Automatic Editing of Spoken Document for Intelligent Speech Archive

Masashi Ito[1], Tomohiro Ohno[2], Shigeki Matsubara[3]
[1]Graduate School of Information Science, Nagoya University
[2]Graduate School of International Development, Nagoya University
[3]Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, 464-8601, Japan
[1]email: masashi@el.itc.nagoya-u.ac.jp

***Abstract*—As typified by World Wide Web, a lot of information became accumulated on the Internet. However, most of the currently distributed information is occupied by written document. Compared with it, spoken document is hardly distributed. Therefore, if the mechanism for distributing them can be built, our human society will be able to share much more information. This paper proposes a technique for editing a sentence in spoken document for the purpose of converting it into the Internet contents equipped with the accessibility and readability. By aligning the recorded video data or speech data with the edited text on a fine level, it can be utilized as the multimedia contents equipped with the accessibility. Our technique consists of the following three sentence technologies: (1) paraphrase, (2) division, and (3) structuration. We implemented a spoken document edit system based on our techniques. We conducted an edit experiment by using lecture speech data and our technique could achieve high accuracy. From the results, we confirmed the availability of our technique.**

I. Introduction

As typified by World Wide Web, a lot of information became accumulated on the Internet and the information produced by human beings became efficiently consumed. One of the key technologies of such information distribution is search technology, and a large amount of texts have been the targets of search. However, most of the currently distributed information is occupied by "written documents". Compared with it, "spoken documents" is hardly distributed. The routinely produced linguistic information is overwhelmingly produced by speaking rather than by writing. Therefore, if the mechanism for distributing spoken documents can be built as well as written documents, our human society will be able to share much more information.

Although we can consider various forms to distribute spoken documents, the form that speech data is only uploaded to the Internet is not suitable enough to reuse them. In order to understand the content of a speech, we have to listen to the speech data and the listening is inefficient in obtaining information. Considering the accessibility of search, it is desirable for the transcribed text data to contain not only speech data but also transcribed text data. Moreover, considering the readability in browsing, it is necessary for the distributed data to be edited into the text which is easy to read.

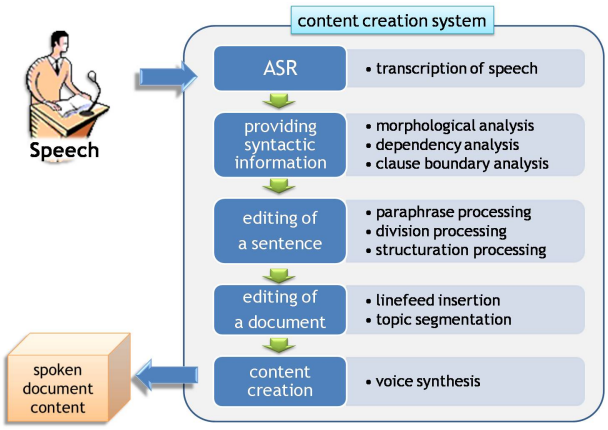

Fig. 1. *Our speech archiving system*

This paper proposes a technique for editing a sentence in spoken documents for the purpose of converting it into the Web contents equipped with the accessibility and readability. The proposal technique consists of the following three sentence technologies: (1) paraphrase, (2) division, and (3) structuration. By these processing, a spoken document can be accumulated as a sharable document which has excellent readability in browsing. In addition, by aligning the recorded video data or speech data with the edited text in detail, it can be utilized as the multimedia contents equipped with the accessibility.

We implemented a spoken document edit system based on our techniques. We conducted an edit experiment by using lecture speech data and our technique could achieve high accuracy. From the results, we confirmed the availability of our technique.

II. Content Creation of Spoken Documents

Figure 1 shows the composition of the content creation system of spoken documents. Our system receives speech data as input, and converts it into a text by ASR, and then provides the syntactic information to the text by morphological analysis, dependency analysis, and clause boundary analysis. By applying our sentence edit technology to the text to which

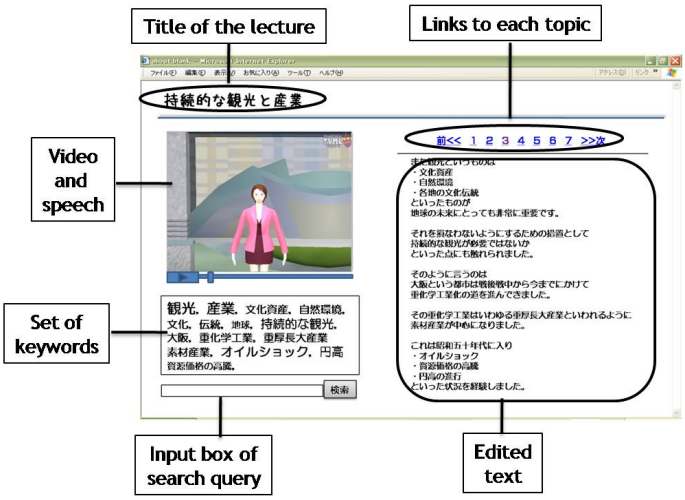

Fig. 2. *Outline of the speech contents on the Web*

the syntactic information was provided, our system converts it into the text which is easy to read. Furthermore, linefeed insertion and topic segmentation are performed to the whole spoken document, and finally a synthesized voice is generated by speech synthesis. It also becomes easy to access directly to the partial spoken document, by aligning the edited text, the synthesized voice, and the recorded video in a unit such as a sentence.

Figure 2 shows the browsing environment of the spoken document content. This environment has the following features.

- **Providing a simulated speech and a simulated video**
  These prevent the outflow of the personal information by which a speaker can be identified.
- **Showing of a set of keywords**
  This set of keywords enables broadly viewing the content of a spoken document.
- **Segmentation by a topic**
  This segmentation enables topic-by-topic browsing and is suitable for selective reuse.
- **Editing of spoken document**
  This editing enables a spoken document to be easy to read and understand.

III. Related Works

In this section, we describes the related works on content creation of speech data and editing of a spoken language sentence.

A. Content creation of speech data

Although some trials of sharing speech data have already been performed, the following problem has been widely recognized[4]. That is, when spoken documents are simply accumulated without editing, its shared data is not easy to browse and is not convenient for a user to access. In response, Lee et al. have actualized the speech archiving based on the summarization and structuration for news[5]. In this research, they have generated the titles and summaries for news speech texts by extracting the important parts from them, and have actualized the structuration of a news database by analyzing the relations between news. Moreover, as the similar research, Nakagawa et al. have developed the content creation system based on the sounds and images of the lecture in which the slides were used[1]. By recognition, summarization, division and indexing of lecture speech, the multimedia contents containing images, sounds and texts are created. On the other hand, the target of our research is the general lecture speech in which news manuscripts or slide data are not used and, thus, our approach for content creation is different from those of the above-mentioned researches.

B. Editing of spoken texts

As a technology for transforming a text into the text which is easy to read, some summarization techniques have been developed for dialogue speech[6], [7] or news speech[8], [9]. On the other hand, our research targets the general lecture speech.

Moreover, Yamamoto et al. have proposed summarization technique for the Diet minutes[2]. In this technique, parenthetical expressions are deleted, redundant expressions peculiar to spoken language are deleted or paraphrased, and honorific expressions are converted into normal expressions. This technique is relevant to our research in the sense that spoken language is converted into the text which is easy to read.

Furthermore, there exist some previous researches about a technology for dividing a sentence. Kim et al. have divided a long sentence for improving the accuracy of machine translation[10]. Hayashi has implemented the support system for polishing sentences in the technical documents, and the system has introduced a technology for dividing a sentence[3]. The technology identifies the splittable points and semantic relations between the sentences generated after the division, for complex sentences including "*renyou-tyuushi*" expressions or conjunctions.

IV. Editing of Spoken Language Sentence

Figure 3 shows an example of a transcribed text of lecture speech and its edited text. Since the transcribed text includes the features of spoken language, if the transcribed text is simply displayed, it is not easy to read. In contrast, the edited text is suitable to quickly understand the whole content. By reducing features of spoken language in the transcribed text and specifying the structure included in the text, we can generate the text excellent in the readability of browsing.

Our technique realizes editing of a sentence by performing paraphrase processing, division processing, and structuration processing in sequence. Figure 4 shows an example of sentence edit processing. The following processing is respectively performed.

1) Paraphrase processing
   Redundant expressions in an input sentence are deleted.

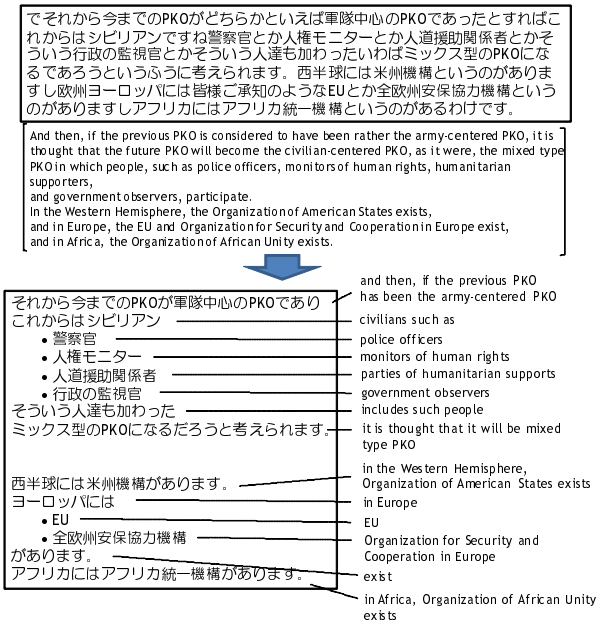

Fig. 3. *Example of a transcribed text and its edited text*

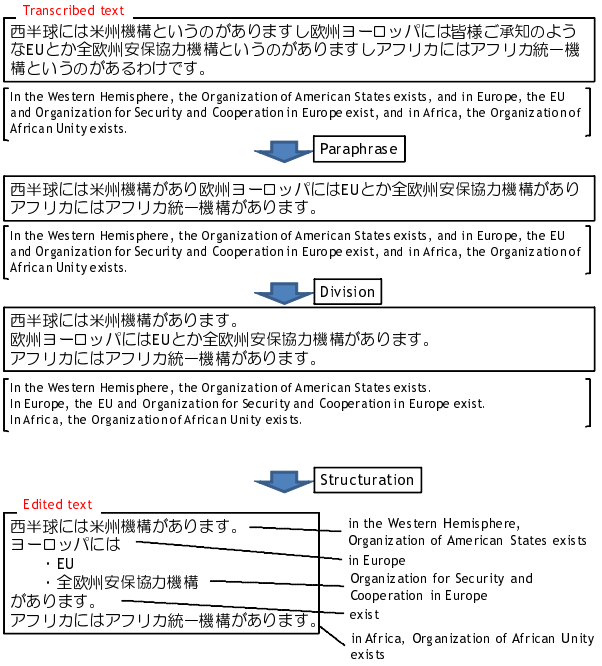

Fig. 4. *Example of sentence editing processing*

Furthermore, spoken language expressions are converted into written language expressions.

2) Division processing
Divisible points in an input sentence are detected, and the sentence are divided into two or more sentences. Moreover, the end of each sentence generated by the division are paraphrased into the proper expressions.

3) Structuration processing
Parallel structures in an input sentence are detected.

TABLE I
RESULT OF ANALYSIS ABOUT PARAPHRASE

paraphrased expression	number of times
Polite expressions expressions includeing “*desu*” or “*masu*”	112
Spoken language expressions peculiar to spoken language	51
Honorific verb verbs which are used as honorific expressions	32
Collocation “*toyu*” expressions including “*toyu*”	26
Completement and correction completement of missing particles or correction of grammar errors	27
Prefix deletion of a prefix	14
Conjunction deletion of a redundant conjunction	14
Demonstrative deletion of a redundant demonstrative	13
renyou-tyuushi bunsetsu paraphrase of *renyou-tyuushi bunsetsu*	13
others	27
total	329

A. *Paraphrase processing*

In order to create the paraphrase rule for automatic sentence editing, we paraphrased the transcribed text of lecture speech by hand. We used 100 sentences extracted from seven lectures of which speaker is different. A total of 329 paraphrases were performed. Table I shows the breakdown of the paraphrased expressions. The examples of the paraphrased expressions are as follows.

- Polite expressions
  しておりました (did) ⇒ していた (did)
  shi-te ori-mashi-ta ⇒ *shi-te i-ta*
- Spoken language
  おっきい (big) ⇒ おおきい (big)
  okkii ⇒ *ookii*
- Honorific verb
  させて頂く (do) ⇒ する (do)
  sa-se-te itadaku ⇒ *suru*
- Collocation “toyu”
  正常化ということが行われた (perform the normalization) ⇒ 正常化が行われた (perform the normalization)
  seijouka-toiu-koto-ga okonawa-re-ta ⇒ *seijouka-ga okonawa-re-ta*
- Complement and correction
  私普段から (I usually)⇒ 私は普段から (I usually)
  watashi fudan-kara ⇒ *watashi-ha fudan-kara*
- Prefix
  お料理を食べる (eat foods)⇒ 料理を食べる (eat foods)
  o-ryouri-wo taberu ⇒ *ryouri-wo taberu*

Especially, “polite expressions”, “spoken language”, “honorific verb”, and “collocation ‘*toyu*”’ included many redundant expressions. Since these expressions appear frequently, paraphrasing them enables the features of spoken language in the whole sentence to be efficiently reduced. In other

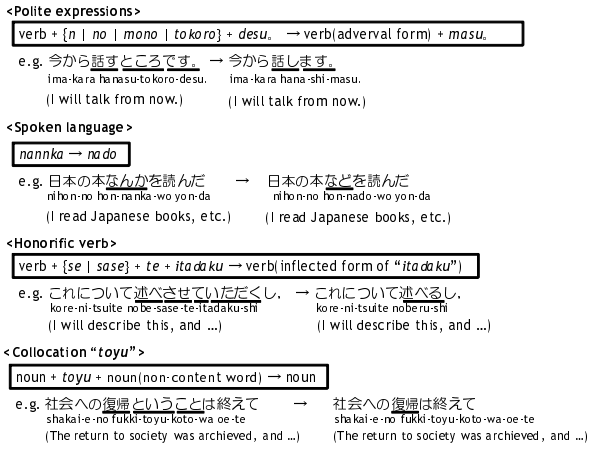

Fig. 5. *Examples of the paraphrase rules and its application*

expressions, the paraphrases were performed in consideration of the sense of a sentence, therefore, it is difficult to create rules based on the surface information. From these reasons, we created paraphrase rules for the expressions of the four above-mentioned types. Figure 5 shows examples of the created rules and its application.

B. Division processing

For analysis about the division processing of a sentence, we divided a sentence in the transcribed text of lecture speech by hand. Size of the created data is 4 lectures consisting of 468 sentences. In the data, the number of times by which a sentence was divided was 299 times. About 89% of sentence divisions are performed just after conjunctive particle *"keredomo", "ga", "node", "shi", "to", "kara", or renyou-tyuushi bunsetsu* [1] . Here, *renyou-tyuushi* bunsetsu is a bunsetsu of which the morphological information is one of the following two cases.

- The inflected form of the rightmost morpheme of the *bunsetsu* is adverbial form.
  e.g. "食べ" *tabe*, "訳で" *wakede*
- The second rightmost morpheme and the rightmost morpheme of the bunsetsu are adverbial form and conjunctive particle "*te*," respectively.
  e.g. "食べ" *tabe-te*, "書いて" *kai-te*

The above-mentioned conjunctive particles and *renyou-tyuushi* bunsetsu become a predicate and, thus, they are considered to be a strong break semantically. Table II shows the appearance frequency of the above-mentioned conjunctive particles and *renyou-tyushi* bunsetsu and the number of times by which a sentence was divided at each point. In our research, the end boundary of the above-mentioned conjunctive particle or "*renyou-tyuushi*" bunsetsu is defined as a **candidate of division points**. However, a sentence can not necessarily be divided at all the candidates of division points. In some cases, the meaning of the original sentence may be lost by dividing at the candidate. Therefore, based on dependency relations, we established the conditions on which a sentence could be divided.

[1] *Bunsetsu* is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and zero or more ancillary words. A *dependency* is a modification relation in which a *modifier bunsetsu* depends on a *modified bunsetsu*. That is, the modifier bunsetsu and the modified bunsetsu work as modifier and modifyee, respectively.

TABLE II
RESULT OF ANALYSIS ABOUT SENTENCE DIVISION

	number of times of division	number of times of appearance
conjunctive particle "*keredomo*"	75	84
conjunctive particle "*ga*"	68	82
conjunctive particle "*node*"	25	30
conjunctive particle "*shi*"	12	12
conjunctive particle "*to*"	10	75
conjunctive particle "*kara*"	4	7
renyou-tyuushi bunsetsu	72	271
others	33	-
total	299	-

- Bunsetsus located before the candidate of division points (except the bunsetsu located just before the point) do not depend on a bunsetsu located after the point.

Figure 6 shows an example of a sentence which fulfills this condition. A sequence from the start of the sentence to the candidate of division points constitutes a independent clause, which basically contains one verb phrase. That is, in this example, it can be said that two simple sentences are connected. Therefore, the sentence can be divided into two independent simple sentences.

When a sentence does not fulfill the condition, the clause containing a predicate located just before the candidate of division points becomes an embedded clause. Figure 7 shows an example of a sentence which includes an embedded clause. In case that a sentence includes an embedded clause, dividing the sentence prevents readers from understanding the modified bunsetsu of the bunsetsu "*watashi-wa*" and, thus, the readers become unable to understand the meaning of the sentence.

However, even if a sentence contains an embedded clause, there are some cases that the sentence can be divided. Therefore, in order to judge it, we added the following two exception conditions. If one of the following conditions is fulfilled, the division can be performed even if the above-mentioned condition is not fulfilled.

1) The morpheme located just after the candidate of division points is a demonstrative, a conjunction, or a conjunctive particle, or the clause located just after the candidate of division point is the topicalized-element"-*wa*" clause.
2) The dependency relation which straddles the candidate of division points has the modifier bunsetsu of which the head morpheme is adverb or conjunction.

When the condition 1) is satisfied, the former and the latter of the sentence are considered to be connected by the coordinative relation. Figure 8 shows an example of a sentence which satisfies the condition 1). Here, the candidate of division points is located just after bunsetsu "*fukamari*," and

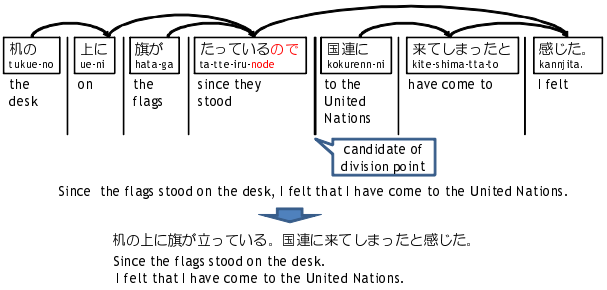

Fig. 6. *Example of a sentence which fulfills the condition*

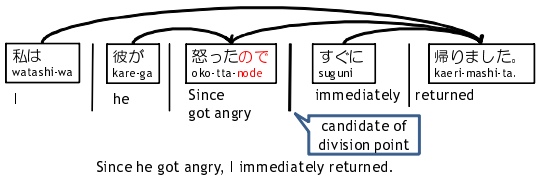

Fig. 7. *Example of a sentence which includes an embedded clause*

bunsetsu "*oosyuu-ni oite-wa*" is the modifier bunsetsu of the dependency relation which straddles the candidate. Therefore, the dependency information is lost by dividing the sentence. However, a reader can understand the meaning of a sentence by complementing it.

When the condition 2) is satisfied, the modifier bunsetsu of the dependency relation which straddles the candidate of division points often modifies the whole sentence. Figure 9 shows an example of the sentence which fulfills the condition 2). In this example, the candidate of division point is located just after the bunsetsu "*henkan-sare*", and the bunsetsu "*soshite*" is the modifier bunsetsu of the dependency relation which straddles the candidate.

C. Structuration processing

In our research, the parallel structures in a sentence are detected and specified. The detection of parallel phrases is performed by using KNP[11]. KNP can judge whether the dependency relation between bunsetsus is a parallel one. Figure 10 shows an example of the detection of a parallel phrase. Based on bunsetsus of which the dependency relation is a parallel one, the parallel phrases are detected as follows. The final bunsetsu of the parallel phrase 1 is defined as the modifier bunsetsu "*koui-ya*" of the parallel dependency relation. The start bunsetsu of the parallel phrase 1 is defined as the leftmost bunsetsu "*wa-ga kuni-ga*", which can reach the final bunsetsu "*koui-ya*" along the dependency relation paths. If there is no bunsetsu which depends on the final bunsetsu, the start bunsetsu is defined as the final bunsetsu itself. Parallel phrase 1 is defined as a sequence of bunsetsus "*wa-ga kuni-ga*" from the start bunsetsu "*koui-ya*" to the final bunsetsu. Next, the final bunsetsu of parallel phrase 2 is defined as the modified bunsetsu "*kutsuu-wo*" of the parallel dependency relation. The start bunsetsu of parallel phrase 2 is defined as the next bunsetsu "*ta-kokumin-ni*" of the final bunsetsu "*koui-ya*" of the parallel phrase 1. The parallel phrase 2 is defined as the

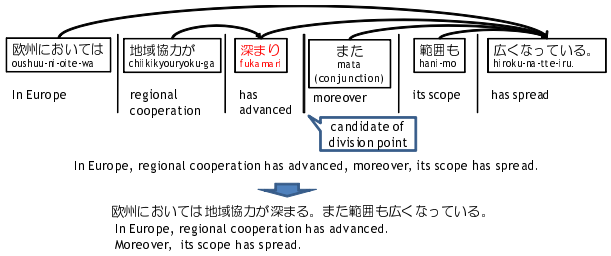

Fig. 8. *Example of a sentence which fulfills the exception condition 1*

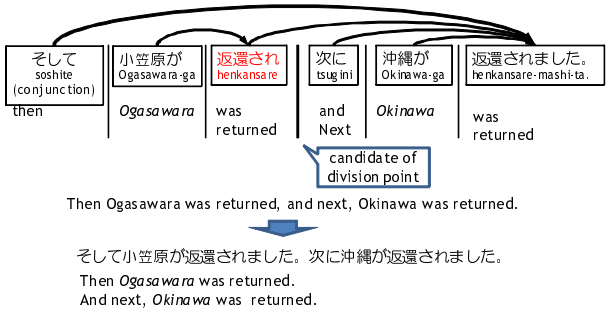

Fig. 9. *Example of a sentence which fulfills the exception condition 2*

sequence of bunsetsus form the start bunsetsu "*ta-kokumin-ni*" to the final bunsetsu"*koui-ya*".

The morphological analysis using JUMAN[12] causes many errors because of spoken language expressions. The errors caused by the morphological analysis influence the detection of parallel phrases. Therefore, among the parallel dependency relations detected by KNP, we finally decided only the parallel dependency relation which satisfies the following condition as the correct parallel one.

- The parallel dependency relation has the modifier bunsetsu of which the rightmost morpheme is one of the following morphemes: "*mata*", "*to*", "*soshite*", "*toka*", "*oyobi*", "*ya*", "*shi*".

V. Evaluation

In order to evaluate the effectiveness of our proposal method, we conducted an experiment. As the correct data, we used 3 lectures (392 sentences) of which sentences were edited by hand. We obtained the precision and recall in each of paraphrase processing, division processing and structuration processing. In the following subsection, we describe the evaluation of each processing.

A. Evaluation of paraphrase processing

The precision was 76.6%, and the recall was 50.7%. In our paraphrase processing, many redundant expressions remained, compared with the correct data. Figure 11 shows an example of the redundant expressions.

These redundant expressions are accompanied by the noun such as "*sidai*" or "*wake*". It is possible to handle them by creating corresponding rules. Moreover, in the correct data, demonstratives or conjunctions are deleted and particles are

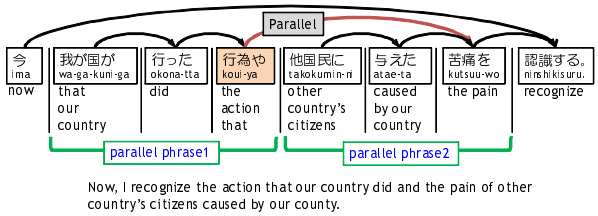

Fig. 10. *Example of detection of a parallel phrase*

原文：それで彼は安心した次第でございます。
(Then, he felt relieved.)
正解データ：それで彼は安心しました。
提案手法：それで彼は安心した次第です。

Fig. 11. *Example of a redundant expression which was not deleted*

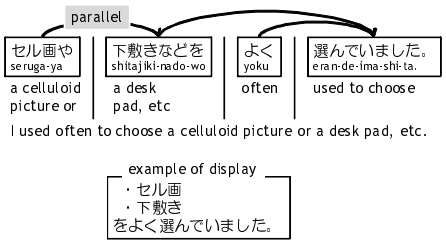

Fig. 12. *Example of failure of the structuration*

complemented. Since many of these processing are paraphrased based on contexts, it is difficult to paraphrase them automatically by using rules. In addition, although we created rules for the spoken language expressions which appeared in analysis data, there is the problem in the coverage.

B. Evaluation of division processsing

The precision was 64.2%, and the recall was 74.5%. Our division processing does not take into account the length of a sentence (e.g. the number of characters). In case that a original sentence is short or each sentence generated by dividing the original one becomes too short, the readability is not necessarily improved by dividing the original one. Moreover, the errors caused by our created rules were also seen. These rules need to be scrutinized. We have to restrict the conditions on which a sentence can be divided, by using the information on part-of-speech and dependency relation which is obtained from a larger size of analysis data.

C. Evaluation of structuration processing

The precision was 34.8%, and the recall was 24.6%. Both the precision and recall were low. As the reason, there is a problem of the accuracy of JUMAN and KNP. Although we created the rules for reducing the influence of errors caused by JUMAN, the accuracy of KNP did not improve enough. Moreover, even if the structure which was detected as a parallel structure was semantically parallel in fact, there existed the case that the display of the parallel structure did not correspond with that of the correct data. Figure 12 shows an example of this case.

VI. Conclusion

In this paper, we proposed the spoken document content which achieves the effective reuse of speech data, and a technique for editing sentences in spoken document in consideration of the readability. Moreover, we described the fundamental technologies for actualizing the edit of a spoken document, and implemented them. The experimental results showed the feasibility of our technique.

Future research includes the extension of the paraphrase rules to improve the precision and recall. Furthermore, in the division processing or structuration processing, it is necessary to consider the readability. Especially, the number of characters in a sentence or a line is thought to influence the readability. We plan to design these techniques in consideration of the factors.

Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 20300058) of JSPS and by the Asahi Glass Foundation.

References

[1] Nakagawa. S, Togashi. S, Yamaguchi. M, Fujii. Y, Kitaoka. N. Useful Contents of Classroom Lecture Speech and a Browsing System (in Japanese), *Transactions of Institute of Electronics, Information and Commmunication Engineers on Information and Systems (Japanese Edition)* , Vol. 91-D, No.2, pp. 238-249, 2008.

[2] Yamamoto. K, Adachi. Y. Informative Spoken Language Summarization of the Diet Minutes (in Japanese), *Journal of Natural Language Processing*, Vol. 12, pp. 3-30, 2005.

[3] Hayashi. Y. A Three-level Revision Model for Improving Japanese Badstyled Expressions, *Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 2, pp. 665-671, 1992.

[4] Lee. L, Chen. B. Spoken Document Understanding and Organization, *Signal Processing Magazine, IEEE*, Vol. 22, No. 5, pp. 42-60, 2005.

[5] Lee. L, Kong. S, Pan. Y, Fu. Y, Huang. Multi-Layered Summarization of Spoken Document Archives by Information Extraction and Semantic Structuring, *Proceedings of 9th International Conference on Spoken Language Processing*, 2006.

[6] Zhu. X, Penn. G. Summarization of Spontaneous Conversations, *Proceedings of 9th International Conference on Spoken Language Processing*, 2006.

[7] Murray. G, Renals. S, Carletta. J. Extractive Summarization of Meeting Recordings, *Proceedings of 9th European Conference on Speech Communication and Technology*, 2005.

[8] Maskey. S, Hirschberg. J. Comparing Lexial, Acoustic/Prosodic, Discourse and Structural Features for Speech Summarization, *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005.

[9] Christensen. H, Kolluru. B, Gotoh. Y, Renals. S. From Text Summarisation to Style-Specific Summarisation for Broadcast News, *Proceeedings of European Colloquium on IR Research*, Vol. 2997, pp. 223-237, 2004,

[10] Kim. Y.B, Ehara. T. An Automatic Sentence Breaking and Subject Supplement Method for J/E Machine Translation (in Japanese), *Transactions of Information Processing Society of Japan*, Vol. 35, No. 6, pp.1018-1028, 1994.

[11] Kurohashi. S, Nagao. M. A syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures, *Computational Linguistics*, Vol. 20, No. 4, pp.507-534, 1994.

[12] Kurohashi. S, Nagao. M. Improvements of Japanese Morphological Analyser JUMAN, *Proceedings of 15th International Conference on Computational Linguistics* , Vol. 2, pp. 1123-1127, 1994.