Sound Retrieval with Intuitive Verbal Expressions

Sanae Wake, Toshiyuki Asahi

NEC Human Media Research Labs.
8916-47, Takayama-Cho, Ikoma, Nara 630-0101, Japan
Tel. +81-743-72-3719

Abstract

A sound retrieval method described in this paper enables users to easily obtain their desired sound. A sound representation experiment was conducted to study how people represent the sounds. Almost all representations were made with verbal descriptions that could be classified into "description of sound itself", "description of sounding situation" and "description of sound impression". The retrieval method, which adopts three keyword types, onomatopoeia, sound source, and adjective, was proposed based on the experimental results. The sound retrieval system's efficiency was discussed based on the subjective evaluation. Users can select a convenient retrieval method and adapt it to their idea of retrieval.

1 Introduction

As the use of sounds for computer interfaces, electronic equipment and multimedia contents, has increased, the role of sound design tools has become more important. In sound retrieval, picking one sound out from huge data is troublesome for users because of the difficulty of simultaneously listening to plural sounds. Consequently, an efficient retrieval method is required for sound databases.

Traditional sound retrieval methods have used acoustic features, for example, pitch, harmonicity, loudness, brightness, and spectral peaks [1], audio databases indexed by using neural nets [2], etc. These methods have been taken automatic indexing approach, and have been great success. However, they have not been verified whether the retrieval method is convenient for users. By developing the most effective and easy retrieval for users, anyone, even novice users, will be able to intuitively and effectively retrieve the sound regardless of the retrieval situation (whether the user has a concrete idea for the sound or not). To realize this, we started our research by observing how people usually represent a sound. We designed the retrieval system based on the results.

2 Experiment of sound representation

2.1 Sound grouping and arranging sound stimuli

In order to arrange the sound stimuli from various kinds of sound, we previously classified sounds into three groups: musical sound, vocal sound, and surrounding sound. Musical sound is the sound made by musical instruments, vocal sound is the sound uttered by a human or other creatures, and surrounding sound is all of the other sounds heard in our daily life. Musical sounds and vocal sounds have some striking semantic and acoustic features, we set these sounds as separate groups. We then made two groups in each of the three groups: consequent sound and signal sound. A consequent sound is "a sound produced as the result of a phenomenon" and a signal sound is "a sound produced with an intention". In addition to these six groups, there is a group that contains compound sounds. We selected 17 kinds of sound stimuli in these groups with considering also the sound sources. Table 3 shows the sound groups and the selected stimuli.

2.2 Procedure

The experiment was conducted with 14 participants (7 pairs) and was recorded on video for analysis.

1. Participant X listens to one sound stimulus.
2. X conveys the sound to a partner Y. X can use any representation to convey the sound.
3. Y concretely imagines the sound. Y can ask any question to X.
4. After imagining the sound, Y listens to the sound stimulus.
5. Y evaluates the similarity between the imagined sound and the sound stimulus.

2.3 Results and Discussion

2.3.1 Three types of sound description

Participant X mostly represented the sound stimuli with verbal descriptions. Pictures and gestures were used as supplements to verbal descriptions. Table 1 lists the descriptions for the sound "wind bell" as an example. The number besides the description shows how many times the description appeared in the experiment.

"Sound itself"		"Sounding situation"		"Sound impression"	
<Onomatopoeia>		**<Source>**		**<Adjective>**	
Chirin Chirin	6	Wind bell	7	Clear	3
Rin Rin	1	Glass	2	Quiet	1
		Glass bell	1	Cool	1
Irregularly	1	a breath of wind	3	not noisy	1
the pitch is "…………"	1	someone is ringing it	4	Relax	1
middle pitch	1	early autumn	1		
higher pitch than the bell of bicycle	1	Evening	1		
		Sounding nearby	1		

Table 1: Description used for the sound "wind bell" (Sound 13)

We listed all descriptions for every sound stimulus, then grouped similar descriptions. In these lists, we could find typical three descriptions: "onomatopoeia", "sound source" and "adjective". They were used to express "sound itself", "sounding situation" and "sound impression". In this point of view, other descriptions could be also classified on these three description types (Table 1).

(1) Description of "Sound ITSELF"

This group describes the sound itself as one hears it. The participants endeavored to express the acoustic features of the sound. "Onomatopoeias" and "acoustic cues" (pitch, duration, etc.) are used in this description.

(2) Description of "Sounding SITUATION"

This group describes sounding situations including sound sources. They can be classified into the groups:

- What (bird) - Where (in a forest) - When (in the morning)
- How (twittering) - Event (at a bird show)

(3) Description of "Sound IMPRESSION"

This group describes one's subjectivity. "Adjective", "Subjective word" and "figurative descriptions" are used.

Kawachi describes the visual information processing model from a medical point of view, that recognition and impression are independently and sequentially processed [3]. From a viewpoint of impression information processing, we assume that the difference between visual information processing and auditory information processing is only human perception. From this assumption, three types of sound description obtained through the experiment are compatible with Kawachi's model, accordingly we are convinced the three types of grouping are valid (Figure 1).

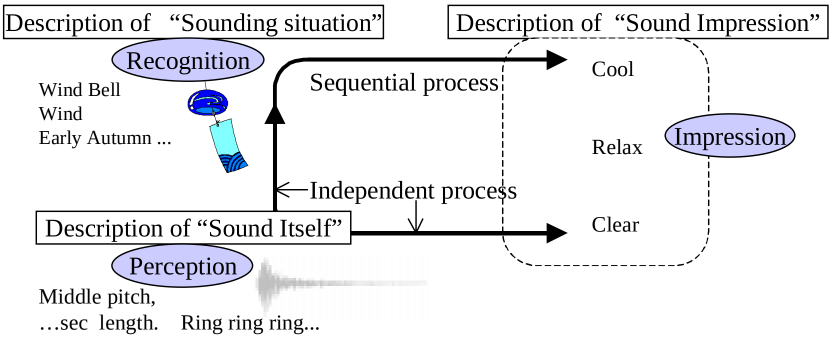

Figure 1: Auditory information processing model

2.3.2 Description types and sound characters

Table 2 shows the ratio of description types that appeared in the first description and from the first to the third descriptions. For example, if a participant expressed the sound in the following way. " It sounds "Puru Puru Puru". It is a telephone call, and it sounds annoying, and ....". We counted the first description "Puru Puru Puru" as "sound itself", counted the second description "telephone call" as "sounding situation", and the third description "annoying" as "sound impression". We can see in Table 2, "sounding situation" was the most frequently used type, and "sound itself" was often used for the "first" description rather than later descriptions.

Table 3 shows the number of description types (from the first to the third descriptions) used for each stimulus. "Sound impression" is frequently used for the stimulus "SF effective sound" (Sound 3) obviously, and also used for "guitar music". It can be said that "sound impression" is useful to represent the sound whose "sound sources" cannot be specified and the sound whose "sound itself " is too complex to be described.

	Sound itself : Situation : Impression (%)
First description	23 : 69 : 8
from first to third descriptions	7 : 86 : 8

Table 2: Description types used to represent sounds

Simple Sound						Compound Sound
Musical Sound		Vocal Sound		Surrounding Sound		
Consequent Sound	Signal Sound	Consequent Sound	Signal Sound	Consequent Sound	Signal Sound	
Musical Timbre	Melody , Harmony	Voice	Language	Daily Sound	Sign	
① a piano note 9:9:0	2. guitar music 1:15:5 ③ SF effective sound 2:2:17 ④.TV jingle 5:16:0	5. laugh 3:15:1 ⑥ cow's mooing 7:14:0	⑦ speech 6:15:0	8.river 3:18:0 9.train 3:17:1 10.foot steps 4:17:0 ⑪.chopping board 8:13:0 12.broken glass 5:15:1 ⑬.wind bell 6:13:2	⑭. telephone call 9:12:0	⑮.fireworks display 6:15:0 16.many insects singing 1:19:1 17. school noise 3:17:1

Table 3: Sound groups and stimuli with number of description types used from first to third descriptions. sound itself : sounding situation : sound impression
There are sound examples for the stimuli with circled number.

2.3.3 Dialog between Participants

Y Participants' questions (procedure 3) showed the following features.

- Questions were also about "Sound itself", "Sounding situation", and "Sound impression". Questions about "Sounding situation" appeared most frequently.
- Questions were asked to obtain details of X's description or confirm the description.

In addition, common knowledge between participants such as "sound like a telephone call in our office" was very useful for easily and correctly imaging the sound.

2.3.4 Imaging the Sound

Participant Y evaluated similarities between the imaged sound and the sound stimulus heard with the subjective evaluation. 5 points are given for an almost perfectly imaged sound, and 1 point is given for the utterly different imaged sound. Table 4 shows the averages of the subjective evaluation. This result shows that daily, real, and simple sounds can be easily and correctly imaged based on the verbal description.

Telephone call	4.7	Guitar music	3.6
Chopping board, Cow's mooing, a Piano note	4.2	Speech, River, Broken glass	3.5
Laugh, Fireworks display	4.1	TV jingle	3.2
Foot steps, School noise	4.0	SF effective sound	2.2
Train, Many insects singing	3.7		

Table 4: Subjective evaluation for similarities between the imaged sound and the sound stimulus

2.4 Summary of the experiment

We conducted an experiment to ascertain how people represent sounds. Almost all representations were made with verbal descriptions, and they could be classified with "description of sound itself", "description of sounding situation" and "description of sound impression". "Description of sound itself" is often used for the first description, "description of sounding situation" is the most frequently used method, and "description of impression" is useful to represent the sound whose sources cannot be specified. Typical examples of these groups are "onomatopoeia", "sound source", and "adjective".

3 Sound Retrieval System

Based on the experimental results, we developed the sound retrieval system. We used the following three types of sound description as input keywords.

• Onomatopoeia • Source • Adjective

3.1 Database Schema

The database consists of 814 sounds, which includes nature sounds, electric sounds, etc. Table 5 shows an example of the keyword labeling. Each sound is labeled with the keywords (onomatopoeias and sources) and adjective values. The adjective value represents a rate of the association between the sound and the predetermined adjective. For example, Table 5 shows that 50% of the people associate this sound with the word "Metallic". The way to calculate an adjective value is described in a later section.

File Name	SIND-015.wav		
Title	The Hero Has Come Onstage!!		
Onomatopoeias (Max. 8)	BaranRanRan, ByanByan, GuanWanWan, BaronRonRon		
Sources (Max. 3)	Reinforcing Bar, Piano, Stringed Instrument		
Adjective Value	Metallic ……….5.00 Annoying ……..3.33 Thick …….…...1.67 Beautiful ……...0	Clamorous ...…..5.00 Hard …………...3.33 Cheerful ……….1.67	Dull ……………1.67 Mild ……………0 Unsatisfactory …0

Table 5: Keyword labeling

3.2 Sound Retrieval

Users may input at least one type of keyword for retrieval. The system uses each keyword to calculate retrieval points that are dependent on the similarity between the input keyword and the labeled keyword. Retrieval points are calculated for each sound, then the sounds are preferentially exhibited according to total points.

3.2.1 Retrieval by Onomatopoeia

Onomatopoeia is frequently used to specify a sound, mostly as an adverb in Japanese. There is a great variety of onomatopoeias, and one sound can be expressed by different onomatopoeias. Thus, a simple keyword matching method is insufficient to cope with these variations of onomatopoeia.

We propose a retrieval method for onomatopoeia, that retrieving labeled keywords (onomatopoeia) with the input keyword itself as well as the varied keywords. In this method, onomatopoeia is treated as a combination of syllables. First, the system retrieves the labeled keywords with the input keyword itself, then by varied keywords composed by cutting one syllable from an input keyword. In addition, three other processes are applied to cope with the properties: gemination, repetition (ex. DunDun), and short onomatopoeia (ex. Pi). Retrieval points (0-10 points) are given for each sound, depending on the similarity between the input keyword and the labeled keyword.

We note that a technique for matching two character string values by comparing their phonic sounds has been proposed [4], and this technique will be useful for evaluating similarities to English onomatopoeia.

3.2.2 Retrieval by Source

The system retrieves the labeled keywords with the input keyword using simple keyword matching. When the input keyword is retrieved in the label, 10 points are given, if not, a 0 point is given for each sound data.

3.2.3 Retrieval by Adjective

Adjectives and Adjective Value

Ten adjectives shown in Table 5 were instituted based on Namba's study [5]. This study dealt with adjectives used for sound, and the similarities of these adjectives were analyzed by cluster analysis. We selected ten distinct words in the sensory dimension. The adjective value represents the rate of those who associate the word with the sound, calculated from experimental data. The following task was conducted with 6 participants.

Task: Listening to a sound and answering whether you associate the adjective (each one from 10 adjectives) with the sound

Adjective Value on Retrieval

A user may select the keyword from ten adjectives on retrieval. The adjective values, which were determined for the retrieval keyword, are set to a retrieval point for each sound. This means more retrieval points are given for a sound that is more generally associated with the input adjective.

3.3 Evaluation

We compared retrieval with the system and retrieval with the sound list. The sound list was classified by the sound source. The tasks are to retrieve given sounds and one's favorite sound in two ways. We observed the 12 participants to investigate the retrieval processes. The characteristics of each retrieval are described below.

Retrieval with the system:

- When a participant had concrete ideas of the sound, he/she could input the keywords directly for quick retrieval.
- When a participant had vague ideas of the sound, or when he/she couldn't specify the sound source, the participant could input the onomatopoeia or adjective as a keyword for a diversified retrieval.

The keyword types used on retrieval (the amount of all participants) were:
Onomatopoeias : Sources : Adjectives = 40 : 33 : 15

Retrieval with the list:

- Even if a participant had no idea to retrieve sound, he/she can listen heuristically to some sounds while referring to the sounds' categories and titles.

3.4 Summary of the Sound Retrieval System

Users can definitely and indefinitely input the keywords with this system. Furthermore, the three ways of retrieval enable sound grouping in three ways: similar sounds, sounds of the same source, and sounds of similar impressions. Therefore, users can select the most convenient retrieval method and adapt it to their idea of retrieval. As a matter of fact, all three types of retrieval were used with onomatopoeia being the most frequent method.

4 Conclusion

We have carried out an experiment to ascertain how people represent sounds. The experiment made clear that a human sound representation uses three basic types of verbal description: sound itself, sounding situation, and sound impression. Therefore, the retrieval system was designed to reflect these three types of keywords: onomatopoeia, sound source, and adjective. This technique enables users to easily, intuitively, and effectively retrieve the sound because users can select the keyword type adapted to their retrieval situation.

We plan to develop the retrieval system for designing auditory icons based on this work. In addition to retrieval by three types of keywords, retrieval by action (ex. pushing) or status (ex. new) should be effective for designing auditory icons.

References

1. Wold E, Blum T, Keislar D. et al. Content-based classification, search, and retrieval of audio, IEEE MULTIMEDIA Vol.3, No.3, FALL, 1996, pp.27-36

2. Feiten B, Gunzel S.: Automatic Indexing of a sound database using self-organizing neural nets, Computer Music J., Vol.18, No.3, Summer 1994, pp.53-65

3. Kawachi: How impression and recognition are processed in the cerebrum, KANSEI / Human / Computer, FUJITSU BOOKS, 1995, pp78-120 (in Japanese)

4. McClelland E. B, Trueblood R. P, C. M. Eastman: Two approximate operators for a data base query language: Sounds_Like and Close_To, IEEE Transactions on Systems, Man, and Cybernetics, Vol.18, No.6, Nov./Dec. 1988, pp873-884

5. S. Namba, S. Kuwano, T. Hashimoto et al.: Verbal expression of emotional impression of sound : A cross-cultural study, The Journal of Acoustical Society of Japan (English) Vol.12, No.1 1991, pp19-29