INTERACTIVE VISUALIZATION AND SEARCH SYSTEM FOR SPEECH CORPORA

Shuichi Itahashi[1,5], Tomoko Kajiyama[2], Kimiko Yamakawa[3], Yuichi Ishimoto[1], and Tomoko Matsui[4]

[1]National Institute of Informatics, Tokyo, Japan, [2]Aoyama Gakuin University, Kanagawa, Japan,
[3]Aichi Shukutoku University, Aichi, Japan, [4]Institute of Statistical Mathematics, Tokyo, Japan,
[5]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

ABSTRACT

We have already reported a corpus similarity visualization method based on the corpus attribute using multidimensional scaling that makes it easy for users to utilize various speech corpora. In this paper, we present a revised visualization method that is based on a ring structure like a planisphere. By using only a mouse, a user can choose appropriate search keys for each of the multiple attributes and can easily filter information by adjusting the keys. Retrieved results are displayed inside the rings, and the user can filter and browse them in real time. This will facilitate efficient searching of the specific corpus that fits user's needs.

Index Terms— *corpus, information, retrieval, specification*

1. INTRODUCTION

Computer performance has greatly increased, and it has become possible to process large amounts of speech data on small computers. Moreover, a great variety of speech corpora have been developed in many countries. Although such diversity gives the user more freedom of choice, users have to select a suitable corpus for the intended purpose from a great variety of corpora.

Several data centers have been set up in the U.S.A., Europe, and Asia to supply various speech corpora. They include the Linguistic Data Consortium (LDC) in the U.S.A., the European Language Resources Association (ELRA), the Chinese LDC / Chinese Corpus Consortium (CCC), the Speech Information Technology & Industry Promotion Center (SiTEC) in Korea, and the Speech Resources Consortium at the National Institute of Informatics (NII-SRC) in Japan. However, the specifications of their corpora vary from center to center. It is difficult for corpus users to select suitable corpora to fit for their purpose. Recently, several initiatives have sought to extend their resource catalogue in new ways. For instance, ELRA's Universal Catalogue [1] and NICT Shachi catalogue [2] both intend to serve as union catalogues like OLAC [3].

Visualizing similarities among various corpora will help users to select suitable speech corpora. In order for users to easily utilize these corpora, we proposed a similarity visualization method based on corpus attributes and multidimensional scaling (MDS) [4-6].

In this paper, we propose a revised method of corpus visualization. It is based on a ring structure like a planisphere. By using only a mouse, users can choose appropriate search keys for each of the attributes, and they can easily filter information by adjusting the keys. Retrieved results are displayed inside the rings, and users can filter and browse them in real time. We incorporate nine groups of attributes, including input devices, input environments, number of speakers, speaking styles, speech modes, data modes, languages, purposes, and S/N ratio [4-6]. Each attribute has a few items, resulting in 61 items in all.

In the following, we discuss the corpus attributes and visualization method of speech corpora in Section 2. Section 3 describes the actual corpus specifications and speech corpora used for the experiment and the results. Section 4 is the conclusion.

2. VISUALIZATION METHOD

2.1. Corpus attributes and items

It would be convenient for corpus users to select suitable corpora for their use if similarities among various speech corpora were able to be visualized. Although there are various metadata sets proposed for universal cataloging of language resources [7], it is difficult to find those suitable for speech resources. Speech corpora can be classified in a number of ways such as by input device, input environment, speaking style, data mode, language, and purpose of corpus utilization [8]. The input device specifies the input apparatus used for recording the speech corpora, such as microphone, telephone, and broadcast. The input environment specifies whether the speech was recorded in a sound-proof room, ordinary office, or noisy environment. The speaking style indicates whether the input is continuous speech or isolated words. The speech mode describes whether the input is read speech, dialogue, or meeting speech. Data mode specifies if the recorded data is speech wave, the analyzed parameters, electroglottogram (EGG), palatogram, or multimodal corpus. 'Language' specifies whether the corpus is a monolingual corpus, bilingual corpus, or dialect. 'Purpose' specifies the purpose of corpus utilization, e.g., speech analysis, synthesis, recognition, speaker identification, and language identification. S/N ratio refers to signal to noise ratio of recorded speech. At present, each data center uses its own specification system of corpora. It is desirable for all data

978-1-4577-0931-9/11/$26.00 ©2011 IEEE

centers to use the same specification system from the users' point of view.

2.2. Concentric Ring View (CRV) system

This is an interactive environment for integrating searching and browsing corpora [9, 10]. It is composed of several concentric rings, each of which corresponds to an attribute of speech corpora (see Fig. 1). The outermost ring expresses the attributes of corpora as shown in Fig. 2. 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 items that correspond to the attribute category specified on the attribute ring. By rotating the item ring, a user can adjust the item and browse filtered results. By specifying more attributes, which cause other rings to appear inside, displayed information can be narrowed down. The current item value is always shown at the bottom of each ring, so that the user can easily check the current position. This technique allows users to specify search items precisely.

A user may not find a suitable corpus because the specified items are slightly different from the attributes the corpus has. Fuzzy matching is necessary to overcome this problem. We implemented this function as follows, the number of candidates decreases naturally as item rings are added; therefore, the breadth of item values depends on the number of item rings if the item rings represent continuous quantities. For example, when a user specifies only one item ring, the attribute values in a three degree range from the bottom of the ring are used as search items. If the user adds another item ring, search items become ones with a six degree range. The system can perform vague filtering automatically, without adjustment by the user. For example, if the "number of speakers" attribute specifies around 300, then 5 corpora will be displayed such as TIDIGITS, CIAIR-VCV, ATR DB-C, CENSREC-3, and JNAS having the number of speakers between 288 and 326; if we add "desk-top microphone" as an item of the second attribute "input device", then 3 corpora will be displayed including CIAIR-VCV, ATR-DB-C, and S-JNAS, which shows that "number of speakers" condition around 300 is relaxed by adding another attribute ring and S-JNAS spoken by 402 speakers are also retrieved. If an item ring represents discrete data, the item value is always the bottom point of the ring to minimize confusion for the user.

2.3. Extraction of retrieved corpora

The retrieval process extracts information that fits the retrieval target and ranks each retrieved results according to attribute value and priority. The item at the bottom of the ring is used for retrieval; the procedure for extracting the suitable information set is as follows [10].

1) If the attribute value is discrete, extract suitable information by using the retrieval item.
2) If the attribute value is continuous, extract suitable information by referring to the nearest value within a certain range from the retrieval item.
3) If the number of selected attributes becomes large, widen the range of the value used for extracting suitable information.

To be able to understand the retrieval items intuitively, only the attribute values at the bottom part of the item ring are used when the attribute value is discrete.

If the attribute value is a continuous one that changes subtly, we extract suitable information having the closest value to the retrieval item within a certain range. We have to adjust the retrieval item suitably, as the number of retrieved corpora decreases if we add more attributes. Therefore, in the case of a continuous attribute value, we can try to loosen the narrowing-down process by widening the range of values used for extracting suitable information as the number of selected attributes increases.

The retrieved corpora are displayed as follows; start from the center of the inner circle, go clockwise from the direction of 3 o'clock to the direction of 5, 7, 9, … o'clock, and then go to the next (outer) circle. The retrieved corpora are displayed inside the ring. If the retrieval conditions include discrete attributes only, then the corpora are displayed according to the order of entry of the corpus to the system. If the retrieval conditions include continuous attributes, then the retrieved results are displayed according to the order of wgt value starting from the corpus with the smallest value.

2.3.1. Discrete attribute value case

Most of the attribute items in Table 1 are discrete data; only two attributes, the number of speakers, and S/N ratio are continuous data. In case of discrete attributes, each data has one or more attribute values. A database is built by applying each attribute values to one bit and creating a bit sequence to represent the attribute values of each data.

Let the number of item rings displayed be denoted by r_n, and the value of the item at the bottom of the j-th ring R_j from the outermost ring be k_j. Furthemore, let the item value for item ring R_j of an attribute I_i be d_{ij}. The degree of fitness of k_i and d_{ij} is calculated as follows.

$$wgt_i = \sum_{j=1}^{r_n} S_j \text{ , where } S_j=1 \text{ when } k_j \ \& \ d_{ij}=k_j, \text{ or } 0, \text{ otherwise.}$$

Here '&' denotes the 'and' operation. I_i is the retrieved result when $wgt_i=r_n$. For example, we have four items for the attribute "speaking style," i.e. continuous speech, isolated words, non-native speaker, and unknown as shown in Table 1. Corpus #1, PASL-DSR, includes both continuous speech and isolated words items. In the case only "speaking style" attribute ring is displayed, corpus #1 will be retrieved by specifying either continuous speech or isolated words as we will get wgt=rn=1 in both cases.

2.3.2. Continuous attribute value case

Each data has only one attribute value in each continuous attribute. The symbols are the same as described above. The degree of fitness of k_i and d_{ij} is

$$wgt_i = \sum_{j=1}^{r_n} | k_j - d_{ij} |$$

I_i is the retrieved result when $wgt_i<x$, where x is the threshold of suitability.

Table 1 Attributes and items of the corpora used for similarity analysis. '#' indicates the number of items in each attribute.

Attribute	#	Item
Input device	7	desk-top mic, close-talking mic, pin-mic, fixed-line telephone, mobile phone, broadcast, unknown.
Input environment	5	sound-proof room, office room, noisy condition, in-car, unknown
Number of speakers	10	male (<10, <100, >=100), female (<10, <100, >= 100), total number (<10, <100, >=100), unknown.
Speaking style	4	continuous speech, isolated words, non-native speaker, unknown.
Speech mode	5	dialogue, read speech, meeting speech, lecture speech, unknown.
Data mode	9	sampling freq. (<8kHz, <16kHz, >16kHz), analysis parameter, multimodal data, electromyogram, palatogram, MRI image, unknown.
Language	4	Monolingual, multilingual, dialect, unknown.
Purpose	14	Analysis, recognition, synthesis, digit recognition, speaker recognition, language recognition, aged person's speech, children's speech, robust recognition, speaker-independent, non-native speech, multimodal, non-speech sound (noise), unknown.
S/N ratio	3	mean, standard deviation, unknown

3. EXPERIMENT

3.1. Corpus specifications

We listed possible attributes showing the speech corpora features based on the classification in Reference [6]. The attributes consisted of the nine groups listed in Table 1: input device, input environment, number of speakers, speaking style, speech mode, data mode, language, purpose, and S/N ratio. We selected a few items for each attribute, resulting in 61 items in all. Table 1 shows the corpus attributes. Each item took '1' or '0' depending on whether the corpus had the attribute or not. However, "the number of speakers" and S/N ratio attributes were continuous data. Each attribute had an item "unknown" to cope with cases in which we could not find a suitable item in the corpus catalogue. The corpus feature was represented as a set of 61 attribute items.

We analyzed the speech corpora distributed by the Speech Resources Consortium (NII-SRC) [11], and some domestic and overseas data centers. Table 2 lists the corpora used in this study. PASL-DSR (#1) and ASJ-JIPDEC (#28) are continuous speech corpora. UT-ML (#2) is a multilingual corpus. GSR-JD (#4) and Tsuruoka91 (#22) are Japanese dialect corpora. The spoken dialogue corpora include RWCP-SP96 (#5), RWCP-SP97 (#6), RWCP-SP01 (#8), PASD (#10), Map Task (#19), and UUDB (#20). In particular, RWCP-SP01 (#8) is a meetings dialogue corpus. RWCP-SP99 (#7) is a speech corpus of broadcast news articles read by professional announcers. RWCP-SSD (#9) contains various sound and noise data. CIAIR-VCV (#11) is a children's voice corpus, and NTT Infant (#27) is an infant voice corpus. The CENSREC series corpora (#12 through #16) are for noisy speech recognition environments. UME-ERJ (#17) is English speech read by Japanese students, while UME-JRF (#18) is Japanese speech read by overseas' students. JNAS (#23) and S-JNAS (#24) are read speech of sentences from Japanese newspaper articles; #24 is read by aged people. TMW (#3), ETL-WD (#21), FW03 (#25), and FW07 (#26) are isolated word corpora. CSJ (#29) is the largest spoken Japanese corpus. ATRDB-A (#30), C (#32), E (#34) and DIC (#40) contain words and sentences. ATRDB-B (#31) and BLA (#39) contain phonetically balanced sentences. ATRDB-D (#33) and F (#35) contain read sentences. ATRDB-SDB (#36), SLDB (#37) and APP (#38) contain played dialogues. ETL-TGT (#41) contains town-guide task sentences. Corpora #42 through 45 are those of LDC. Corpora #46 and #47 are from ELRA. Corpus #48 is from Chinese LDC, #49 from CCC of China, and #50 from SiTEC of Korea.

3.2. Results

Figures 1, 3 and 4 show three examples of retrieval results. Figure 1 illustrates 22 corpora retrieved by specifying 'close-talking microphone' for the input device attribute. By rotating the input device ring, a user can place the desk-top microphone at the bottom and the corpora recorded by desk-top microphone will be displayed inside the ring, as shown in Fig. 3. By clicking the input environment sector, another item ring appears inside. Figure 4 shows the corpora recorded with a close-talking microphone in a sound-proof room. By specifying more attributes, other rings appear inside, and the displayed information is narrowed down. In this way, the system retrieves the desired corpora and shows them visually through a series of simple operations.

Regarding the evaluation of the system, image retrieval test was performed in reference [9]. It showed that the proposed ring structure is effective for a search based on the ambiguous information, but we need the evaluation of the present system in the future.

4. CONCLUSIONS

This paper showed the effectiveness of the Concentric Ring View system that integrates searching and browsing speech corpora. Candidates can be browsed while key items are adjusted; thus, fine adjustment of the key item is simple. This technique treats various data uniformly: i.e., it uniformly deals with discrete data and continuous data. The ring structure can treat attributes of three or higher dimensions. A user can grasp the current search condition easily because the selected key

items can always be seen at the bottom of the displayed rings. The calculation process is simple, and the system works quite fast. This system can also be applicable to visualizing text corpora. The present searching and browsing system based on references [4-6] can be accessible at the following URL: http://corpus-search.nii.ac.jp/?lang=en

5. REFERENCES

[1] ELRA: http://www.elra.info/

[2] Shachi: http://www2.shachi.org/index.php?lang=english

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

[4] K. Yamakawa, T. Matsui, and S. Itahashi, "Visualization of Various Speech Corpora by Multidimensional Scaling," Proc. Oriental COCOSDA 2007, pp. 112-115, Hanoi, Vietnam (Dec. 2007).

[5] K. Yamakawa, T. Matsui, and S. Itahashi, "MDS-based Visualization Method for Multiple Speech Corpora," Proc. Interspeech 2008, pp. 1666-1669, Brisbane, Australia (Sep. 23-26, 2008).

[6] K. Yamakawa, T. Matsui, H. Kikuchi, and S. Itahashi, "Utilization of acoustical feature in visualization of multiple speech corpora," Proc. Oriental COCOSDA 2009, pp. 135-139, Beijing, China (Aug. 2009).

[7] C. Cieri and K. Choukri, "From roadmaps to plans: Towards the design and cost-benefit analysis of a universal language resources catalogue," Preprint, FlaReNet Forum 2010, Barcelona, 5 pages (Feb. 11-12 , 2010).

[8] S. Itahashi, K. Yamakawa, T. Matsui, and Y. Ishimoto, "A proposal for standardizing catalogue specifications of speech corpora," Proc. Oriental COCOSDA 2010 (CD-ROM), 5 pages, Kathmandu, Nepal (Nov. 24-25, 2010).

[9] 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, and 5th International Symposium on Advanced Intelligent Systems (CD), 6pages, Yokohama, Japan (Sep.2004).

[10] T. Kajiyama and S. Satoh, "Construction of image retrieval systems focused on user knowledge interaction," Proc.ACM Multimedia2010, pp. 1673-1676, Firenze, Italy (Oct. 2010).

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

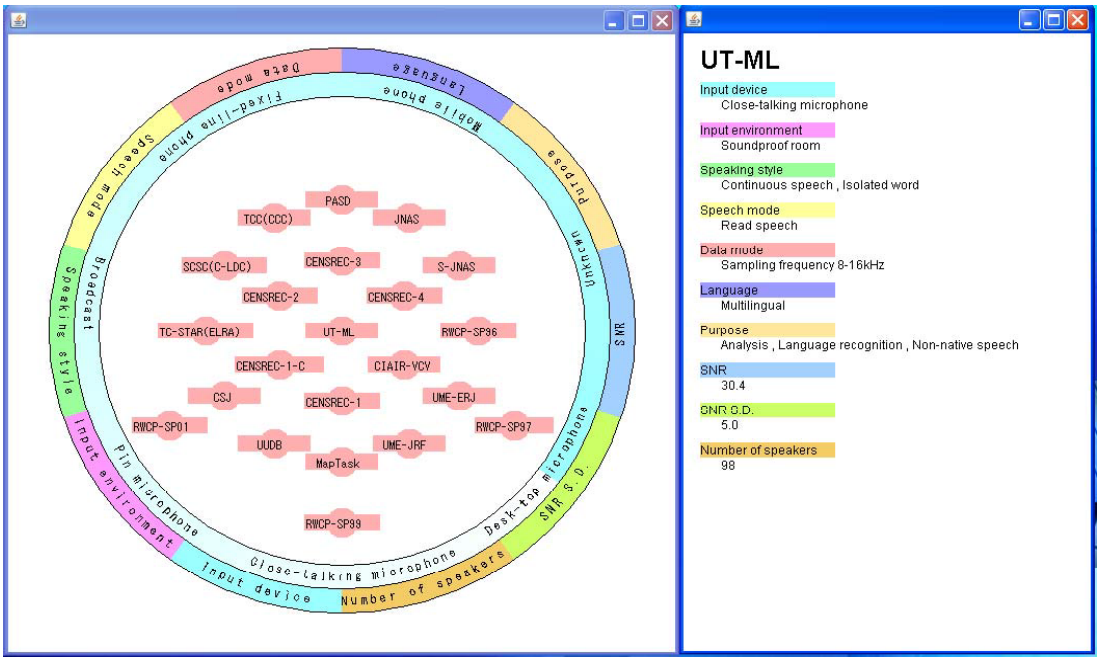

Figure 1 A screenshot of our system's 'Concentric Ring View'. The outermost ring shows the attributes, and the inner ring indicates the items of the "input device" attribute. Corpora retrieved by specifying "close-talking microphone" for the "input device" attribute are shown inside the rings. All attributes and attribute items of the corpus located in the center of the ring are automatically shown in the right window. By clicking any corpus shown in the circle, details are displayed in the right window .

Figure 2 Initial screenshot showing only the attribute ring.

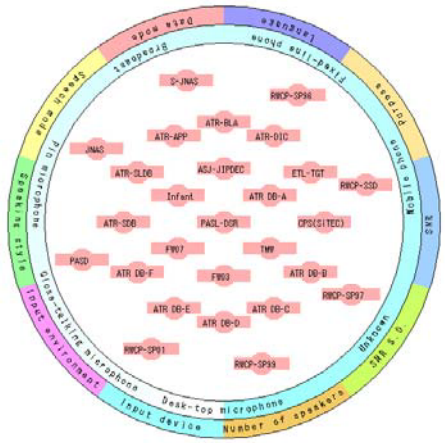

Figure 3 Screenshot of displaying retrieved corpora by specifying “desk-top” microphone for the “input device” attribute.

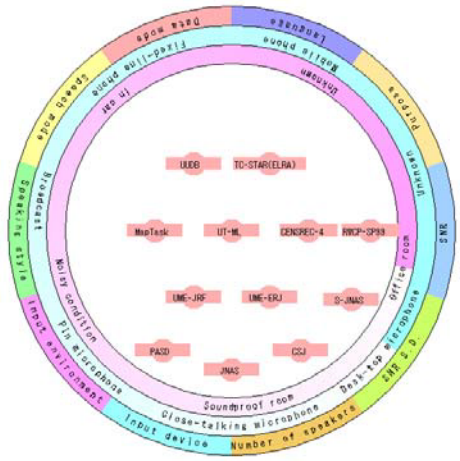

Figure 4 Screenshot of displaying retrieved corpora specifying “close-talking microphone for the “input device” attribute and “sound-proof room” for the “input environment” attribute.

Table 2 Corpora used for the analysis

#	Corpus name	#	Corpus name
1	PASL-DSR	26	FW07
2	UT-ML	27	NTT Infant
3	TMW	28	ASJ-JIPDEC
4	GSR-JD	29	CSJ
5	RWCP-SP96	30	ATR-DB-A
6	RWCP-SP97	31	ATR-DB-B
7	RWCP-SP99	32	ATR-DB-C
8	RWCP-SP01	33	ATR-DB-D
9	RWCP-SSD	34	ATR-DB-E
10	PASD	35	ATR-DB-F
11	CIAIR-VCV	36	ATR-SDB
12	CENSREC-1	37	ATR-SLDB
13	CENSREC-1-C	38	ATR-APP
14	CENSREC-2	39	ATR-BLA
15	CENSREC-3	40	ATR-DIC
16	CENSREC-4	41	ETL-TGT
17	UME-ERJ	42	TIMIT(LDC)
18	UME-JRF	43	TIDIGITS(LDC)
19	Map Task	44	NTIMIT(LDC)
20	UUDB	45	YOHO(LDC)
21	ETL-WD	46	TC-STAR(ELRA)
22	Tsuruoka91	47	O-French(ELRA)
23	JNAS	48	SCSC(C-LDC)
24	S-JNAS	49	TCC(CCC)
25	FW03	50	CPS(SiTEC)