# REVISED CATALOGUE SPECIFICATIONS OF SPEECH CORPORA WITH USER-FRIENDLY VISUALIZATION AND SEARCH SYSTEM

*Shuichi Itahashi[1], Tomoko Ohsuga[2], Yuichi Ishimoto[3], Hiroaki Kojima[4], Kiyotaka Uchimoto[5], and Shunsuke Kozawa[6]*

[1]Tsukuba University, Tsukuba, [2]NII, Tokyo, [3]NINJAL, Tokyo, [4]AIST, Tsukuba, [5]NICT, Kyoto, and [6]Hatena Co., Ltd., Kyoto, Japan

## ABSTRACT

It is well known that speech corpora are indispensable to speech research; several data centers of speech corpora have been set up worldwide in order to meet this demand that serve as a repository for various speech corpora. However, they use different specification systems for their corpora, and so it is difficult for speech corpora users to compare and select suitable corpora. It would be more convenient for the users if each data center used a common specification system for describing its corpora. Based on this idea, we have already proposed a set of specification attributes and items as the first step towards standardization, but the scale of the retrieval system was limited. This paper introduces a revised version of the speech corpora specification attributes and items to be connected with the large-scale metadata database "SHACHI" combined with the "Concentric Ring View (CRV) System" to improve the user interface.

***Index Terms—*** Database, retrieval, corpus, language, resource.

## 1. INTRODUCTION

A large amount of various kinds of speech data is required for speech research. This is partly because a large quantity of speech data is necessary to train speech recognizers, most of which are based on statistical methods and partly because it is necessary to compare and evaluate the performance of any new methods of speech processing. With this background, it is indispensable to develop a large amount of various kinds of speech data to be used in common and to establish a system for its effective utilization not only for the purpose of utilizing them for research and development but also for the purpose of performance assessment of speech processing systems.

In order to meet this demand, several datacenters have been established in the U.S.A., Europe, Asia, and elsewhere. The amount of data distributed from such datacenters is quite abundant, as classified in Table 1. Such diversity gives users more freedom of choice, but it has become difficult to select suitable corpora for the intended purpose from the great variety of corpora that has been made available. Although it is possible to search for these data from the websites of each datacenter, the metadata of the corpus description are not unified and so it is not easy for the users to find the necessary corpus. Therefore, we think it would be more convenient for the corpus users if the catalogue specifications of the corpora were standardized among all the various data centers.

We have already proposed a scheme for standardizing the speech corpora catalogue specifications [4]. We have also adopted an interactive visualization and search system of speech corpora called the "Concentric Ring View (CRV)" system, which simultaneously creates its visual display and data retrieval [1]. In addition, a large-scale database system called "SHACHI" was developed to collect the metadata, such as tag sets, formats, and recorded contents, from the language resources worldwide [7]. We have come up with an idea this time that combines these two systems with the revised corpus specifications that will greatly improve the search and display of speech corpora [3]. Section 2 introduces the outline of the SHACHI metadata database, Section 3 explains the revised version of the attributes for the speech corpora specifications, Section 4 presents an outline of the CRV system, Section 5 describes the experiment and its results, and finally, Section 6 concludes the paper.

## 2. SHACHI, LARGE-SCALE METADATA DATABASE OF LANGUAGE RESOURCES

SHACHI is a large-scale metadata database of language resources developed jointly by the National Institute of Information and Communications Technology (NICT) and Nagoya University of Japan [7]. It collects detailed metadata information on the language resources worldwide. The metadata set adopted by SHACHI conforms to the OLAC [9] metadata set, which is based on 15 fundamental elements of the Dublin Core [10] and constitutes an extended version of OLAC with 19 newly-added metadata elements that were considered to be indispensable for describing the characteristics of language resources. It semi-automatically collects 55 kinds of detailed

metadata information and contained about 3000 compiled language resources in July 2013. ELRA also has a similar system, i.e., a universal catalogue, but the most important feature of SHACHI is that the automatically collected metadata is manually corrected [7]. It also has its own search function based on keywords, facets, and ontology. However, it would be much more convenient if SHACHI could be combined with a visual search system such as CRV, which we will mention in Section 4.

## 3. REVISION OF ATTIBUTES AND ITEMS FOR SPEECH CORPORA DESCRIPTION

It is necessary and more convenient if we could use a set of attributes that describes a corpus to search through the various speech corpora. We proposed eight attributes, each of which contains 4 - 14 items, in our previous paper [4]. SHACHI also uses several attributes to describe the language corpora. However, SHACHI focuses mainly on text corpora and it does not have enough suitable attributes to describe the speech corpora. We have revised the set of attributes and items discussed in Reference 4 to connect SHACHI with the CRV search system. Figure 1 illustrates the structure of the present research.

We have re-examined the attributes proposed in our previous paper [4] so that we can more suitably describe the speech corpora. The revision is two-fold; first, a SHACHI attribute was changed so that it fits the CRV system, and second, we have changed some of the items used in CRV in order to conform to the specifications used in SHACHI. The new set of nine attributes is itemized in Table 2.

Regarding the first revision, the “speaking style” attribute concerns whether the corpus contains read speech, acted speech, or spontaneous speech, although formerly it had such items listed as continuous speech, isolated words, and non-native speaker; former two items “continuous speech” and “isolated words” will be covered by the new “sentence length” attribute and the item “non-native speaker” is covered by the new “speaker” attribute.

As for the second revision, we have added two new attributes, “sentence length” and “speaker attribute”. The “sentence length” attribute indicates whether the corpus contains isolated words, or short or long sentences. The “speaker attribute” indicates whether the speaker in the corpus is a native speaker of the language, a professional, child, senior, or others. In addition, the “language” content was changed from the former one; formerly it contained such designations as monolingual, multilingual, and dialect, but now it indicates the area where the languages are spoken, such as Asia, Europe, Africa, etc., which were originally used in SHACHI. Actually, the specific language name is shown in the detailed information area on the right side of the screen, as shown in Fig. 2. The “multilingual” and “dialect” items are covered by the new “characteristics” attribute.

The new “characteristics” attribute concerns both types of revisions and includes such items as multilingual, dialect, dialogue, emotional, non-speech, and others.

The “input device”, “input environment”, “sampling rate” and “number of speakers” attributes remain the same as before, but the “sampling rate” attribute is taken as a separate one because it is a very important parameter for a speech corpus; although formerly it belonged to the “data mode” attribute.

Furthermore, we have included more information from SHACHI such as the title, the publisher, the price, and the URL for more information on the corpus in the detailed information area on the right side of the screen, as shown in Fig. 2. In this way, the new set of attributes will more suitably reflect the content of the corpora. The extended SHACHI in Fig. 1 has already been adopted in this new set of attributes.

## 4. INCORPORATING RING-TYPE VISUAL SEARCH

The authors have adopted a novel search and display system of speech corpora called the “Concentric Ring View (CRV)” system [1]; it is an interactive environment for integrating searching and browsing of various corpora [5, 6]. It is composed of several concentric rings, each of which corresponds to a speech corpora attribute, as shown in Fig. 2. Initially, only the outermost ring was displayed, which expresses the attributes of the corpora, as shown in Fig. 3. It is divided into several sectors, each corresponding to an attribute.

By clicking a certain sector, another ring, an item ring appears inside. This item ring holds the category items that correspond to the attribute category specified on the attribute ring. Those corpora specified by the attribute and items on the rings are displayed inside the ring. A user can rotate the item ring and adjust the item by dragging a suitable sector of the item ring and browse the searched results shown inside the rings. The current item value is always shown at the bottom of each ring in a highlighted color, so the user can easily check the current position or condition. The displayed information can be narrowed down by specifying more attributes, which causes other rings to appear inside. This technique allows users to easily and precisely specify each item. The details of a specified corpus are displayed on the right side of the screen by clicking on the desired corpus shown inside the rings. This is an attribute-based search and users can search corpora by any attribute in any order.

Table 1　Number of LRs of several data centers

| Organization | # | Category |
|---|---|---|
| LDC, U.S.A. | 596 | Text, Speech, Lexicon |
| ELRA, Europe | 1121 | Text, Speech, Lexicon |
| C-LDC, China | 97 | Text, Speech |
| CCC, China | 35 | Text, Speech, Image |
| SiTEC, Korea | 26 | Speech |
| GSK, Japan | 20 | Text, Lexicon, Speech |
| NII-SRC, Japan | 43 | Speech, Sound, Image |
| ALAGIN, Japan | 38 | Text, Speech, Tools |

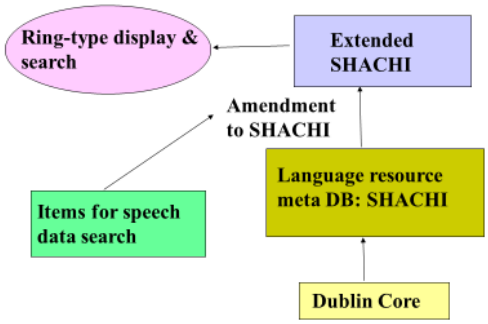


Fig.1　Structure of search and view system.

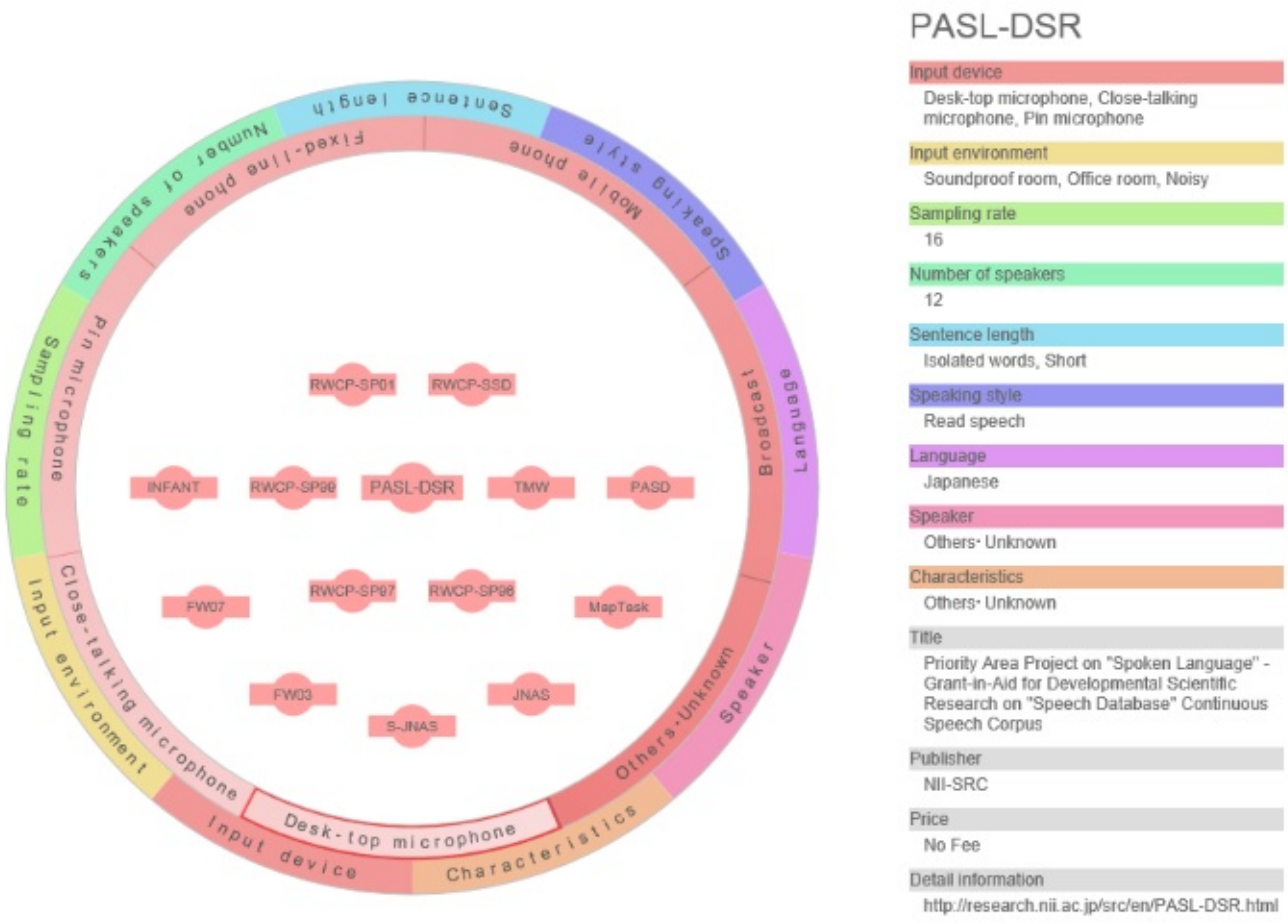


Fig. 2　Screenshot of Concentric Ring View (CRV) System. The outermost ring shows the attributes and the inner ring indicates the items of the "input device" attribute. The corpora searched for by specifying "desk-top microphone" for the "input-device" attribute are shown inside the rings. All the attributes and items of the corpus located in the center of the ring, i.e., PASL-DSR in this case, are automatically shown on the right side. The details of a given corpus are displayed on the right side by clicking on the desired corpus shown within the ring.

Table 2 New set of attributes and items for speech corpora specifications.

| Attribute | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Input device | Desk-top microphone | Close-talking microphone | Lapel microphone | Fixed-line phone | Mobile phone | Broadcast | Others/ Unknown |
| Input environment | Sound-proof room | Office room | Noisy | In-car | Others/ Unknown | | |
| Sampling rate | SR < 10 kHz | SR < 20 kHz | 20 kHz ≦ SR | Unknown | | | |
| Number of speakers | No < 10 | No < 100 | No < 1000 | 1000 ≦ No | Unknown | | |
| Sentence length | Isolated words | Short | Long | Others/ Unknown | | | |
| Speaking style | Read speech | Acted speech | Spontaneous speech | Others/ Unknown | | | |
| Language | Japan | Asia | Europe | Africa | America | Oceania | Others/ Unknown |
| Speaker | Non-native | Professional | Child | Senior | Others/ Unknown | | |
| Characteristics | Multilingual | Dialect | Dialogue | Emotional | Non-speech | Others/ Unknown | |


Fig. 3 Initial screen shot showing only attribute ring.

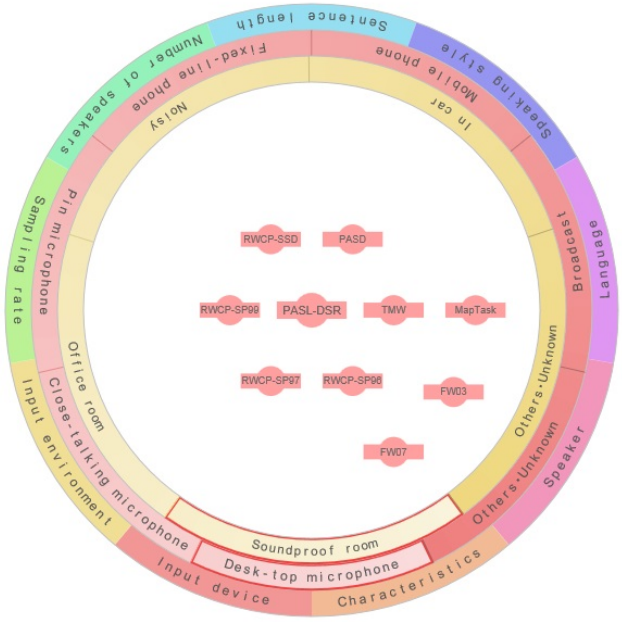


Fig. 4 Screen shot of CRV displaying searched corpora specifying "desk-top microphone" for "input-device" attribute and "sound-proof room" for "input environment" attribute.

## 5. EXPERIMENT AND RESULTS

The proposed attributes consists of nine groups, as specified in Table 2: input device, input environment, sampling rate, number of speakers, sentences length, speaking style, language, speaker attribute, and characteristics. We selected a few items for each attribute, resulting in 50 items in all. Each item takes on a "1" or "0" depending on whether the corpus has the attribute or not. Each attribute has an "others/unknown" item to cope with cases where a suitable attribute is not found in the corpus catalogue.

We have tentatively developed software for extracting the corpus metadata from SHACHI and transferring them to the CRV system. We have presently analyzed the speech corpora distributed by the Speech Resources Consortium (NII-SRC) [2], which includes 43 corpora in all. Figure 2 shows a screenshot of the Concentric Ring View (CRV) system. Initially, the outermost ring is displayed on the screen as shown in Fig. 3 indicating there are nine attributes. By clicking, for example, the input device attribute, another ring appears inside showing seven items such as a desk-top microphone, close-talking microphone, etc; by dragging this "input device" attribute ring, you can rotate this ring and place the desk-top microphone sector at the bottom; then this sector becomes enclosed in a bright contour and then the corpora recorded by desk-top microphones are displayed inside the rings as shown in Fig. 2. By clicking the "input environment" sector, another item ring appears inside. Figure 4 shows the searched corpora recorded by selecting the "desk-top microphone" in a "sound-proof room". By specifying more attributes, other rings appear further inside the circle, and the displayed search results are thus narrowed down. All the attributes and items of the corpus located in the center of the ring, i.e., PASL-DSR in this case, are automatically shown on the right side. By clicking any corpus in the ring, the details of that corpus are displayed on the right side. In this way the system searches the desired corpora and visually presents them through a series of simple operations using the mouse.

## 6. CONCLUSION

We discussed the importance of standardizing corpus specification attributes and items to enable corpus users' easy access to the speech corpora catalogue. We have proposed a revised set of attributes after a few years of testing. We think the proposed specification attributes and items are language-independent that can also be applicable to the classification of text-corpora and images, etc. by changing or adding/removing some items, if necessary. We have also shown the effectiveness of the Concentric Ring View (CRV) system, which simultaneously performs the search for and display of speech corpora. We are developing a corpus search and visualization system for searching for speech corpora for users in various research fields for use as a web application system.

## REFERENCES


[1] S. Itahashi, T. Kajiyama, K. Yamakawa, Y. Ishimoto, and T. Matsui, "Interactive visualization search system for speech corpora," Proc. Oriental COCOSDA 2011, pp. 157-161, Hsinchu, Taiwan (2011).

[2] S. Itahashi and T. Ohsuga, "Introduction of NII-Speech Resources Consortium," Proc. Oriental COCOSDA 2006, pp. 38-43, Penang, Malaysia (2006).

[3] S. Itahashi, T. Ohsuga, and H. Kojima, "A proposal for standardizing metadata specifications of speech corpora," (in Japanese) Preprints, Fall Meeting of Acous. Soc. Jpn., Paper 1-P-4, pp. 389-390 (2013).

[4] S. Itahashi, K. Yamakawa, T. Matsui, and Y. Ishimoto, "A proposal for standardizing catalogue specifications of speech corpora," Proc. Oriental COCOSDA 2010, 5 pages, Kathmandu, Nepal (2010).

[5] T. Kajiyama and S. Satoh, "Construction of image retrieval systems focused on user knowledge interaction," Proc. ACM Multimedia 2010, pp. 1673-1676, Firenze, Italy (2010).

[6] T. Kajiyama, K. Nakamaru, Y. Ohno, and N. Kando, "Concentric Ring View: An interactive environment for integrating searching and browsing," Proc. Joint International Conference on Soft Computing and Intelligent Systems, 6 pages (CD), Yokohama, Japan (2004).

[7] H. Tohyama, S. Kozawa, K. Uchimoto, S. Matsubara, and H. Isahara, "Construction of a metadata database for efficient development and use of language resources," Proc. LREC 2008, pp. 1687-1692, Marrakech, Morocco (2008).

[8] K. Yamakawa, T. Matsui, and S. Itahashi, "MDS-based visualization method for multiple speech corpora," Proc. Interspeech 2008, pp. 1666-1669, Brisbane, Australia (2008).

[9] OLAC, http://www.language-archives.org/

[10] Dublin Core, http://dublincore.org/