pop-up balloons on 2-D maps.

II. Related Work

Many studies have been carried out on the retrieval and summarization of life-logs. Aizawa proposed a system that retrieves life-log videos by obtaining retrieval keys from sensor information, such as brain waves, user accelerations, and GPS signals along with information stored in a PC, such as Web addresses and emails [2]. A Life Pod, which is a life-log system that involves the use of a mobile phone, has also been proposed [3]. Life Pod manages memos inputted by a user in addition to image and location information acquired by a camera and a GPS-enabled mobile phone. Moreover, it can obtain information on surrounding objects by using RFID tags.

Several methods for the clustering and segmentation of life-log data have been proposed for their easy retrieval. For example, color histograms of personal video archives are clustered in [4]. Video data are recorded for 62.5 h in the MPEG-4 format and labeled with 34 locations such as staircases, corridors, and office rooms corresponding to the location where the data are recorded. The data is applied TCK-means clustering in such a manner that the data recorded in near time are classified in the same cluster. Moreover, the results of TCK-means clustering are compared to those of k-means clustering. A method for the segmentation of daily events is suggested. In [5], life-log videos comprising 1785 images per day are handled. First, the sequences of these images are divided into groups. A new group is created when the boundary device begins operation after having been switched off for at least 2 h. Each group corresponds to images that were collected for an entire day. The groups are further divided into

subgroups. The color and edge information in the MPEG-7 format is used for the segmentation. The peaks of dissimilarity of two neighboring images that obtained this information are boundaries of events. The experiments performed in [5] were carried out using 271163 images captured by five users over a period of one month.

Methods of retrieval and summary for audio life-logs are also studied. In [6], location information and speech recognition for conversation data are used as memory aids in a retrieval system. However, since speech recognition from conversations has word error rate of 30% to 75%, the system aids a user in recalling past events by also presenting confidence scores of speech recognition results. In [7], audio data are used for archive user' locations, actions, conversations and people the user met. For minimizing the burden on a user, only a nondirectional microphone and GPS are used. Additionally, 62 hours audio data obtained at a library, restaurant, lecture room, meeting room, etc., are classified by using a spectral clustering algorithm. Clustering accuracy is approximately 60%.

Audio life-log expected speech is also useful in numerous applications. For example, a desk job and a meeting taking place in an office are discriminated by the occurrence rate of sounds of a page turning, keyboard typing, and speech in [8]. In [9], the scenes of life-log videos in railway stations are divided by the identification of three sounds corresponding to a waiting train, a departing/arriving/passing train, and the inside of a train. These situations are difficult to be discriminated only with image/video life-log information and then environmental sounds help to discriminate or cluster events/scenes.

III. Utilization and Processing of Audio Life-log

Audio life-log contains various sounds. We collected over 59 h of audio life-log for 11 days. It contained speech, machine sounds, background noise, broadcast sounds, warning tones, etc. The major sounds that are contained in the log are sorted into their recorded locations in Table I.

TABLE I
Main recording locations and sounds.

Locations	Sounds
Laboratory	Speech, page turning PC (mouse and keyboard) Fan (air-conditioner)
Class	Speech, Fan (air-conditioner, PC)
Hallway	Footfall, speech
Campus(Outdoors)	Speech, construction work Air duct
Home	TV, music
Video shop	Speech, music
Fast food shop	Speech
Convenience store	Speech, music
Supermarket	Speech, music
Street	Car, speech, beep tones of rail crossing

Characteristic features of acoustic information such as gain levels, frequency responses, sampling rates, and quantization bit rates are varied according to the recording device such as IC recorders and the capturing device such as microphones.

For speech processing applications such as speech recognition, it is common that the capturing and recording devices are uniformed in order to achieve high accuracy or performance. However, this assumption is not realistic for life-log applications, because life-log archives may have a longer life than recording/capturing devices. From a view point of sharing multiple life-logs, the use of a uniform device is also not realistic. Therefore, the processing of life-log should be robust in variable recording/capturing conditions.

Speech in audio life-log is useful for many applications such as a personal memorandum, and it is one of the major contents in audio life-log. In the three-hour part investigation from the above-mentioned life-log, about one-half segments (91 one-minute segments among 180 segments) contained speech. Most segments that do not contain speech were recorded in solitary situations such as operating a computer in the laboratory or at home.

For personal memorandum or diary, playback of desired parts is required. This requires retrieval or summarization for quick browsing by indexing segments or tagging/annotating on segments.

In life-logs, time stamps, speaker identification data, location information, and speech contents are considered as indexes. Speech contents are created by transcription manually or automatically. Manual transcription is costly. Automatic transcription can be done by speech recognition systems. However, the recordings in an audio life-log are difficult to recognize accurately; further, the spontaneity of speech in a life-log is also difficult to recognize.

In this study, we propose a clustering method and a browsing method of audio events using both acoustic information and GPS information. Audio events cannot be clustered accurately by only acoustic information; GPS information may improve clustering performance. GPS information also helps in browsing speech segments.

IV. Data Collection

The audio part of the life-log used in this study was recorded by three kinds of IC recorders (EDIROL R-09, EDIROL R-09HR, YAMAHA POCKETRAK CX) and a binaural microphone (Adphox BME-200). The recorder used is varied from day to day in order to investigate the effect of devices. Binaural microphones are earphone-type microphones and are normally worn on the ears. In life-log recording, since wearing microphones on the ears for a long time is a burden to a user, they are worn around the neck and positioned close to the user' s chest (Figure 1). In the experiments, two microphones were fixed at a regular distance by a wire. The users recorded sounds heard in their daily life. The sampling rate was 48 kHz and the quantization bit rates were 24 bits or 16 bits that differed among recorders.

The GPS information part of the life-log, which contains location information and time stamps, was captured at intervals of five seconds by a GPS logger (GlobalSat DG-100). In the recording session, the recording by the IC recorder and the GPS logger started simultaneously.

Fig. 1. Binaural microphone worn for recording data.

A. Utilization of Location Information

A segment of audio life-log is corresponded to a location on a map by a latitude and a longitude that are captured by the GPS device. An example of a GPS life-log is shown in Figure 2. In Figure 2, distances between a user and each place which are calculated by the GPS life-log are shown in chronological order. The example is a part of the log from a university to home. There are three convenience stores and one supermarket on the way home. Latitudes and longitudes of these six places are exported from the map data. Distances between the user and places are obtained using latitudes and longitudes of the two places. Latitudes and longitudes of each place are obtained by Geocoding[1]. Since an area of a movement of the user is sufficiently small, distances are obtained as Euclidean distance as an assumption that the area is approximated as flat surface. Time when GPS devices could not trace are interpolated by straight line approximation. This figure shows the user is in the place where is the smallest distance. Distances of Figure 2 almost correspond to the actual movement at that day. Thus, rough location information of the user can be obtained. There is a possibility that this information is useful for clustering locations by audio information in Section 5. Rough location information by GPS is used for helping clustering detailed location such as lab, hallway etc. Clustering detailed location by acoustic information is carried out after clustering rough locations by GPS information. Since locations of clustering by acoustic information are limited, clustering accuracy is counted on improve.

V. Location Clustering Using Acoustic Information

In this section, we describe a method and an experiment of clustering location of audio life-logs by acoustic features.

A. Clustering of Audio Segments

Acoustic information is clustered for obtaining location information of rooms that cannot be captured by GPS. A same room has a constant background noise. Thus, an audio life-log is divided into one-minute segments. Moreover, features extracted from the segments are clustered.

[1]Geocoding http://www.geocoding.jp/

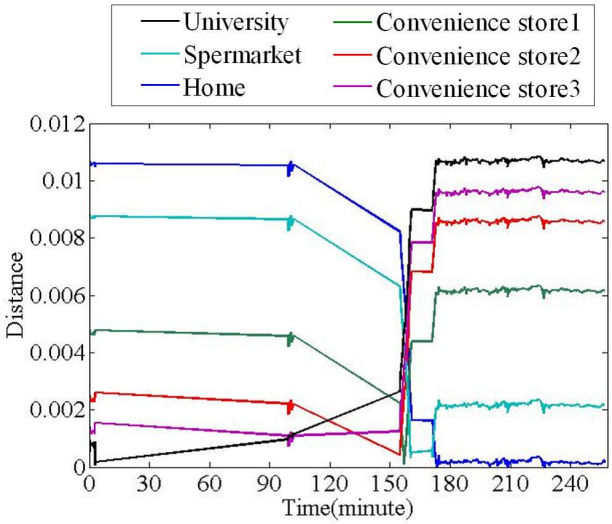

Fig. 2. Distances between the user and the six places.

In [7], an average duration of recorded segments that contain a single location and/or situation was 26 min. Because the shortest event should have a duration of 15 min, data are divided into one-minute segments and each segment is processed. In this paper, one-minute segments are also used because the locations clustered are similar to [7].

A normalized average spectrum envelope is used as a feature in this paper. The spectrum envelope is obtained by applying filter bank analysis to a short time spectrum on the mel-frequency axis. The short time spectrum is obtained by applying FFT to a wave extracted by 85.3 ms Hanning window shifting. The shift is 42.7 ms. In the filter bank analysis, a spectrum is obtained by using a fixed-length triangular window shifting on a mel-scaled frequency axis, and summations of spectrum in each band are calculated. The width of the triangular window is 600, and the shift is 300 along the mel-scaled frequency axis. The mel frequency is near an auditory scale of humans and is obtained from Equation (1) [10]. The filter bank analysis combines 2048 FFT values in 12 energy spectrum bins (Figure 3). Since several spectrum envelopes are obtained from a segment, their average is a feature of a segment. Normalization is the subtraction of the average of a spectrum envelope from the total spectrum envelope.

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

This feature is classified by k-means clustering. K-means clustering is a process given below. The variable k denotes the number of clusters.

1) k segments are randomly selected as the first centroids.
2) Euclidean distances are calculated between the features of the first centroids and those of all segments.
3) Each segment is assigned to the closest cluster.
4) The centroids of each cluster are calculated as new centroids.

TABLE II
THE RESULT OF CLUSTERING ALL DATA. THE CLUSTERS ARE LABELED BY HAND. FOR EXAMPLE, LAB CLUSTER INVOLVES 113 LABS, 1 HALLWAY, 1 OUTDOOR, 4 HOMES, 2 CONVENIENCE STORES, 2 STREETS, AND 3 SUPERMARKET SEGMENTS.

	Lab	Hallway	Outdoors	Home	Convenience store	Street	Supermarket	Precision	Recall
Lab	113	1	1	4	2	2	3	89.7%	27.0%
Hallway	37	4	1	0	1	1	1	8.9%	33.3%
Outdoors	40	2	8	0	1	13	2	12.1%	42.1%
Home	0	0	0	78	0	0	0	100%	94.0%
Convenience store	51	0	2	0	2	0	0	3.6%	33.3%
Street	67	1	0	1	0	10	0	12.7%	29.4%
Supermarket	100	4	7	0	0	8	3	2.5%	33.3%

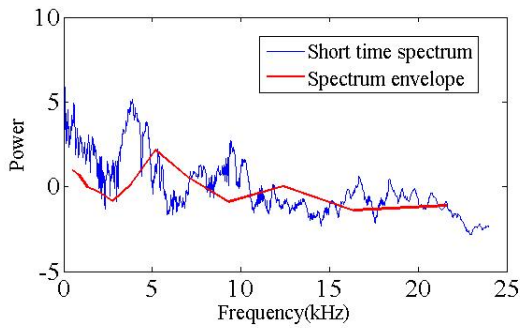

Fig. 3. Spectrum envelope and short time spectrum.

5) Euclidean distances between the centroids and all segments are calculated. Each segment belongs to a cluster of the least distance.
6) Steps 4 and 5 are repeated until the centroids do not move or until clustering has been performed for a predetermined number of times.

In this paper, the value of variable k is 7 for clustering all data and 3 for clustering the data in a university.

B. Experiment

Experiments were carried out on the clustering of places in an audio life-log. The data used for the experiments were collected over a period of two days (nine hours thirty minutes). The data for one day are recorded by YAMAHA POCKETRAK CX, and the data for the other day are recorded by EDIROL R-09HR. Locations in the data are a laboratory, a hallway, a campus (outdoors), a street, a home, a convenience store, and a supermarket. The total number of segments is 517. All segments are labeled about above locations by hand.

Two experiments are carried out using the data. One is clustering all locations on the basis of the presumption that GPS is not used, the other is clustering locations of a university on the basis of the presumption that GPS identifies the university.

C. Results of Clustering

Results of clustering are showed in Table II and III. Each cluster is labeled by hand. The results are evaluated by recall and precision.

Each row is the number of segments involved in the cluster. For example, lab cluster involves 113 labs, 1 hallway, 1 outdoor, 4 homes, 2 convenience stores, 2 streets, and 3 supermarket segments, as shown in Table II. Precision and recall of home and precision of lab were high. However, other precision and recall were low.

Table III shows the result of clustering the data in a university. The use of location information improved precisions and recalls. The precision and recall of a lab were improved by 8.4% and 24.5%, respectively. The precision and recall of outdoors were improved by 34.1% and 52.6%, respectively. The recall of hallway was improved by 16.7%. However, the precision of hallway was deteriorated by 24.5%.

TABLE III
THE RESULT OF CLUSTERING THE DATA IN A UNIVERSITY. THE CLUSTERS ARE LABELED BY HAND. EACH ROW IS THE NUMBER OF SEGMENTS INVOLVED IN THE CLUSTER AS WELL AS THE DATA IN TABLE II.

	Lab	Hallway	Outdoors	Precision	recall
Lab	210	3	1	98.1%	51.5%
Hallway	180	6	0	3.2%	50.0%
Outdoors	18	3	18	46.2%	94.7%

D. Discussion

In lab and outdoors clusters, capturing location information improved recall and precision. Since lab and outdoors clusters are confused with the supermarket cluster, capturing location information makes it possible for improving the accuracy of clustering. Although the recall of hallway is also better by using the location information, the precision of hallway is deteriorated by using the location information. A percent of the lab segments that are populous and confused with other segments is higher by confining the location to a university. In the data of these experiments, since supermarket, convenience store, home, and street clusters are not classified according to rooms, these locations captured by GPS devices are used as a cluster. Therefore, the location information can complement the accuracy of clustering by acoustic information. Although the location information can be captured, a user often moves in the street. Thus, clustering the street segments is considered. The data of this study involves a situation in which a user walks; however, this situation lasts for approximately ten minutes. In this case, classifying motion as one cluster is not a problem. Clustering methods in a case of street segments involved considering long time motions.

It would appear that a causality of deploying the lab segments is that the sounds recorded in the laboratory depend on different situations . The sounds recorded in the laboratory are shown in Table I. In these sounds, acoustic characteristics of segments involving speech entirely differ from segments not involving speech. Conversations are often recorded for several minutes. Thus, the feature of segments involving conversations differs from the one not involving conversations. Especially, since the speech of a user recording is loud, a spectrum is significantly affected. For an application of an audio life-log, classifying segments involving conversations or not by this distinction of acoustic features may be effective. A cluster does not almost change other clusters after another at one-minute interval. Therefore, lab segments may be classified into one cluster by [4] methods in such a manner that the data recorded in near time are classified in the same cluster.

In this experiment, two days data recorded by a different IC recorder was used. To use a normalized feature, clustering was not affected by recorders. This is confirmed from outdoor clusters nearly classified into one cluster shown in Table III. However, acoustic features may change if conditions of weather or air conditioning are different. Experiments using data recorded over a long duration is required to verify effects by variations in these conditions.

VI. Method of Data Presentation

In this section, presentation of audio life-logs are described. We propose presenting speech part of audio life-logs on 2-D map.

A. Speech Presentation on 2-D Map

A browsing system assumed as a memorandum application of an audio life-log is suggested. This system presents speech data classified by times, speakers, and locations on a 2-D map. Requests of information presentation from a user are as given below.

1) The conversation with A in the lab on May 1st.
2) The conversation with A in the evening (date unknown).
3) The conversation with A and B on May 1st.

Clustering according to rooms is useful for these requests. About clustering speakers, one-minute segments often involved several speakers. Thus, shorter segments should be used for clustering. Ideally, segments should not be fixed-length but flexible-length that extracted speech parts. To index segments after clustering by these processes, speech of A on May first can be presented for request 1. For request 2, speech of A at evening can be presented. For request 3, desired information can be presented by searching parts of speech of A, B, and the user appearing very often by an additional process.

B. Example of Audio Life-log Browsing

An example of browsing an audio life-log by times, locations, rooms, and speakers is shown in Figure 4. Google Maps API[2] is used in this system. First, a user selects a marker of location captured by GPS. When the user selects a room, the data of tree structure is displayed in the left part. When the user selects the date, the speakers who present at the day are displayed. Moreover, the speech of each speaker is displayed in time series. Speech is played to obtain a time label.

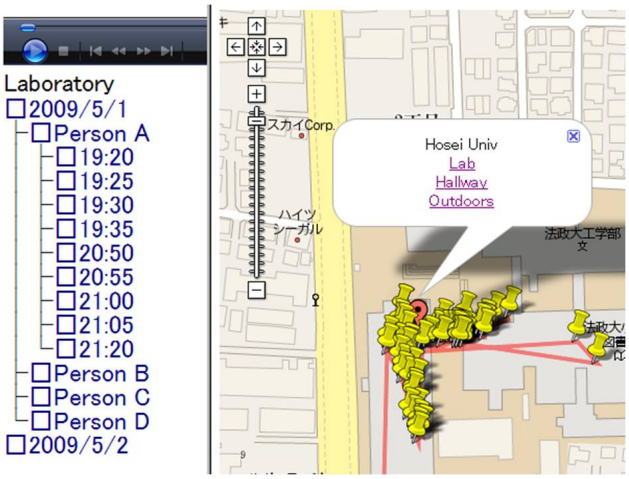

Fig. 4. Audio life-log browsing.

VII. Conclusion

In this paper, a method of information presentation using time, speaker, and location information was suggested as an effective way of using audio life-logs. Clustering location using acoustic information was also suggested as a method of capturing location information in a building. For evaluating proposal methods, audio life-log for two days was divided into one-minute segments, and the segments were clustered by a spectrum envelope. Experiments were assumed for two situations. One was a situation in which the location information is used; another one was a situation in which the location information is not used. As a result, the accuracy of clustering was improved by the location information in this experiment. However, experiments using long duration data are required.

The experiment in this study was assumed such that the location information by a GPS device does not involve errors. However, the location information by a GPS device sometimes involves errors of a few meters to a dozen meters. Although the information rarely involved errors up to a few kilometers, such errors are not consecutive. Thus, large errors can be eliminated. For identifying the location and building by GPS devices, the places that GPS signal break up or observation points converge are important. Experiments for identifying the location by GPS information must be carried out in the future. Since GPS devices rarely could not receive signal outdoors, the frequencies of this phenomenon must be researched.

Appropriate features for speaker and location clustering are also considered. Although k-means clustering was used in this study, the number of cluster is unexplained in the actual data. Therefore, clustering methods deciding the number of cluster automatically are considered.

[2]Google Maps API http://code.google.com/intl/ja/apis/maps/

REFERENCES

[1] J. Gemmell, G. Bell and R. Lueder, "MyLifeBits: A PERSONAL DATABASE EVERYTHING", *COMMUNICATIONS OF THE ACM*, Vol.49, No.1, pp.88-95, Jan. 2006

[2] K. Aizawa, "Digitizing Personal Experiences: Capture and Retrieval of Life Log", *Proceedings of the 11th International Multimedia Modelling Conference*, pp.10-15, Jan. 2005

[3] A. Minamikawa, N. Kotsuka, M. Honjo, D. Morikawa, S. Nishiyama and M. Ohashi, "RFID Supplement for Mobile-Based Life Log System", *Proceedings of SAINTW'07*, pp.50-50, Jan. 2007

[4] WH. LIN and A. HAUPTMANN, "Structuring Continuous Video Recordings of Everyday Life Using Time-Constrained Clustering", *SPIE Symposium on Electronic Imaging,*, Jan. 2006

[5] AR. Doherty and AF. Smeaton, "Automatically Segmenting Life-Log Data into Events", *In WIAMIS 2008*, pp.20-23, May 2008

[6] S. Vemuri, C. Schmandt, W. Bender, S. Tellex and B. Lassey, "An Audio-Based Personal Memory Aid", *Ubicomp 2004*, Vol.3205, pp.400-417, Oct. 2004

[7] DPW. Ellis and K. Lee, "Minimal-impact audio-based personal archives", *CARPE'04*, pp.39-47, Oct. 2004

[8] S. Shimura, Y. Hirano, S Kajita and K Mase, "Experience Movie Presentation Method Using Action Situation Query" ,*Proc.68th National Convention of IPSJ*, pp.4_81-4_82, 2006, (in Japanese)

[9] K. Yamano and K. Itou, "Detecting Scenes in Lifelog Videos based on Probabilistic Models of Audio data", *Acoustics08*, Jul. 2008

[10] D. O' Shaughnessy, "Speech Communication: Human and Machine", Reading, MA: Addison Wesley, 1987.