Repeat Recognition for Environmental Sounds

Yuya Hattori, Kazushi Ishihara, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku,
Kyoto 606–8501, JAPAN
E-mail {yuya, ishihara, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

This paper focuses on recognition of repeats in continuous environmental sounds. Environmental sounds, that are very informative in our daily lives, often repeat several times. Repeat recognition for environmental sounds is essential for compact representation of those sounds and for forecasting the future. In our method, input environmental sound signal is partitioned into several units according to the shapes of the power envelopes, and the auditory distance between every pair of units is computed. Repeating parts are then detected by using the approximate matching algorithm. Experimental results showed the 73.3% of recognition rate in counting the repeats of 30 environmental sounds, the length of each of which ranged from 20 seconds to 2 minutes.

1 Introduction

Auditory signals are one of our most important sources of information. We obtain much information from not only voices but also non-verbal sounds, such as a railroad crossing alarm or the cry of an animal. These various kinds of non-verbal sounds in our daily lives, called "environmental sounds," are important clues to understand the surroundings. However, they have been little studied except as noises interfering with speech recognition. In recent years, as the development of robots which can behave in the real world, several studies on recognition of environmental sounds appeared [1] [2] [3] [4]. These studies mainly focused on recognition of sound sources. The kinds of sound sources are indeed essential clues for robots to understand their environments. Additionally the temporal structure of the sounds has to be apprehended as well as the kind of the sound sources because there are many kinds of sounds with structures.

This paper presents a method for recognizing repeats in continuous environmental sounds. Repeat is the most typical structure of environmental sounds. A railroad crossing alarm, for example, may continue for several minutes. Many environmental sounds in the real world repeat several times so that entire information can be represented by one of the repeating parts. When we communicate incidents using sound imitation word, we often omit the repeating parts except a few times. In human-robot interaction, the method is useful for enabling a sound-to-onomatopoeia transformation system [5] to avoid redundant verbatim expressions.

Repeat recognition for environmental sounds also enables robots to forecast the future from auditory signals and to select their actions: robots have little necessity of reaction when hearing the repeating sound again; even if the sound source moves into a blind spot, it is expected to continue working as long as the same sound continues; in collaborative tasks between robots and humans where exact timing is important, repeat recognition is the most important clues to plan the next action.

For repeat recognition, environmental sounds have the following problems to be solved:

1. Environmental sounds are so various and changeful that they are hard to learn previously.
2. The environmental sounds are not regular in time.
3. For repeats with structure, some errors should be allowed. For example, we can recognize the character sequence "ABCDABBCDABCD" as three repetitions with one error. Similarly, we recognize repeats of environmental sounds with a few errors.

Problem 1 means that statistical methods which are dominantly used in speech recognition are not applicable to environmental sounds. Existing algorithms for recognizing repeats in character strings (including DNA sequences) or music cannot be applied to environmental sounds. The algorithms for recognizing repeats in character strings depend on the simplicity of the problem. Though repeats of music are

more complex, algorithms for recognizing repeats in music, which take advantage of the orderliness of music, are hard to apply to environmental sounds because of Problem 2. Since problem 3 is common with algorithm for character strings or music, we can take use of their acquaintances. In consideration of problem 1–3, we designed a repeat recognition algorithm suitable for environmental sounds.

The remainder of this paper is split into four sections. Section 2 of this paper describes our basic approach to repeat recognition for environmental sounds. Section 3 explains the entire process of our method in detail. Section 4 reports the results of evaluation experiments, and section 5 concludes the paper by briefly summarizing it.

2 Repeat recognition for environmental sounds

Although there are repeat recognition algorithms for character string or music, repeat recognition algorithms applicable to environmental sounds have not yet been developed because of the problems mentioned in Section 1. We thus need to design the followings:

1. unit of the repeat
2. distance between units
3. algorithm to detect repeating parts

in order that they solve the problems. In this section, we design each of them.

2.1 Definition of the unit of the repeat

Repeats of environmental sounds do not occur regularly in time as the repeats of music do. For environmental sounds, it is more important that the same sound (or a sound perceived to be the same) has been heard than that sounds are heard at a constant rhythm. We define a unit of repeats of environmental sounds so that each unit corresponds to a human beings' perception of "sounding one time." This, however, is a psychoacoustic concept and thus hard to use in computing directly. To extract a unit, we therefore take advantage of a physical correlate of the concept: a peak in power envelope. The extraction procedure is described in detail in Section 3.1.

2.2 Definition of distance between units

Distances of every pair of the units are computed by dynamic time warping (DTW) with MFCCs. MFCCs, which are made to adapt to the human perception, are reported to be the most effective feature for recognition of environmental sounds [4]. Although the shapes of the collision sounds of simple substances can be almost the same in every instance, the shapes of the sounds of complex substances vary from instance to instance. For example, Fig.1 shows the shapes of the sounds of two water drops. Although the shapes of these two units are roughly similar in damping with time, they differ in details from each other. DTW absorbs the difference by aligning power modulation in time.

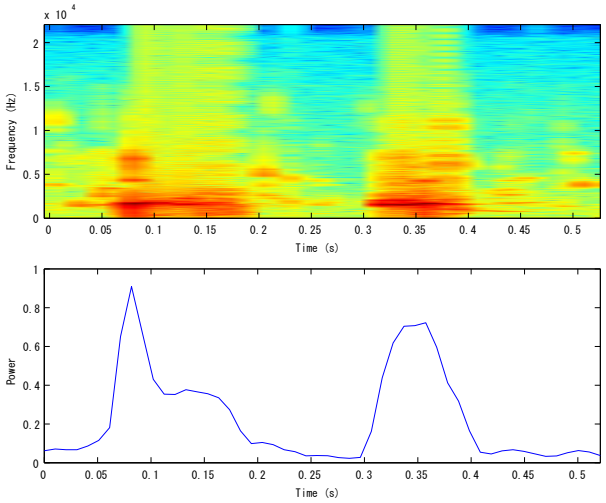

Figure 1: Sounds of two pebbles dropping into water: they sound similar but the shapes of their power envelopes differ.

Repeating environmental sounds are composed not only of single units but also of multiple units. For example "cuckoo, cuckoo;" it is divided into four units, and the parts composed of two units repeats two times. Such repeats are detected from sequence of units by the algorithm represented in Section 2.3.

2.3 Algorithm detecting repeating parts

Shamoto *et al.* developed a birdsong search system which allows search by human vocal mimicry [6]. Their experimental results indicated that the number of voiced segments is not conveyed correctly if it is more than four. This means that people can recognize a pattern to be the repeat of another pattern that actually has a different number of units.

Therefore, to recognize the repeat of a pattern that is composed of multiple units, some errors of their structures need to be allowed. It is better to use units as the input than to use the auditory signal, because, as shown on the left side of Fig.2, using units enables correspondence between voiced segments. On the other hand, as shown on the right side of Fig.2, using auditory signals requires assigning correspondence between voiced and unvoiced segments.

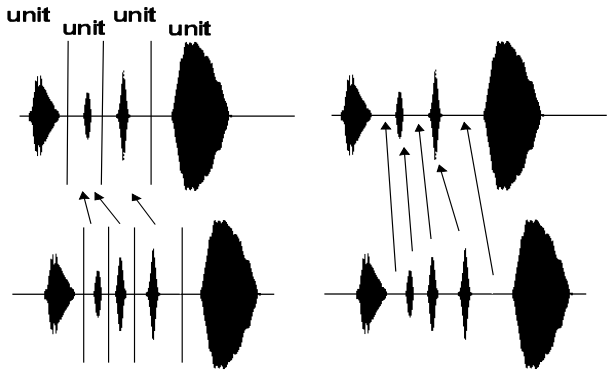

Figure 2: Correspondence of environmental sounds (Left: using units, voiced segments correspond to other voiced segments. Right: using auditory signal, the central voiced segment corresponds to an unvoiced segment).

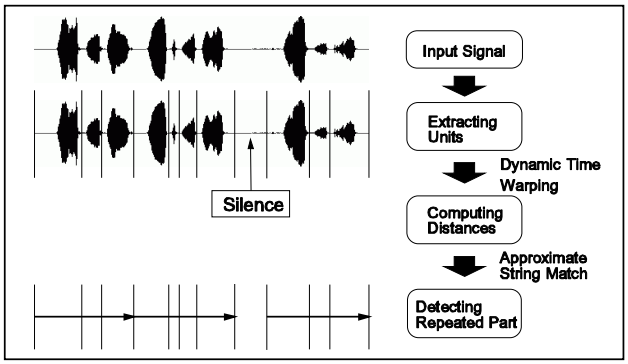

Figure 3: Process flow of repeat recognition

3 Details of the whole process of repeat recognition

Repeating parts of environmental sounds are recognized using units defined in Section 2. The whole process flow is shown in Fig.3. First, units are extracted from input of environmental sound signals by finding peaks of the power envelope. Silent segments, which affect human pattern recognition, are extracted if their lengths exceed a threshold length and treated as the separators of the pattern. Auditory distances between every pair of the units are computed by DTW. Finally, repeating patterns, which can be composed of single or multiple units, are detected (allowing some errors) from the sequence of the units.

3.1 Extracting units

In most physical phenomena, one sound is equivalent to one peak in the power envelope of an input audio signal. Therefore, each unit is extracted by extracting each peak. This is done by extracting local maximums of power envelope that do not meet the following conditions:

1. The power is below the threshold T_1.
2. The interval between the maximum and the neighbor maximum is shorter than the threshold T_2.
3. The ratio of the power to the power of the neighbor local minimum is smaller than the threshold T_3.

Condition 1 removes sounds imperceptible to the human ear, and conditions 2 and 3 concatenate local maximums that are recognized as one sound because of their closeness in time or continuity of power. This is the method previously proposed for a sound-to-onomatopoeia system [5].

3.2 Separating patterns by silent segment

We use the fact that silent segments affect human perception and are recognized as the separators of perceived patterns [7]. Whether a silent segment is long enough to be recognized as a separator depends on the lengths of other silent segments. For example, a 400-ms silent segment among continual sounds at 200-ms intervals is recognized as a separator, but a 550-ms silent segment among continual sounds at 500-ms intervals is hard to recognize as a separator.

Threshold length is therefore determined dynamically from a histogram of lengths of silent segments by using the thresholding method of Kittler [8]. For constants M_1 and M_2, the threshold is detected from M_1 ms to M_2 ms. That is, silent segments longer than M_2 are always treated as separators and silent segments shorter than M_1 are never treated as separators. Then, special units that mean "silent segments" are assumed to exist where silent segments longer than the threshold lies. Distances between two silent units are set to zero and distances between silent units and normal units are set to a high value. We also add a rule that silent units are never included within repeating pattern. In the evaluation experiments reported in Section 4, we chose $M_1 = 300$ ms and $M_2 = 1000$ ms. Fig.4 shows this thresholding method.

In this study, silent segments are defined as follows:

- Silent segment exists between (and only between) neighbor units.

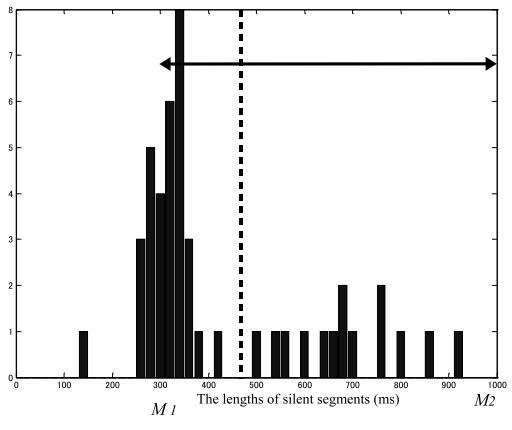

Figure 4: Thresholding of silent segments: The bar graph is a histogram of the lengths of the silent segments. Threshold is detected only between M_1 and M_2 (indicated by the arrow line). The dotted line represents the computed threshold.

- Silent segments continue as long as the power of input signal is lower than the threshold.
- The length of a silent segment can be zero.

The thresholding method of Kittler is adequate for the purpose of this study, because the result of this method is not biased even if the variances or the numbers of elements of two classes differ widely as the case of Fig.4 [8].

3.3 Using dynamic time warping to compute auditory distances between units

Distance between every pair of the units is computed by DTW. The 26-dimensional features (12-dimensional MFCCs, power, 12-dimensional ΔMFCC, and Δpower) are extracted for every frame in each unit. The frame size is 25 ms and the frame shift is 10 ms. Distances between frames are defined as an Euclidean distance of all of the 26-dimensional features. Finally distances between every pair of the units are computed by DTW, which minimizes the total distance between units by aligning their time series.

Distance $D(A,B)$ between units A and B (whose lengths are I and J) is defined as follows:

$$D(A,B) = \frac{g(I,J)}{I+J}$$

$$g(i,j) = \min \left\{ \begin{array}{l} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{array} \right\}$$

$$g(1,1) = 2d(1,1)$$

where $d(i,j)$ means Euclidean distance between all the features of i-th frame of unit A and them of j-th frame of unit B.

3.4 Repeat recognition based on approximate matching

An algorithm based on approximate matching is used to detect repeating parts in sequences of the units with the distances computed in section 3.3. The algorithm assumes the head of the unit sequence to be a base part of repeats and detects repeating parts by approximate matching algorithm. If a part matches the base part, namely, if the distance between the part and the base part is below the matching threshold, the part is registered as a repeating part, not allowing an overlapped registration. We use a general edit distance [9] for this detection, although normally approximate string matching algorithms use simple edit distance where distances between units are one or zero. In our method, the distances between units are set as corresponding to the auditory distances, which are computed in section 3.3, and the errors of deletions or insertions have constant costs.

The approximate matching is formalized as follows using dynamic programming: For assumed base part B and the remainder R, C_i and S_i defined below are calculated.

$$\begin{aligned} C_i &= \frac{g(i,J)}{J + (i - h(i,j) + 1)} \\ S_i &= h(i,J) \end{aligned}$$

$$\begin{aligned} t(i,j,1) &= g(i-1,j-2) + 2d(i,j-1) + d(i,j) + P \\ t(i,j,2) &= g(i-1,j-1) + 2d(i,j) \\ t(i,j,3) &= g(i-2,j-1) + 2d(i-1,j) + d(i,j) + P \\ \alpha^*_{(i,j)} &= \arg\min_{\alpha=1,2,3} t(i,j,\alpha) \end{aligned}$$

$$\begin{aligned} g(i,j) &= t(i,j,\alpha^*_{(i,j)}) \\ h(i,j) &= \left\{ \begin{array}{ll} h(i-1,j-2) & \text{if } \alpha^*_{(\mathrm{i,j})} = 1 \\ h(i-1,j-1) & \text{if } \alpha^*_{(\mathrm{i,j})} = 2 \\ h(i-2,j-1) & \text{if } \alpha^*_{(\mathrm{i,j})} = 3 \end{array} \right. \end{aligned}$$

$$\begin{aligned} g(i,1) &= 2d(i,1) \\ g(i,0) &= 0 \\ h(i,1) &= i \\ h(i,0) &= i+1 \end{aligned}$$

where $d(i,j)$ is the distance computed by DTW between i-th unit of B and j-th unit of R, and P is the insertion and deletion cost. Next, repeating parts

are detected. While $\min_i C_i$ is below threshold, units $S_i..m$ are registered as a repeating part and $C_{S_i}..C_m$ are removed, where $m = \arg\min_i C_i$.

In changing the length of the base part and shifting its position, the best base part that gives the lowest total distance is picked up. The total distance D_T is defined as follows:

$$D_T = \sum_i D_i + I \times B$$

where D_i is the distance of i-th registered part, I is the cost of inserting a unit, and B is the number of unregistered unit.

Although this algorithm can detect only the repeats of the head part of the sequence of units, we can extend it to the environmental sounds composed of multiple patterns. When the input environmental sounds include repeats of two or more kinds of patterns, the algorithm processes recursively. The repeats of head part are detected by the preceding algorithm, and the repeats of the rest are detected by applying the same algorithm to the unregistered units.

Additionally, using a sound-to-onomatopoeia system [5] can make a humanlike output, such as "the sound of ⟨onomatopoeia⟩ occurred N times."

4 Experiments

4.1 Experimental conditions

We evaluated the effectiveness of our method with the system implemented as described in Section 3. Experiments were executed for birdsongs in CDs for sound effects [10]. We evaluated the accuracy of the number of repeats reported by the system. It was regarded as accurate if it was the same as the number of repeats in the track counted by person. Thirty tracks of the CDs were used in the experiments. Each of the tracks ranged from 20 seconds to 2 minutes and contained multiple birdsongs of one kind of bird. In these experiments, tracks that contain fluctuating birdsongs were not used even if they were sung by one kind of bird. This is because

- fluctuating patterns can be recognized as repeating or nonrepeating, depending on the listener.
- Whether patterns are perceived as repeating or nonrepeating also depends on the purpose. For example, errors should not be allowed when user hopes to know the variation of the birdsongs.

4.2 Results of experiments

The results of the experiments are listed in Table 1. The numbers of repeats were correctly recognized in 22 tracks, so the accuracy rate was 73.3%. The specifications of the result are listed in Tables 2 and 3 separately according to whether the base of the repeat is composed of single unit or multiple units.

Table 1: Result of evaluation experiments

total tracks	30
accurate tracks	22
accuracy rate	73.3%

Table 2: (Specification of Table. 1) composed of multiple units

total tracks	22
accurate tracks	18
accuracy rate	81.8%

Table 3: (Specification of Table. 1) composed of a single unit

total tracks	8
accurate tracks	4
accuracy rate	50.0%

In the results, accuracy rate for patterns which are composed of single units was much lower than that for patterns composed of multiple units. This is because the number of repeats of a single unit is relatively more than that of multiple units as long as the length of the track is the same. Consequently, there are likely to be repeating parts that are less similar to the base part. Even if the part is similar to the neighboring part, the part is not regarded as a repeating part as long as the distance between the part and the base part is below the threshold.

Meanwhile, the base parts of repeats were accurately reported in 24 tracks. If the track is regarded as recognized accurately when the base part of repeats is valid, the accuracy rate is 80%.

5 Conclusions

In this paper, we proposed a method for repeat recognition of environmental sounds. In our method, repeats are recognized by extracting units, each of which is corresponds to "one sounding," instead of by processing untouched input audio signal of environmental sound. Our method works by extracting units and calculating distances between every pair of them. Additionally, silent segments are treated as pattern separators, and repeating parts are detected by an algorithm based on approximate matching.

The effectiveness of the method was evaluated in

experiments with birdsongs on CDs. In these experiments, the numbers of repeats recognized were compared with the numbers counted by a human. Twenty two of thirty tracks were recognized correctly, yielding accuracy rate of 73.3%.

In order to apply repeat recognition to robots in the real world, we need to improve in some ways. First, an on-line real time algorithm is required. Second, self-produced noises of robots must be considered.

6 Acknowledgements

This study was partially supported by the Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Science, Sports and Culture (No. 15200015), and by the JPSP 21st Century COE Program.

References

[1] Goldhor, R. S., "Recognition of Environmental Sounds," *Proceedings of ICASSP*, Vol. 1, pp.149–152, 1993.

[2] Martin, K., "Sound-source recognition: A theory and computational model," *Ph.D. Thesis, MIT Media Lab*, 1999.

[3] Ashiya, T., Hagiwara, M. and Nakagawa, M., "IOSES: An Indoor Observation System Based on Environmental Sounds Recognition Using a Neural Network," *Transactions of the Institute of Electrical Engineers of Japan*, Vol.116-C, No.3, pp.341–349, 1996.

[4] Cowling, M., and Sitte, R., "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, Vol.24, No.15, pp.2895–2907, 2003.

[5] Ishihara, K., Tsubota, Y., and Okuno, H. G. "Automatic Transformation of Environmental Sounds into Sound-Imitation Words Based on Japanese Syllable Structure," *Proc. 8th Europ. Conf. on Speech Communication and Technology*, pp.3185–3188, 2003.

[6] Shamoto, M., Fujita, T., Ueno, M., Tabuchi, H., and Muraoka, Y., "Analysis and Recognition of Human's Imitation" (in Japanese), *IEICE Technical Report*, SP91–139, pp.63–70, 1991.

[7] Moore, B. C. J., "*An Introduction To the Psychology of Hearing,*" Academic Press, 1989.

[8] Kittler, J., and Illingworth, J., "Minimum Error Thresholding," *Pattern Recognition*, Vol.19, No.1, pp.41–47, 1986.

[9] Navarro, G., "A guided tour to approximate string matching," *ACM Comput. Surv.* Vol.33, No.1, pp.31–88, 2001.

[10] "ANIMAL TRAX"(CD), AT7–8, The HOLLYWOOD EDGE.