電気学会論文誌Ｃ（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.136 No.12 pp.1712–1720 DOI: 10.1541/ieejeiss.136.1712

Paper

Sound–effects Exploratory Retrieval System Based on Various Aspects

Kahori Okamoto* Student Member, Ryosuke Yamanishi** Non-member
Mitsunori Matsushita* Non-member

(Manuscript received March 25, 2016, revised July 27, 2016)

The goal of this study is to develop an exploratory search system for sound effects (SEs). SEs appreciably influence viewers' impressions of a scene in a movie; thus, SE editors must skillfully select the most appropriate SEs for each scene. Existing SE search methods, however, have three main difficulties involving the diversity of SE purposes, the representation of sound using text, and the conceptualization of SEs for a given scene. These difficulties lead to inefficient SE searches because the SE editor is forced to perform repeated searches. To solve the problem, this paper proposes a framework for SE exploration under multiple perspectives. In the framework, similarities among SEs are provided to the searcher as clues for exploration. The similarities are defined by three types of features: context, acoustic features, and symbol of onomatopoeia. This paper presents the details of the framework, a system developed with the framework, function and interaction provided by the system, and the results of user observation with the system.

Keywords: Sound effects, Music informatics, Onomatopoeia, Exploratory Search

1. Introduction

In the production of visual content such as movies and games, attaching suitable sound effects (SEs) to a scene in the content is an important task because a viewer's impression of the scene greatly depends on the SEs used. To attach a suitable SE to the scene, an SE searcher seeks and selects an appropriate one from various candidate SEs. Since an SE has reusable, one SE can be used in several scenes and occasions. For example, "Rain Sound" might be appropriate for any rainy scene. From this characteristics of SEs, various SEs are stored in repositories to reuse and the SE searcher (e.g., SE editor, visual creator) seeks and selects appropriate SEs from large SE repositories.

Finding an SE that is well suited to the searcher's intention is more difficult than expected. When the searcher searches an SE, he/she verbalizes his/her intended SE prior to conduct the search. Generally, the search is performed by matching metadata given to the SE (i.e., titles, descriptions) and the verbalized keywords that express the target SE. However, there are several SEs with the same title, and a completely different titles and/or descriptions are often given to similar SEs in the database. The impression associated with the keyword may differ from the actual impression that the SE searcher obtains by listening to the SE. The search criteria used in the search could favor certain SEs, hence some SEs might be excluded from the search result before trial listening takes place, even if they would be appropriate for the scene. For these reasons, the SE search task requires the SE searcher to access a number of SEs, and to sift through SEs while listening to them one by one. As a result, a significant amount of time is required to find a suitable SE.

The search task is more difficult if the searcher only has a vague idea or supposition for the target SE. In such a situation, he/she has to search a suitable SE by performing a lot of trial-and-error searches. Thus, the SE search task can be regarded as an extremely time-consuming task. The goal of this study is to relieve the difficulties of the task.

1.1 Contribution To meet the goal, this paper proposes an SE search system that intends to facilitate "exploratory search" of SEs. The proposed system visualizes SE's sound symbolism by using onomatopoeia and visually presents different types of similarities between the SEs. We believe that visualizing the relations contributes to establish a smooth exploration environment and it contributes to the searcher's efficient SE search.

Figure 1 shows a general SE search process. In the beginning of the process, the SE searcher inputs a sound source and/or description as a query. The intended SE, however, may not be searched successfully. The SE searcher would confuse after the step (a) in the figure, because she does not know what to do. Even when the searcher examines trial many times, she fails to meet a suitable SE. It devotes much time on the step (c) in the figure; time is up. Finally, she gives up searching and employs an SE as the target SE without her satisfaction. Such a bad case of an SE search is caused by the characteristics of SE; because SE is "sequential" media, it is difficult to grasp the SEs before listening to the SE, hence it takes too long to listen to various SEs. To solve these problems, this paper proposes an approach that "sequential" information (i.e., the SE list) is presented as a "spatial" infor-

* Graduate School of Informatics, Kansai University
2-1-1, Ryozenji, Takatsuki, Osaka 569-1095, Japan
** College of Information Science and Engineering, Ritsumeikan University
1-1-1, Noji-higashi, Kusatsu, Shiga 525-8577, Japan

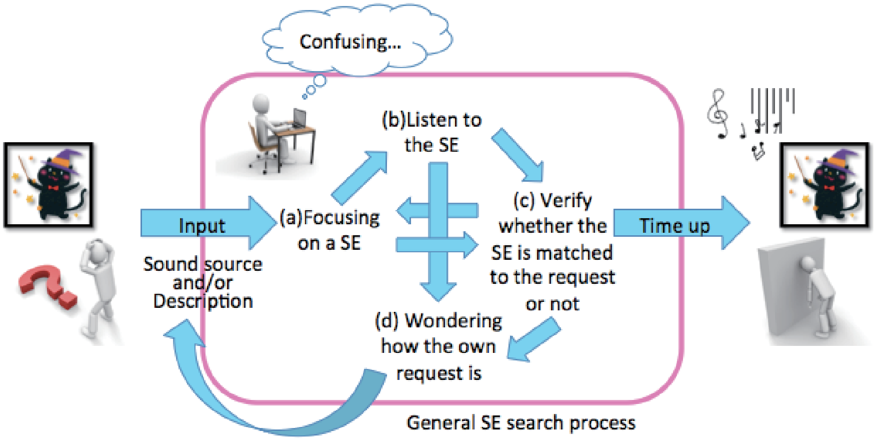
Confusing...
(b)Listen to the SE
Input
(a)Focusing on a SE
(c) Verify whether the SE is matched to the request or not
Time up
Sound source and/or Description
(d) Wondering how the own request is
General SE search process

Fig. 1.　General SE search process.

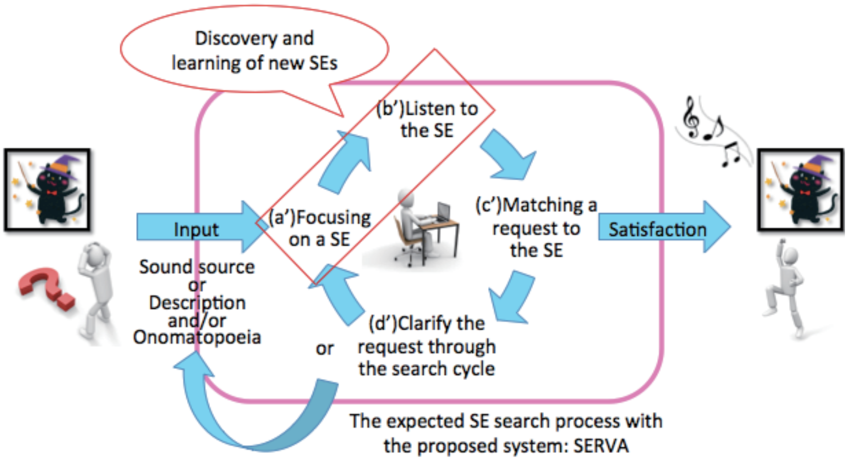
Discovery and learning of new SEs
(b')Listen to the SE
Input
(a')Focusing on a SE
(c')Matching a request to the SE
Satisfaction
Sound source or Description and/or Onomatopoeia
or
(d')Clarify the request through the search cycle
The expected SE search process with the proposed system: SERVA

Fig. 2.　The organized search process proposed in this paper.

mation on a screen. It is expected that the spatially structured relationships among SEs enable the SE searchers to smoothly and quickly understand what the focused SE is.

Visualizing the relationships among SEs alters the SE search process as shown in Fig. 2. Figure 2 shows the SE search process being iteratively conducted. The proposed system enables the SE searcher to search SEs while clarifying his/her own request through the cycle of the search as step (d') in the figure. It is expected that an efficient search in a short time would be achieved with the proposed system. In addition, using three different types of similarities is expected to give creative experiences such as the discovery and association of new SEs in the search process. Moreover, the proposed system enables users to naturally retrieve SEs by using the words by which the user verbalizes the sound effects. The characteristics of the proposed system would be more effective in supporting the users in retrieving SEs by their intuitive representation.

This paper proposes a framework for exploratory search for SEs and develops a prototype system based on the framework. The proposed system utilizes existing techniques in analyzing each of the different media processing fields; that is, there are no brand-new techniques used to correlate the similarities among SEs in each aspect. However, improved retrieval of SEs is realized by the combinations of these techniques. The effectiveness of the interface is evaluated by discussing the user's observed behavior with think–aloud method. This paper conducts user observation as the usability test and the effectiveness of the proposed system is discussed through the user-observations.

2. Problem of Searching for SEs

An SE search consists of three processes: selection, comparison, and decision. In our previous study, we have organized the following three difficulties that the SE searcher may face in an SE search [1].

- **Provision of a concept by SE's title and/or description**
  An SE used for a scene can be applied to another scene: for example, an SE for "Knocking on a wooden door" can also be applied to "Cutting vegetables on a wooden cutting board." Although an SE essentially has flexible concepts, the SE searcher might feel one concrete concept for the SE from its SE's title and/or description. The concept given by title and/or description can make an SE search difficult; the SE searcher may lightly determine SEs with a simple search based on the given concept. That is, it is suggested that a given concept prevents SE searchers from determining the appropriate SEs before listening to them.
- **Using text to represent sound**
  It is difficult to express the detailed nuances of SEs with titles alone (e.g., "wind 1") or descriptions (e.g.,

"wind blowing softly"). SE nuances such as "reverberation" and "strength" cannot be clearly obtained from the text information. The text information might provide SE searchers with the arbitrariness. It is difficult for SE searchers to correctly recognize the titles and/or descriptions that is given by the SE editor; all SE searchers do not describe a same description from listening to the same SE. Therefore, SE searchers must listen to all searched SEs to determine the most appropriate example; as a result, a significant amount of time is expended listening to irrelevant SEs.

- **Conceptualizing SEs for a given scene**
  The SE searcher may have vague ideas for the SE required because the creativity is necessary when selecting SEs for non-existent objects. Vague ideas for the SE might be clarified through the search cycle; the relationships among SEs may help the SE searcher to conceptualize the desired SE. In order to conceptualize the desired SE, they must listen to many SEs. During an SE search process, their request for SEs, might change as a result. In an effective exploratory environment, the SE searcher should be allowed to repeat the search by reformulating the query.

To reduce the expense to select the most appropriate SE for a scene, this paper designs the framework for an SE search system to solve the above three difficulties in the following section.

3. Design Concepts

The target user in this paper is a user searching for an SE who cannot clearly define his/her requirements; hereafter, this paper refers to the target user as an "SE searcher." The design concepts will be discussed in subsection 3.1. Subsequently, detail designs of an SE search system based on these concepts will be shown in subsections 3.2, 3.3, and 3.4.

3.1 Model Case Discussions for SE Search Figure 3 shows an example case of searching for an SE. In this example, John first focuses on "Walking" in "Footsteps" (Phase 1). Second, John changes his mind and searches for "Running"; the SE nuances he imagines becomes clearer as he seeks a "lighter sound" (Phase 2). John cannot find the best SE in "Running" and begins a more extensive search of the SE database. Then, John expresses his ideas using onomatopoeia, and selects SEs similar to the onomatopoeia (from Phase 3 to Phase 4). The SE that he was finally satisfied with was different from what he originally thought in this example. The characteristics of the ideal SE became clearer while searching and listening to various SEs; he listened to SEs in three categories – "Walking," "Running," and "Footsteps–comical." That is, he listened to six sound effects until he finally decided on the SE "Footsteps Comical 03." In order to facilitate SE searches, the SE search system should enable users to grasp the nuances of SEs and the similarities among SEs without listening.

Based on the discussion above, this paper focuses on the following five functions as the design concepts of the SE search system:

- Narrowing down the SEs in the database
- Finding the SEs by detail of nuances
- Grasping the nuances of SEs without listening
- Suggesting similar SEs
- Responding to the search with onomatopoeia

John is creating a new animated scene. In the animation, a fictional small creature walks along with the hero.

—–Phase 1—–

John uses an SE Web site to find the sound of the creature's footsteps. At first, John checks the sounds labeled as "Walking" in the category "Activities of people." The scene shows a creature walking in the woods. John listens to the SE labeled as "Walking on gravel." However, the sound is not suited for the footsteps of the creature, because the creature walks with shorter steps though the listening sound is a little bit louder.

—–Phase 2—–

Then, John changes his mind, and searches for the sounds labeled "Running." He listens to the sound of "Running on gravel," but he thinks "A lighter sound is better. I do not want a heavy sound." He listens to the sound of "running on soil or sand." However, the lack of lightness in the sound is similar to "Running."

—–Phase 3—–

Next, John deeply thinks "It is a fictional character, so the footsteps sound may not suit the footsteps of the creature." Then he finds the sound labeled as "Footsteps–Comical 01." After listening to the sound, he thinks "This sound rings like *Buni-buni*. It does not fit the character's image. *Pon-pon* is better." Next, he listens to the sound labeled as "Footsteps Comical 02," but the sound is not what he has imagined – the sound makes him image "heavy creature is walking."

—–Phase 4—–

Finally, John listens to the sound labeled as "Footsteps Comical 03" and is satisfied with the sound; he feels "This sound is not like *Pon-Pon* but more like *Poko-poko*. This sound is better suited to the scene. This sound should match the walking speed of the character."

Fig. 3. A model case of SE search

These functions can be marshaled as three requirements for the SEs search interface. The SEs search interface should:

(a) provide users with a search using vague expressions

(b) reduce the trial listening time, and

(c) enable users to exploratory and heuristically search for SEs.

3.2 SE Search Using Onomatopoeia Even though some SEs have the same title, the differences of SEs' nuances can give us different impressions of the SEs. For example, there are two SEs that are labeled with the same title "Wind sound." The listening impressions are greatly different from each other depending on the characteristic of wind "Breeze" or "Typhoon." Moreover, there are some levels of wind strength in "Breeze." Even if the wind strength has been described as text information, the actual listening impression might be different from the expected one. Therefore, it is hoped that nuances of SEs can be delicately expressed in text. This paper focuses on *the onomatopoeia* as the word expressing such ambiguity in detail.

Onomatopoeia is a term that emotionally and delicately describes sounds and conditions [(2)]. Because the onomatopoeia is the sensuous word, we believe that sensuous requests for SEs can be expressed by using the onomatopoeia. In particular, this paper would like to emphasize the idea that SEs can be expressed with onomatopoeia, which is described in an existing study [(3)]. Using onomatopoeia enables the user to retrieve SEs by using not only the visualized acoustic similarity but also the similarity in text, which is the word that the SE searcher uses to verbalize SEs. The similarity in text enables the users to retrieve SEs via two types of aspect — similarity of the source itself (it does not include human recognition), and the user' s input, that is, his/her recognition (how the user

Table 1. Purpose of each characteristic used to represent SEs.

Characteristic	Purpose
Acoustic features	SE searchers can recognize the detailed nuances. Acoustic features are used for finely adjusting the nuances of SEs.
Context	SE searchers understand causality of source and sound, and roughly narrow down SEs.
Symbol of onomatopoeia	SE searchers capture the characteristic sound of an SE. Symbol of onomatopoeia enables SE searchers to roughly narrow down SEs. The relationships among SEs can be spatially expressed.

interprets and verbalizes the SE). Moreover, three important elements of SEs, i.e., acoustic features, onomatopoeia, and the listening impression, are compatible. The existing research has reported that onomatopoeia is effective for capturing the acoustic impressions of SEs (4).

From these facts, it is suggested that the onomatopoeia can grasp the SE contents before trial listening. Users intuitively and specifically request target SEs' contents of the system by using onomatopoeia. Using onomatopoeia satisfies the aforementioned requirement (a) in section 3.1.

3.3 Visualization of SEs Based on Various Similarities

An SE searcher (i.e., the target user) who does not have a clear target SE in mind has to listen to SEs one by one, since the target user does not have concrete images in mind for SEs. A great deal of listening time for the work would lead to the user's burden. To solve this problem, the similarity among SEs should be grasped before listening. Visualization of the similarities would be effective in reducing the trial listening time; the target user might more quickly recognize whether an SE is the target or not.

By using onomatopoeia, SEs can be expressed via three types of information: "Acoustic features," "Context," and "Symbol of onomatopoeia." "Acoustic features" are the waveform features of an SE. "Context" is the SE's title and description. "Symbol of onomatopoeia" is the textual features of onomatopoeia. Based on these three different aspects, the relationships among SEs are visualized, and the similarities of SEs are presented. Table. 1 shows the purpose of using each aspect. Visualizing the relationships among SEs with the three different similarities enables the SE searcher to grasp how an SE is by using the similarity to other SEs.

The proposed system is developed with a database where relationships among SEs based on the three different aspects are stored. As shown in Fig. 4, the three aspects are used to form relationships among SEs. The blue node is related to other nodes by each of the three aspects. For example, "ザーザー (1): *Za– Za– (1)*, rain," which is the blue node, is similar to "サーサー (1): *Sa– Sa– (1)*, rain" based on its relation to "symbol of onomatopoeia," and is similar to "ザッパーン : *Zappa–n*, sea" from its relation to "acoustic features." Our proposed system enables the SE searcher to perform an exploratory search while visually surveying the SEs' similarities to each other. The aforementioned requirement (b) in section 3.1 is realized by visually grasping SEs prior to trial listening; it aims to shorten the trial listening time when selecting SEs.

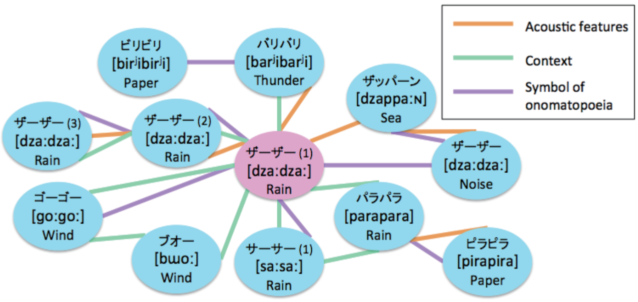

Fig. 4. Schematic diagram of the relationship among SEs. In the node, original Japanese is described and its pronunciation is described below. The context is described below the pronunciation. The nodes are linked with any types of similarity as shown as each colored line.

3.4 Exploratory Search Interface to Facilitate Awareness

When a user searches SEs, his/her ambiguous requirements must be clarified. The ambiguous requirements would be gradually clarified through listening to various SEs in the search process. Therefore, a search environment where the SE searcher can easily explore various SEs should be provided in the SE search system. As shown in Table 1, three different aspects are selectively used in the proposed system. The proposed system presents not only similar SEs but also a variety of SEs for each aspect; the opportunity to find a wide variety of SEs would be increased. Using the proposed system, the user can search SEs while confirming the similarity of SEs. Consequently, the SE searcher additionally has the following three experiences while efficiently searching for SEs: 1) identifying what query is effective for searching the SE, 2) getting an associative idea showing the similarity between SEs, and 3) discovery of new SEs. Moreover, a linked SE structure such as the example shown in Fig. 4 may save many steps when regenerating queries; SE searchers only have to follow the link. The aforementioned requirement (c) in section 3.1 is realized by selectively using the similarities from the three different aspects; the user exploratory and heuristically searches for SEs.

4. Development of SERVA

Based on the design concepts described in section 3, **SERVA** (Sound effects Exploratory Retrieval system based on Various Aspects) is developed. SERVA visualizes the relationships among SEs with three type of similarities: "Acoustic features," "Context," and "Symbol of onomatopoeia." Fig. 5 shows the interface design of SERVA. The SEs are searched as inputing query: onomatopoeia and/or a sound source. After clicking the search button, an SE matched to the query is placed in the center as a pink node (hereafter, the center node). The center node is linked to each of the other SEs. The links are derived by the similarities of the three aspects: the center node is similar to orange nodes in acoustic features, green nodes in context, and blue nodes in symbol of onomatopoeia. The length of the link is changed at random to prevent the overlap of texts.

SERVA is a Web application, which is developed by using HTML, CSS, JavaScript, and jQuery†. The SEs are visualized by the Force layout function of D3.js (5). The database of

† http://jquery.com/

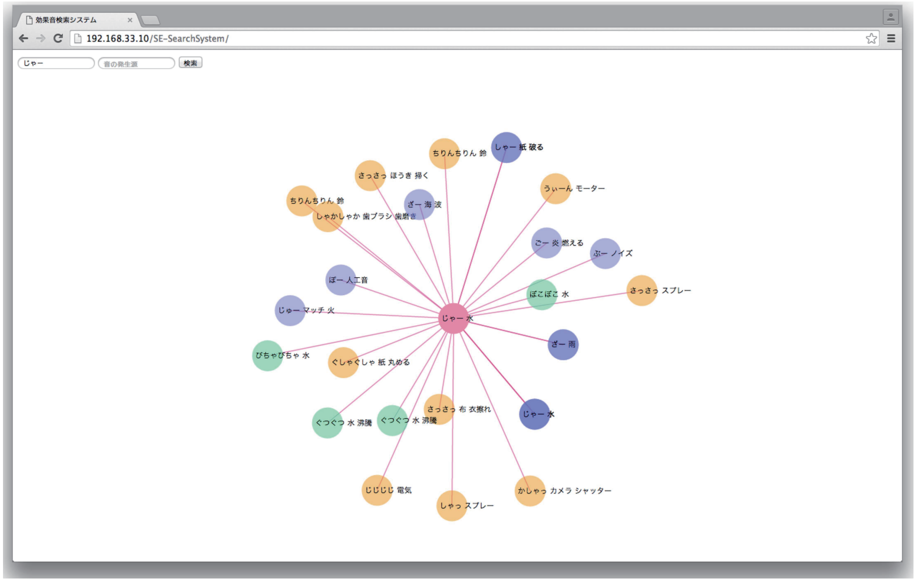

Fig. 5. The interface design of SERVA.

Table 2. SE's information examples of SERVA

Onomatopoeia	Description 1	Description 2
Kachan	Pot	Open
Karan	Fall	
Koro Koro	Pencil	Roll
⋮	⋮	⋮

SEs is managed in the JSON. Currently, the SEs registered in SERVA are 100 samples that were associated with the onomatopoeia in previous research [1]. Each of these SEs has one onomatopoeia that represents the SE and at most two links that represent the sound sources and/or detailed descriptions (Table 2).

The visualization methods will be shown for each similarity type; similarity in acoustic features, context, and symbol of onomatopoeia. Each will be detailed in sections 4.1, 4.2 and 4.3, respectively.

4.1 Acoustic Similarity Among SEs As shown in Section 3.2, the acoustic features and the onomatopoeia are compatible. The acoustic features and symbol of onomatopoeia are expected to have a certain correlation; these are essentially different indicators though "acoustic features" and "symbol of onomatopoeia" have a certain correlation. A variety of SEs with similarities are presented by dealing with both features in order to improve the retrieval of SEs.

Humans feel some kind of impression (e.g., powerful, bright, dry) when listening to an SE. In addition, humans recognize what creates the sound and where the sound comes from. The impression and identification aspects, that is, the tone features [6], might provide us with such abilities. The impression aspect refers to the characteristics of tone in that the tone features can be represented by using adjectives. The impression aspect can be arranged into three or four factors: "aesthetic," "metal", "powerful," and sometimes, additionally, "softness." Each factor can be represented by each corresponding adjective pair, e.g., "unclouded–turbid," "sharp–dull," and "weakness–strong" that correspond to aesthetic, metal, and powerful factors, respectively [7]. The identification aspect refers to the characteristic where the listener recognizes what the sound is and how it is created. Human listens to and recognize sounds because of this aspect. The SE searcher also uses these two types of aspect, thus the SE searcher should be assumed as tone. Focusing on the impression aspect, the following eight kinds of acoustic features related to impression are extracted from each SE: (1) RMS energy, (2) Low energy, (3) Tempo, (4) Zero cross, (5) Roll off, (6) Brightness, (7) Inharmonicity, and (8) Mode. The detail motivation to use these features, that is, how the features are related to each factor in the impressive aspect, is shown below.

- Aesthetic factor

  "Unclouded–turbid" is a general expression used to represent the presence or absence of dissonance. Therefore, we use (7) Inharmonicity and (8) Mode such that each indicates the Inharmonicity and Mode features as the acoustic features for the aesthetic factor.
- Metal factor

  Low tone affects the impression of soft and hard: the bigger the amount of low tones the harder the sound is. As the acoustic features for the metal factor, we use (4) Zero cross, (5) Roll off, and (6) Brightness which are related to the pitch, the ratio of low tone, and the ratio of the treble tone, respectively.
- Powerful factor

  Humans have strong feelings from sound when the volume is large and the rhythm is fast. This paper uses the features (1) RMS energy and (2) Low energy for volume and (3) Tempo for rhythm.

We selected the above features referring to the related work [8]. To extract the features, we use the MIRtoolbox[†] [9], which is a music information-processing tool. Acoustic similarities among SEs are developed based on the results of analysis in our previous research [1].

The ranges of the acoustic feature values differ. Thus, the

[†] `http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox` Confirmed in January 11, 2015

*i*th feature value of an SE *s* is normalized using the following equation:

$$f_i'(s) = \frac{f_i(s) - f_i^{\min}}{f_i^{\max} - f_i^{\min}}, \cdots\cdots\cdots\cdots \tag{1}$$

where $f_i^{\min}$ and $f_i^{\max}$ each shows the minimum and maximum values for f_i in all SEs, respectively. The normalized *i*th feature value, $f_i'(s)$, is regarded as the *i*th feature of *s*. If an SE does not have periodicity, it is impossible to extract the tempo feature, i.e., f_3; in that case, f_3 is set to 0. Then, hierarchical clustering with Ward method[10] is conducted for the SEs. The resulting dendrogram is divided into clusters according to the rule that "a cluster cannot consist of only a single SE." The 100 SEs are clustered into nine groups. In this clustering, it is confirmed that 50% of the SEs in the group have similarity through the subjective evaluation experiment.

Based on this classification, it is confirmed that SEs belonging to the same group are regarded as similar SEs in acoustic features. The SEs similar to the center SE are linked with *orange* nodes.

4.2 Context Similarity Among SEs Context similarity is calculated by using related words of SEs' titles and descriptions. The related words are obtained by using the *word2vec*[†] method. *Word2vec* is a quantification method that detects similar words by using the vector of the word obtained by using the *word2vec* method, with each word being represented as an (approximately) 200-tuple vector. This method represents the semantics of the word as a mathematical model. The method is based on the assumption that words in the same context have similar semantics.

The semantics of the word can be calculated because of the vector representation; this characteristic suggests a future expansion of the proposed system. To represent the sound in a word, Wake *et al.* have revealed that the subjective expressions (e.g., "clear sound of glass" and "hard wood sound") are used with the sound source and onomatopoeia[3]. Subjective expression is a type of intuitive representation; it might enable us to ambiguously represent the target SE. The users would be able to retrieve SEs with a query such as "the sound like word *A* minus *B*" using the calculation of the semantics of the words using *word2vec*. Such an improvement would provide users with a more intuitive interaction in the retrieval of SEs, which is an ambiguous task. This paper looks ahead to such future work, and uses *word2vec* as the context modeling method.

Since the word2vec method requires a learning corpus, two types of corpus are prepared in this study — Aozora Bunko and blog entries in the Ameba blog[††]. The documents as the learning corpuses are morphologically analyzed and separated into terms using MeCab[11], a Japanese morphological analyzer. Each term is converted to the basic form, and the basic form is used in the learning process. For the related words for one object in the title of an SE, the word2vec method obtains five terms from each corpus. A maximum of 20 related words are given to the SE as related word tags. The following terms are excluded from the related word tags: onomatopoeias, synonyms (e.g., "Sound of the bell" to "Bell"), the different representation between Hiragana and Kanji in Japanese, terms that are equal to the object in the title, and adjectives.

To represent context similarity, SERVA refers to the related word tags of the center node. The SEs with an object that matches the related word tags are regarded as similar SEs in context; the similar SEs are presented as *green* nodes and linked with the center node.

4.3 Similarity Among SEs with Symbol of Onomatopoeia The similarity of symbol of onomatopoeia is visualized by calculating the string distance between onomatopoeias.

The string distance is measured based on the Levenshtein distance[12]. The Levenshtein distance calculates the edit cost and the edit cost is the minimum number of steps necessary to convert from a given string to another string by insertion, deletion, and substitution. The Levenshtein distance can be calculated using the Dynamic Programming method[13], which is often used to calculate the distance between the phonemes in speech recognition. In order to calculate the Levenshtein distance, prior to the calculation, Japanese onomatopoeia that is attributed to the SE is converted into Roman characters.

SERVA calculates the Levenshtein distance between the center node's and the other SE's onomatopoeias. The SEs that have a Levenshtein distance of two or less to the center node are presented as similar SEs in symbol of onomatopoeia. The similar SEs are presented as *blue* nodes and linked to the center node.

4.4 How to Use SERVA In SERVA, "onomatopoeia" and/or "sound sources" are available as the search query inputs (see Fig. 6). A querying box has a function to suggest queries. Depending on the input, the system suggests onomatopoeia and/or sound sources of the SEs in the database. With the suggestion function, our system provides query candidates to the users by taking the following two types of availability into account: the prior confirmation of SEs in the database, and the auxiliary input to the query for the beginner who does not know what query will be appropriate.

The nodes group is generated by clicking the search button; several SEs, each of which has a similarity to the center node, are radially presented (see Fig. 7). These SEs can be played when a cursor is placed on the node (i.e., mouse-over event), and stopped when the cursor leaves the node (i.e.,

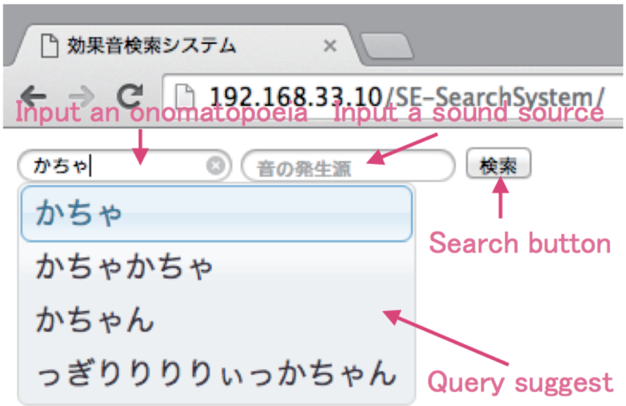

Fig. 6. Screen example of inputting query.

[†] https://code.google.com/p/word2vec/

[††] Confirmed in December 15, 2015

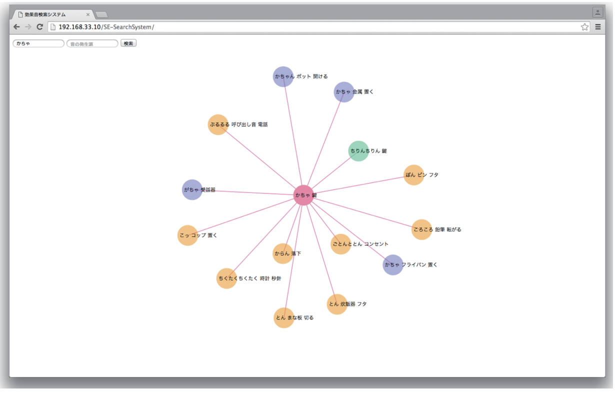

Fig. 7. Result screen after the search button.

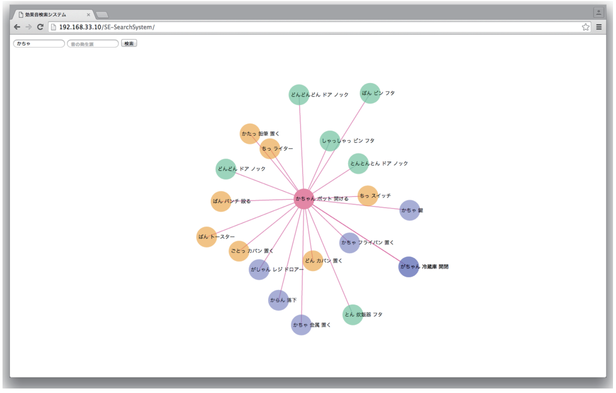

Fig. 8. Nodes group is updated as clicking a node.

mouse-out event). This playing function provides not only for smooth operability of the "play" and "stop" functions but also the chance of accidentally playing of unintended SEs. We believe that the users should have a chance to listen to new SEs easily, hence the discovery and learning of the unintended SEs is expected. This function amplifies the opportunity to have secondary experiences; the proposed system is designed to provide users with serendipity in a sense. When clicking a node, it moves to the center of the interface and the clicked node becomes the new center node. Then, new nodes that relate to the new center node appear and surround it (see Fig. 8). With this system, the user can explore SEs by examining similar SEs.

5. Usability Test

User observation was conducted as the usability test of SERVA. Through the observation of SERVA users, the process of how to find SEs was discussed. In this study, interaction between a participant and SERVA was observed until they were satisfied with the searched SE. Five persons were employed as participants: four participants (two males and two female) in their twenties and a female in her fifties.

The SEs database in SERVA was not exactly complete. For example, the database of SERVA did not have the SE entitled as "Wind chime" in its 100 SE samples. However, the database had the alternatively usable SE with an onomatopoeia "ちりんちりん : *Tirin Tirin*"; this onomatopoeia is typically used to express the sound of a wind chime in Japanese.

Three images were prepared as sample images to be used with an SE: "A dripping tap," "Switch," and "Wind chime." The users searched the SEs for the each image using SERVA where sound source and onomatopoeia were available as the query inputs. Then, the users searched SEs while speaking what he/she thought, i.e., think aloud [(14)]. The search process was recorded from the beginning to the end. After the search, the users were interviewed as to whether they selectively used the three aspects.

How did the users search SEs from such an incomplete database? What queries did the users input to the system? Which type of similarity did the users focus on? What process did the users follow in the search? These are the points to be discussed in this paper.

5.1 User-SERVA Interactions Approximately 5–10 min were spent by the users to finish the search for SEs. All users defined the object that should have a SE, and generated the query. The query generations were summarized into four patterns: a) the name or sentence concerning the target (e.g., wind chime), b) onomatopoeia converted from the sound caused by the object (e.g., ちりんちりん: *Tirin Tirin*), c) scene where we often hear the sound caused by the object (e.g., summer), and d) the alternative query such as a synonym (e.g., from "ちりんちりん: *Tirin Tirin*" to "からんからん: *Karan Karan*"). The query was regenerated only in the case that the target SE was not found by the initial query, that is, d) above.

All users searched for SEs following the same process after inputting the initial query by: 1) listening to all of the presented SEs, 2) narrowing those down to some candidates, and 3) selecting the SE for the image while listening and comparing the candidates. The process was repeated until the users determined the SE. In process step 1), the users were interested in several SEs and listened to them even though the resulting SEs are different from the target SE. In process step 2), the difference between one SE and an SE of the same title were confirmed by listening and comparing the SEs many times. The users compared a given SE with other SEs that had the same title many times, and they clarified the differences. The listened to SEs were correlated with his/her experiences, as they stated, "This sound is similar to another sound I checked." The users often listened to other candidates in the cycle of comparison, and the exploratory search was repeated after the target SE was found. It was confirmed that the three types of similarities were all used in the search process.

One user had learned the SEs throughout the search process. The user searched SEs while recalling the already learned SE in the next search: "I heard this sound a little while ago." It was confirmed that 67% of SEs finally given to the images had a different title from the initial query. One user quit the search when he could not find SEs that had his intended sound source in the title even though he was not satisfied with the SE. The user selected the SE that had the intended onomatopoeia without any searching at all.

In the interview, four out of five users answered, "I did not mind the selective usage of various aspects with the system." The user who selectively used the different types of similarity answered, "To expand my own idea for the SE, I focused on "acoustic features," and I used "context" and "symbol of onomatopoeia" when I clearly figured out the target SE."

5.2 Discussions for User Behaviors Exploratory search of SEs was conducted using SERVA. The users understood what kinds of SE were stored in the database while repeating the search process. Then, they listened to all of the presented SEs and focused on some of the candidates. This process can be regarded as a typical information retrieval behavior of exploratory search[15]: "Exploratory Browsing (Search to expand the search space)" and "Focused Searching (Search to narrow the search space)." It was suggested that SERVA provided users with the environment based on the Exploratory search model.

SERVA might change a user's initial ideas for SEs. The initial query was mostly different from the finally selected SE's title. It was suggested that the initial request was changed or clarified through the search process. Some of the users searched SEs while learning what and how SEs were. To present similar and variety of SEs should contribute to the subsequent searches. The users continued the exploration after the target SE was found. From their behaviors, it can be considered that the users verified whether the selected SE was the optimal SE or not. From this fact, it suggests that the satisfaction with their own selection was necessary and important in completing the SE search.

5.3 Discussions for System Equipments Exploratory search of SEs selects and listens to various SEs, and repeats the process of acceptance and rejection. However, the current version of SERVA does not display the nodes group of the previous stage after a new nodes group is formed. To listen to the SEs presented in the previous stage, the users needed to search the SEs again. It seems that the determination of the target SE would be facilitated by looking back to a previous search. The development of a search history or a list that holds the SE candidates would be a future work.

SERVA presents words related with the sound source and description, and the onomatopoeia. Based on this information, the users have a sense whether the target SEs are in the database or not. Although this information might be useful to easily conduct exploratory searches for SEs, one user quit the search without exploring SEs by looking at this information only. It seems that this information would be of less use than interrupting the search process; a given concept might be given to the SEs.

In the current SERVA system, the users have to explore a group of nodes in order to listen to the SE candidates. The comparison of the SEs in the search process might be conducted for only a certain nodes group. That is, the SEs might be searched from only the nodes group where similar titles were given to the SEs. By only visually selecting SEs, the number of searches can be reduced. In the future, we will consider the appropriate information required to better facilitate the exploratory search.

6. Related Work

The related work of this study is marshaled from each aspect below. This study is closely related to the following three types of researches discussed below: SE search system, music exploratory search service, and visualization of onomatopoeia.

6.1 SE Search System Wake *et al.* have proposed an SE search system where the queries were (a) onomatopoeia, (b) sound source, and (c) subjective representation[3]. The queries are based on three representation methods to represent the SEs, which is presented in the same paper. The three representation methods are (1) description of the waveform information, (2) the source of the sound, and (3) subjective description. The system proposed by Wake *et al* accepts at least one type of query to search SEs, and then the search score is calculated. The search score is the degree of similarity between an input query and SE information in the database.

In their system, similar SEs to the input query are also presented based on the degree of similarity; the search result is just shown in high-search-score order via a list. The system does not cover the exploratory search function.

6.2 Music Exploratory Search Service Hamasaki *et al.* have proposed a music exploratory search service[16]. With the service, users can discover a new song while grasping the relationship between the song and others. A song is connected to the other song by an arrow tag that shows the relationship of the songs. The arrow tag can be freely given to the songs by general users of the service. Five types of availability are denoted by the arrow tag: (1) understanding the positioning of the content before viewing, (2) discovering the relationship between the contents, (3) attention to the content groups that have a similar relationship, (4) culling the arrow tag that has less support, and (5) inspiring the creativity for new contents based on the relationship. The user discovers a new song by tracing the nodes connected by the arrow tags.

If a song is within the scope of the preference, most songs can be accepted in music retrieval. Thus, a lot of information would be effective in satisfying the user intent to get a list of songs. On the other hand, only *one* sample directing a scene should be selected in an SE search. In an SE search, a diverse and large amount of information about SEs is not necessarily effective. As mentioned in section 2, a concrete concept given to SEs might interrupt the search process. We consider the flexibility to freely interpret the sound should be a requirement in the SE search results.

6.3 Visualization by Onomatopoeia Tomoto *et al.* have proposed a search system visualizing similarity of deserts using onomatopoeia[17]. The system displays the onomatopoeia related to the texture of the dessert and dessert of the corresponding image. The users can grasp the similarities between desserts from a map showing the positional relationship. Onomatopoeia is a word that specifically shows the impression of the object.

In this study, not only onomatopoeia but also acoustic features and context are used to represent the relationships between SEs. The similarities based on various aspects should allow users to clearly and visually search for SEs.

7. Conclusions

This paper proposed a sound-effect exploratory retrieval system — SERVA. SERVA visualizes the relationships among SEs in terms of three types of similarity: "acoustic features," "context," and "symbol of onomatopoeia." Visualizing SEs allows the user to frequently conduct exploratory search. Through user observations, it was confirmed that the proposed system enabled the users to learn and expand upon the ideas for the SE in the search process. The experiences in the previous search were effectively referred in the subsequent search. It was suggested that the more the user searches for SEs with the proposed system the better their skills for searching SEs would become. From the discussions of the usability test, we believe that the proposed system not only provides an intuitive SE search environment but also educates users on how to efficiently search for SEs.

In the usability test, one user stopped searching SEs without the exploration if SEs that have a title certainly matched with the query were not found. The kind and amount of information used for the search process will be reviewed in our future work. We will develop a more effective SE retrieval system to promote more intuitive and efficient searches.

Acknowledgment

This work was supported by JSPS KAKENHI Grand Number 15H02780.

References

(1) K. Okamoto, R. Yamanishi, and M. Matsushita: “Exploratory searches for sound effects: Verification of similarity based on the acoustic features of sound effects”, In Proceedings of The Fourth Asian Conference on Information Systems, pp.MS1–3 (2015)

(2) M. Ono: Giongo Gitaigo 4500 Nihongo Onomatope Jiten (Onomatopoeia Dictionary). Shogakukan (2007) (in Japanese)

(3) S. Wake and T. Asahi: “Sound retrieval with intuitive verbal expressions”, In Proceedings of the 5th International Conference on Auditory Displa (1998)

(4) S. Iwamiya: “Science of sound color: evaluation and creation of timbre and sound quality”, Vol.1 of Sound science series. Coronasha (2010) (in Japanese)

(5) M. Bostock, V. Ogievetsky, and J. Heer: “D3: Data–driven documents”, Visualization and Computer Graphics, IEEE Transactions on, Vol.17, No.12, pp.2301–2309 (2011)

(6) H. Helmholtz: “On the Sensations of Tone”, Dover Publications (1954)

(7) O. Kitamura, S. Ni-i, J. Kuriyama, and N. Masuda: “Extraction of timbre factor for the youth of 1975”, Proceedings of the auditory research meeting (1978) (in Japanese)

(8) K. Kusama and T. Itoh: “Muscat: a music browser featuring abstract pictures and zooming user interface”, In Proceedings of the 2011 ACM Symposium on Applied Computing, pp.1222–1228 (2011)

(9) O. Lartillot, P. Toiviainen, and T. Eerola: “A matlab toolbox for music information retrieval”, Data Analysis, Machine Learning and Applications, pp.261–268 (2008)

(10) J.H. Ward: “Hierarchical grouping to optimize an objective function”, Journal of the American Statistical Association, Vol.58, No.301, pp.236–244 (1963)

(11) T. Kudo, K. Yamamoto, and Y. Matsumoto: “Applying conditional random fields to japanese morphological analysis”, In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230–237 (2004)

(12) V.I. Levenshtein: “Binary codes capable of correcting deletions, insertions, and reversals”, SOVIET PHYSICS–DOKLADY, Vol.10, No.8, pp.707–710 (1966)

(13) S.B. Needleman and C.D. Wunsch: “A general method applicable to the search for similarities in the amino acid sequence of two proteins”, Journal of Molecular Biology, Vol.48, pp.443–453 (1970)

(14) P.C. Wright and A.F. Monk: “The use of think–aloud evaluation methods in design”, SIGCHI Bull., Vol.23, No.1, pp.55–57 (1991)

(15) R.W. White and R.A. Roth: “Exploratory Search : Beyond the Query-Response Paradigm”, Morgan and Claypool Publishers (2009)

(16) M. Hamasaki and M. Goto: “Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community”, In Proceedings of the 9th International Symposium on Open Collaboration (2013)

(17) Y. Tomoto, T. Nakamura, M. Kanoh, and T. Komatsu: “Visualization of onomatopoeia based on phonemic features”, Transactions of Japan Society of Kansei Engineering, Vol.11, No.4, pp.545–552 (2012) (in Japanese)

Kahori Okamoto (Student Member) received a Bachelor of Informatics from Kansai University in 2015. She enrolled in Graduate School of Informatics, Kansai University in 2015. Her research interest includes Human-Computer interaction, Music informatics, and Onomatopoeia Computing.

Ryosuke Yamanishi (Non-member) received each B.E., M.E. and Ph.D. degree from Nagoya Institute of Technology, Japan, respectively 2007, 2009 and 2012. He has joined to College of Information Science and Engineering, Ritsumeikan University as a Research associate 2012-2013, a Research assistant professor 2013-2014, and an Assistant professor 2014-present. His research interest includes Affective computing, Multimedia processing including music, language and image, and Human-Computer interaction. The Japanese Society for Artificial Intelligence, Information Processing Society in Japan, Japan Society for Fuzzy Theory and Intelligent Informatics, Japan Society of Kansei Engineering and Association for Computing Machinery member.

Mitsunori Matsushita (Non-member) received a Ph. D. degree in engineering science from Osaka University in 2003, and is presently a professor working at the faculty of informatics, Kansai university. He has worked on interaction design, multimodal information access and development of interactive systems. The Japanese Society for Artificial Intelligence, Information Processing Society in Japan, The Virtual Reality Society in Japan, and Association for Computing Machinery member.