Automatic Indexing of Lecture Presentations Using Unsupervised Learning of Presumed Discourse Markers

Tatsuya Kawahara, *Member, IEEE*, Masahiro Hasegawa, Kazuya Shitaoka, Tasuku Kitade, and Hiroaki Nanjo

***Abstract*—A new method for automatic detection of section boundaries and extraction of key sentences from lecture audio archives is proposed. The method makes use of 'discourse markers' (DMs), which are characteristic expressions used in initial utterances of sections, together with pause and language model information. The DMs are derived in a totally unsupervised manner based on word statistics. An experimental evaluation using the Corpus of Spontaneous Japanese (CSJ) demonstrates that the proposed method provides better indexing of section boundaries compared with a simple baseline method using pause information only, and that it is robust against speech recognition errors. The method is also applied to extraction of key sentences that can index the section topics. The statistics of the presumed DMs are used to define the importance of sentences, which favors potentially section-initial ones. The measure is also combined with the conventional tf-idf measure based on content words. Experimental results confirm the effectiveness of using the DMs in combination with the keyword-based method. The paper also describes a statistical framework for transforming raw speech transcriptions into the document style for defining appropriate sentence units and improving readability.**

***Index Terms*—Automatic indexing, discourse marker, language model, spoken language system, spontaneous speech.**

I. Introduction

AUTOMATIC indexing of audio materials is one of the applications of large vocabulary continuous speech recognition. Even if recognition performance is not very high, it is often possible to detect topics or to segment speech by topic boundaries so as to help users efficiently find segments that they might want to listen to from audio archives.

Previous studies of this kind of application have addressed topic classification for broadcast news [1] and voice mails [2]. Most of these studies extract a set of keywords (KWs) that characterize topics for classification [3]. This approach is effective when there are many short recordings such as news clips and voice messages, each of which might contain a few minutes of speech at most. However, this method is not easily applied to the indexing of materials such as lectures and discussions, where each recording might contain dozens of minutes of speech, during which one broad topic remains unchanged while closely related small subtopics succeed each other. The KWs that characterize the whole topic appear throughout the recording, so that a broad classification based on such KWs is meaningless. Instead, a browsing function is needed for this kind of material [4], [5]. Specifically, exact time indices for boundaries of subtopics or 'sections' are required, since such indices can be used to locate segments to be replayed.

The structure of sections and paragraphs is also known to be useful for extracting 'key sentences' that can be used to index text materials, because many key sentences appear at the beginning of sections. In audio materials, however, there is no explicit definition of sections and paragraphs such as the line-breaks and indentation of text.

We approach the problem of indexing lecture audio archives by automatically detecting the boundaries of 'sections.' We focus on 'discourse markers' (DMs), which are rather topic independent. In this work, DMs are defined as expressions that are characteristic to the beginning of new sections in lectures and oral presentations. Unlike in earlier studies of DMs, this study adopts unsupervised training. That is, the proposed method defines DMs solely in terms of their distribution in the lectures, and extracts these 'presumed' DMs automatically, without any manually tagged information about topics and boundaries. While the method is initiated by a pre-selection based on pause information, the final indexing is done based on the statistics of the derived DMs.

We then can make use of the DMs that suggest the beginning of sections for extracting key sentences that can be used as index labels in browsing through the lectures. The proposed method is complemented by the conventional method that focuses on topic-dependent KWs.

The measure that we use to define DMs is based on the comparison of the first sentences of potential sections with other sentences, and thus requires an appropriate definition of the unit 'sentence'. An appropriate definition of sentence units is vital also for identifying the key sentences that provide users with efficient access to the audio archives. It is not easy to define and automatically segment such units for spontaneous Japanese, which is characterized by hesitations and pauses in seemingly arbitrary places, as well as by spoken-style inflectional endings and other morphological structures not used in the written language. Therefore, we also present a statistical framework that 'translates' raw transcriptions into document-style sentences as a pre-processing step.

The outline of the paper is as follows: Section II describes the model of discourse structure and DMs that is the background to our approach to indexing, as well as the corpus and baseline speech recognition system used in this study. Section III describes the statistical translation for cleaning raw transcriptions.

Manuscript received May 2, 2003; revised February 20, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mary Beckman.

The authors are with the School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan (e-mail: kawahara@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TSA.2004.828701

Section IV presents the methods of unsupervised training of presumed DMs and indexing of section boundaries using them, which are then evaluated in Section V. Section VI describes the application of the presumed DMs to extraction of key sentences, which is then evaluated in Section VII. Section VIII concludes the paper.

II. Approach to Indexing of Lecture Audio Archives

A. The Corpus of Spontaneous Japanese (CSJ)

The research described in this paper is part of the "Spontaneous Speech Corpus and Processing Technology" project sponsored by the Science and Technology Agency Priority Program in Japan [6]–[9]. The CSJ [10][1] developed by this project is one of the largest spontaneous speech databases: approximately 7 M words in terms of text size and 500 h in terms of speech. The CSJ mostly consists of two styles of monologue: academic presentation speeches at technical conferences and extemporaneous public speeches on topics such as hobbies and travel. These monologues are manually given orthographic and phonetic transcriptions, but they are not segmented at all, i.e., one large file corresponds to a talk. In this study, we mainly deal with the academic presentation speeches with the purpose of providing efficient access to this kind of lecture audio archives.

B. Modeling the Structure of Lecture Presentations

There is a relatively clear prototype for the flow of these presentations. Typically, the introduction provides an overview of the background. Next, the specific problem and approach are described. Then comes an explanation of concrete algorithms and systems, followed by an experimental evaluation. When a speaker uses slides or viewgraphs, the sequence of descriptions of the slides or viewgraphs typically constitute the units that are relevant for listeners, including later users wishing to browse through the recording of the talk. We call this flow a "slide-based discourse structure".[2] Thus, the definition of 'section' in this paper almost corresponds to the slide, except that several slides on the same topic are merged into one unit. The unit in turn usually corresponds to the numbered (sub)sections in the proceedings paper.

Our first goal is to segment the lecture audio material into these units based on the structure, or to find the boundaries between them. The time indices for the boundaries are useful for skipping over the demarcated audio segments in search of relevant sections. If these audio segments are aligned with slides, though that is not done here, multi-media browsing can be realized.

Finding boundaries between sections is also known to be a useful heuristic for extracting key sentences in text-based natural language processing [11]. However, the text-based methodology cannot be simply applied to spoken language because the boundaries of sections are not explicit in speech. Thus, the second goal of the study is to apply the proposed method of discourse segmentation to the problem of extracting key sentences from the lectures, e.g., to generate content-based tags for the audio segments.

[1]The official release of the CSJ to the public is due in early 2004.

[2]The naming was by Prof. R. Grishman at SSPR-03 workshop.

Previous works on discourse segmentation of speech include Passonneau and Litman [12], who addressed discourse segmentation of narrative monologues using various combinations of prosodic features, cue phrase features, and noun phrase reference features. They crafted decision rules combining these features both by hand and by machine learning. However, deterministic rules are not robust against inputs containing speech recognition errors, especially if they rely on reference patterns of noun phrases. Haase *et al.* [13] also tried discourse segmentation of monologue news reports using several prosodic features. Preliminary analysis on our corpus suggests, however, that prosody alone is not sufficient for detection of section boundaries. Therefore, we adopt a statistical framework that is based on a "bag-of-words" model of DMs as well as on pause information in order to robustly handle speech input.

We focus on typical patterns in the first utterances of section units corresponding to lecture segments. We capitalize on the fact that, in the initial utterances of sections, speakers try to briefly tell what comes next and attract the audience's attention by saying, for example, "Next, I will explain how it works." and "Now, let's move on to the experimental evaluation". This typical pattern means that key sentences that can be used as tags for indexing the lectures often appear at the beginning of sections. We define DMs simply as characteristic expressions that appear at the beginning of sections.

A more conventional definition of DMs, which are often referred to as cue phrases or clue words, is that they convey explicit information about the structure of discourse rather than any literal semantic information [14], [15]. Typical examples of general DMs are the words "now" and "next" used metaphorically to refer to discourse time and position as in the two examples above. Hirschberg and Litman [15] proposed a prosodic model to disambiguate DMs such as "now" from their literal counterparts. Kawamori *et al.* [16] broadened the class of DMs for Japanese to include responsive interjections such as "*hai*" (literally "yes") and fillers such as "*e*" and "*ano*:" (corresponding roughly to "uhm" in English), differentiating their various discourse functions on the basis of prosodic features such as pitch patterns. Quimbo *et al.* [17] also explored the use of prosodic features to disambiguate these fillers from similar-sounding words or parts of words. For example, the filler "*ano*:" contains the same segments as the deictic adjective "*ano* (that)", but is typically produced with a lengthening of the final vowel. However, these studies mainly deal with dialogue data. As Kawamori *et al.* pointed out, fillers and their prosodic patterns have an important role as DMs in dialogue because they often appear at the beginning of speaker turns. But this does not apply to the monologues that we deal with in this study. Moreover, we observed that many of the responsive and modal DMs in Japanese dialogue are so colloquial that they are rarely used in public speaking such as lectures. Therefore, we rely on lexical DMs, and we also include in this set some words and phrases that are particular to lecture presentations. That is, since technical presentations generally address problems and evaluation results of proposed solutions, the word "problem" and "result" also signal parts of the discourse in this style of speech. Thus, even in their literal use, they can also be regarded as DMs specific to the lecture presentation style.

Moreover, unlike in previous studies, where DMs are defined *a priori* based on linguistic analysis, our method automatically

U1: *e: soredewa*

(well, then)

U2: *saQsoku ano: naiyou ni haira shi te itadaki masu @@ e:to tekisuto oNsei gousei*

(let me directly go into the main topic @@ well, as a text-to-speech)

U3: *houshiki to shi te @ ma: ano: kono mae no ima seNsei kara choQto rebyu: ga ari mashi ta kedomo @ e: tsui*

(method @ as the previous speaker just made another review @ well, just)

U4: *juu neN ijou mae made wa ma: kisoku gousei Qte iu houshiki ga (shu) shuryuu daQ ta wake desu ga @@ saikiN ko:pasu ni motozuku oNsei gousei to iu no ga ano: shuryuu ni naQ te ki te ru to @@*

(ten years ago or so, the rule-based synthesis method was common @@ but recently the corpus-based speech synthesis is, well, getting popular @@)

Fig. 1. Example of speech units automatically segmented based on pause information. In this romanization of the transcription (which is in Japanese orthography), ":" stands for a long vowel, "Q" for a long consonant, and "N" for the moraic nasal. "@@" stands for a sentence boundary, and "@" for a clause boundary. The English translation is intended only to give a rough sense of the meaning, and is not an exact word-by-word gloss.

derives a set of DMs without any manual tags. Moreover, we do not assume that correct segmentations are given for training because it costs too much to manually tag the large database. Thus, our DMs, which are automatically defined from a set of lecture transcriptions, are presumptive and not necessarily exact in the linguistic sense. However, they are expected to be useful for indexing lecture archives through automatic speech recognition (ASR).

C. Automatic Lecture Transcription System

The ASR system for the lecture corpus (CSJ) has also been developed in our laboratory under the "Spontaneous Speech Corpus and Processing Technology" project.

All transcribed data (as of January 2003) including extemporaneous public speeches are used for language model training. There are 2592 presentations and talks by distinct speakers. The text size in total is 6.7 M words (=Japanese morphemes). A trigram language model is trained with a vocabulary of 24 K words. For acoustic model training, only male speakers of academic presentations are used in this work. By using 781 presentations that amount to 106 h of speech, we set up a gender-dependent phonetic tied-mixture (PTM) triphone model that consists of 25 K Gaussian components and 576 K mixture weights. We also revised our recognition engine Julius [18] so that very long speech files can be handled without prior segmentation [7].

With the baseline system, the word error rate is 30.9% for the ASR evaluation set of 15 academic presentation speeches [9]. Adaptation of acoustic and language models based on the initial recognition result together with the speaking-rate dependent decoding strategy [19] improves it to 22.0%.

III. Automatic Insertion of Periods and Particles in Speech Transcriptions

Transcriptions of spontaneous speech include many disfluency phenomena and do not have linguistic punctuation such as periods. In read speech, a long pause is regarded as a mark of the end of a sentence or clause, and thus can be converted into a period or comma. In spontaneous speech, however, this assumption does not hold. Speakers put pauses in other places for certain discourse effects, and disfluency causes irregular pauses. Therefore, we make use of linguistic information as well as pause information in order to insert periods. The period insertion procedure is necessary for segmenting speech into appropriate sentence units to be indexed. Especially, indices to audio segment boundaries should not be assigned in the middle of sentences, as this would make it difficult for listeners to follow. Fig. 1 shows an example of utterances that are segmented with a pause-based algorithm which is deployed in our dictation system. Clearly, the segmented utterance units do not correspond to sentence or clause units in spontaneous Japanese. Note in particular that the pause between U2 and U3 breaks up the compound word "*tekisuto oNsei gousei houshiki* (text-to-speech method)".

Another problem in spoken Japanese is that particles such as the "*-o*", which marks the preceding noun phrase as the object of the clause, are often omitted, particularly when the syntactic role of the noun phrase is clear from the context. Recovering these omitted particles is also needed to properly define compound nouns for which statistics need to be computed in the following stages of processing. This pre-processing will also help the sentence segmentation (period insertion) by making explicit the grammatical cases of nouns.

We approach these problems by using a statistical framework that has become popular in machine translation. We treat the spoken and written styles of Japanese as different languages and apply the translation methodology to automatic transformation of the former into the latter. With this framework, deletion of fillers and correction of colloquial expressions can also be handled in an integrated manner [20]. The method will be useful for archiving transcripts of lectures, because spoken Japanese is quite different from the written language and thus needs heavy cleaning for documentation. Conventional approaches to this problem typically have relied on heuristic rules and simple pattern matching. Recently, Murata *et al.* [21] explored automatic extraction of such rules. However, such rule-based approaches cannot control the application of rules by referring to the naturalness of the output and they do not incorporate context dependent effects as readily as can be done with the statistical machine translation framework.

The statistical machine translation method [22] is formulated in the same way as statistical speech recognition; i.e., find the best output sequence Y for an input sequence X, such that

the *a posteriori* probability $P(Y|X)$ is maximum. It is equivalent to maximization of the product of $P(Y)$ and $P(X|Y)$, where $P(Y)$ is the probability of the source language model and $P(X|Y)$ is the probability of the transformation model. The transformation model represents the correspondence of input and output word sequences. In the case of the transformation problem addressed here, the input X is a word sequence of spoken language transcription that does not have periods but marks pauses with their durations. The output Y is a word sequence of the cleaned text. The term $P(Y)$ is effective for considering contextual effects and naturalness of the output. For its calculation, we use a word 3-gram model trained with a written language corpus. Since applying the conversion of one word affects neighboring words in the N-gram model, decoding is performed for a whole input word sequence with beam pruning.

The specific procedures are explained in the following subsections by illustrating how they are applied in two examples of the kinds of problems that are addressed in the 'translation' process: the insertion of periods and of particles.

A. Insertion of Periods

Transcriptions of the CSJ have pause marks with their durations instead of periods, and the speech recognizer using a language model trained with the CSJ does not output periods.

A word 3-gram language model is used to judge whether a period should be inserted at the position of a pause. We made use of another language model trained with punctuated texts of lecture archives that had been collected via the World Wide Web and consisting of 1.7 M words. Since the texts had been edited for public readability, the model is not matched to spontaneous lectures [23]. For a word sequence around a pause, $X = (w_{-2}, w_{-1}, pause, w_1, w_2)$, a period is inserted at the place of the pause if $P(Y_1) = P(w_{-2}, w_{-1}, period, w_1, w_2)$ is larger than $P(Y_2) = P(w_{-2}, w_{-1}, w_1, w_2)$ by some margin. In this baseline method, the decision to insert a period or not is based on the probability of $P(Y)$ only.

Then, we introduce a more elaborate model that converts pauses into periods selectively by considering the duration information and the adjacent words in the statistical transformation framework. Specifically, we introduce a pause duration threshold of X, which further conditions whether a pause should be converted into a period depending on its contextual words $(w_{-2}, w_{-1}, w_1, w_2)$. This is realized by introducing a term $P(X|Y)$. Although the value of $P(X|Y)$ is binary (0/1) given by a simple thresholding function, the threshold value in the pause length is dependent on the sequence pattern $(w_{-2}, w_{-1}, w_1, w_2)$. Concretely, the threshold should be determined by observing the minimum duration when there is a corresponding period in the sequence Y, since the final judgment is done by the source language model $P(Y)$.

Thus, a pause following typical Japanese end-of-sentence expressions, as in "*-desu (pause)*" and "*-masu (pause)*", can be converted into a period even if its duration is short. On the other hand, if a pause follows end-of-sentence expressions peculiar to spoken Japanese as in "*-to (pause)*", "*-nai (pause)*" and "*-ta (pause)*", or a pause precedes a conjunction peculiar to spontaneous speech "*(pause) de-*", it can be converted only when the duration is relatively long.

A preliminary evaluation of this procedure was done using four lectures that were excluded from the training corpus. A professional editor cleaned the transcriptions and inserted periods. The following three methods were compared.

1) Zero threshold: any pause can be converted into a period.
2) Setting the threshold to the average pause duration in a talk, which was most effective as a fixed threshold
3) The proposed method that uses different threshold values depending on the context

TABLE I
RESULTS OF PERIOD INSERTION

threshold of pause duration	recall	precision	F-measure
1) zero	83.2%	75.4%	0.791
2) average	64.4%	93.7%	0.763
3) depending on expressions	76.3%	92.3%	0.835

The recall, precision rates, and F-measure for these methods are listed in Table I. The F-measure is the combination of the recall and precision rates defined as

$$\mathrm{F-measure} = \frac{2 * \mathrm{recall} * \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}. \quad (1)$$

When we use the zero threshold, a large number of erroneous insertions are caused, and the precision rate is low. In contrast, fixing the threshold at the average value improves the precision considerably, but results in a much lower recall rate. With the context-dependent threshold, both higher recall and precision rates are obtained.

B. Insertion of Particles

Japanese particles are often omitted in spontaneous speech, and the frequency of omissions is apparently different depending on the particle and its adjacent words; namely, the omission is more likely to occur when it would not cause misunderstanding. Therefore, we introduce a statistical model and define the deletion probabilities of particles $P(X|Y)$ for triplets of the preceding part of speech, the particle itself, and the following part of speech, such as "*Noun Particle Noun*", "*Noun Particle Verb*", and "*Noun Particle Adjective*", as listed completely in Table II. The probability estimation is done using the parallel corpus of original transcriptions and their professionally edited clean versions of 18 lectures from the CSJ.

The particle insertion procedure was evaluated using the same test-set as in the previous subsection. The proposed statistical method was compared with the conventional rule-based method, in which the transformation score $P(X|Y)$ was not used or set to 1 in all permissible cases. Results are given in Table III. Although both methods achieved the same recall rate of 89.4%, the statistical transformation model successfully suppressed a significant number of false alarms, and thus improved the precision rate.

Most of the remaining false alarms were insertions of the attributive/genitive particle "*-no*" (meaning "of") within compound nouns, resulting in errors such as "*Kyoto daigaku*" (the proper noun "Kyoto University") being changed to "*Kyoto no daigaku*" (a phrase meaning "a university located in Kyoto"). The deletion pattern of "*no*" was derived from cases of concatenation of many nouns such as "*buN* (sentence) *kyoukai* (boundary) *jidou* (automatic) *keNshutsu* (detection)

TABLE II
PATTERNS AND PROBABILITIES OF PARTICLE DELETION

pattern Y	deletion probability
Noun wa Noun	0.073
Noun o Noun	0.032
Noun ni Noun	0.0046
Noun ga Noun	0.0028
Noun no Noun	0.0017
Noun wa Verb	0.056
Noun o Verb	0.040
Noun ga Verb	0.012
Noun to Verb	0.0097
Noun ni Verb	0.0016
Noun wa Adjective	0.02
Noun ga Adjective	0.024
Noun wa Conjunction	0.16

"*wa*", "*o*", "*ni*", "*ga*", "*no*", and "*to*" are Japanese particles.

TABLE III
RESULTS OF PARTICLE INSERTION

method	recall	precision	F-measure
rule-based	89.4%	58.3%	0.706
statistical	89.4%	65.9%	0.759

arugorizumu (algorithm)" where "*no* (of)" should be inserted after "*kyoukai*" for better readability. Although its deletion probability $P(X|Y)$ is very small $(= 0.0017)$, insertion of "*no*" is favored in many cases by the language model score $P(Y)$. However, the large number of false insertions suggests that "*no*" should be treated differently, and that oblique case particles such as the attributive/genitive "*-no*" and the locative "*-ni*" can be treated differently since they are much less likely to be deleted than the subject and object case particles, as suggested by the relatively low $P(X|Y)$ values associated with them (see Table II).[3] The problem of false insertions of "*no*" could also be eased by adding entries of proper nouns to the lexicon. If we exclude these errors, the precision rate should be about 79.4%.

IV. AUTOMATIC DERIVATION OF DISCOURSE MARKERS (DMs)

The procedure for extracting DMs is illustrated in Fig. 2. First, candidate section boundaries and their first utterances are extracted. Then, we compute the statistics of word frequency and sentence frequency, which are the basis for selecting DMs. These processes use various information sources such as pause, N-gram language model, and statistics of word occurrences.

A. Use of Pause Information

Pause information is used for pre-selection, that is extracting candidate section boundaries. It is expected that speakers put relatively long pauses in shifting topics or changing slides, although a long pause does not always mean a section boundary. Hirschberg and Litman [15] also confirmed that DMs are very much likely to be preceded by orthographic markers of comma, periods, and paragraph breaks in transcription, which apparently correspond to pauses in speech. Also, in the work of Haase *et al.* [13], a longer pause was identified as the most distinct feature for paragraph boundaries. Here, we set a relatively low threshold not to miss correct hypotheses, which will be selected by the following process. The threshold value varies from person to person, depending mainly on the speaking rate. Therefore, we use the average pause length during a talk as the threshold for distinguishing long pauses.

B. Use of Language Model

In order to judge whether the detected pauses are actually at the ends of sentences, a word 3-gram language model is used in combination with the transformation model, as described in the previous section. Here, the existence of a short pause at the end of the utterance is assumed. When a short pause is found in the transcription, we test whether a period can be put in or not by considering the neighboring word sequence and the length of the pause. Thus, we obtain a candidate set of first utterances of sections.

C. Use of Word Frequency and Sentence Frequency

From these candidate first sentences, we extract characteristic expressions. That is, we select DMs that can be used to locate section boundaries. As a pre-processing step, we exclude function words and proper nouns because the function words appear in any utterance and proper nouns would appear only in a limited set of lectures.

DMs should frequently appear in the first utterances of sections in all lectures, but should not appear in subsequent utterances so often. Word frequency is used to represent the former property and sentence frequency is used for the latter. For a word w_j, the word frequency wf_j is defined as the number of occurrences in the set of first sentences. The sentence frequency sf_j is the number of sentences in all lectures that contain the word. The larger wf_j is and the smaller sf_j is, the more appropriate the word is as a DM for indexing. We rate each word's DM potential using (2), which is the essentially same function as the tf-idf measure used in information retrieval

$$S_{\mathrm{DM}}(w_j) = wf_j * \log\left(\frac{N_s}{sf_j}\right). \tag{2}$$

Here, N_s is the total number of sentences in all lectures. A set of n DMs is selected by ranking the words that occur in candidate first sentences according to $S_{\mathrm{DM}}(w_j)$ and taking the top n words. We also investigated the effect of weighting wf_j and sf_j, and found that the above simple formula, in which both weights are 1, was the most effective.

D. Indexing Using DMs

For a given new lecture, automatic indexing using the presumed DMs is done by almost the same procedure as the training phase as depicted in Fig. 2. First, the candidate section boundaries are chosen based on long pauses. Next, their initial utterances are cut out based on short pauses and the language model for punctuation. Then, they are evaluated using the same function that was used in rating the DM potential of words in the

[3]It was pointed out by the AE, Prof. M. Beckman.

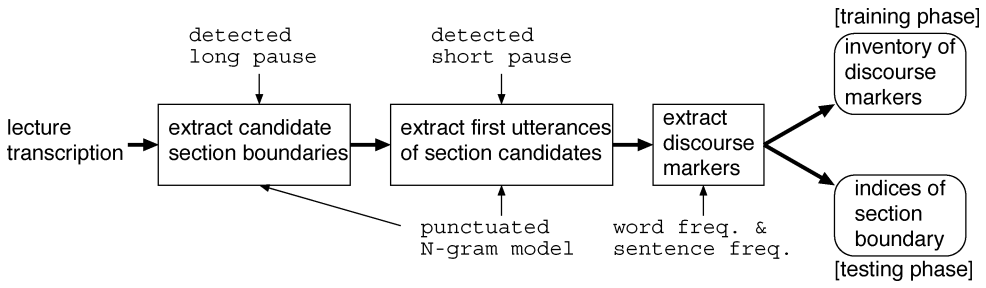

Fig. 2. Flow of extracting DMs for indexing.

training phase, along with a threshold value θ. Suppose a sentence s_i contains several DMs ($s_i \ni \{w_1, \ldots w_m\}$), we compute the following summed score:

$$S_{\mathrm{DM}}(s_i) = \sum_{w_j \in s_i} S_{\mathrm{DM}}(w_j). \tag{3}$$

Then, the start time of the utterance is indexed as a section boundary if $S_{\mathrm{DM}}(s_i) > \theta$.

V. Evaluation of Section Boundary Detection

For choosing the DMs in the training phase, we used most of the academic presentations in the CSJ, which seem to follow the model described in Section II-B. These presentations were collected at a variety of technical conferences. We set aside an evaluation set of seventeen presentations that are not included in the training set. The duration of the lectures is 11–15 min. The 'correct' section boundaries for the test-set were determined by human observation by considering the structure of the slides or the corresponding proceedings paper. The number of boundaries per lecture varied between 7 and 16.

The evaluation measure is based on the recall rate of the correct boundaries and the precision rate of the detected boundaries. Here, we modify the F-measure so that the recall rate is weighted more highly, since the correct boundaries should not be missed in indexing, while false alarms can be simply skipped over during searching

$$\mathrm{F-measure}(\alpha) = \frac{\left(1 + \frac{1}{\alpha}\right) * \mathrm{recall} * \mathrm{precision}}{\frac{1}{\alpha} * \mathrm{recall} + \mathrm{precision}}. \tag{4}$$

We set $\alpha = 10$, which puts a ten-times larger weight on the recall rate.

A. Effect of DMs

We first evaluated our segmentation method with the manual transcriptions of lectures. Based on evaluation function (2), 75 DMs are selected. The recall rate, precision rate, and F-measures ($\alpha = 1$ and 10) are plotted in Fig. 3 as functions of the threshold θ.

For comparison, we also tried a simple indexing method using the pause length only, where locations of pauses longer than a threshold duration are indexed as section boundaries. This corresponds to the method employed in current audio tape recorders. The operation curve obtained by varying the threshold length is plotted in Fig. 4.

By comparing the two graphs, we can see that the proposed method has better indexing performance. In particular, for high recall rates (the left-most region of the graph), it has much higher precision rates. When the F-measure(10) that puts priority on the recall rate is used, the superiority of the proposed method is even more apparent. Therefore, the use of the presumed DMs is effective.

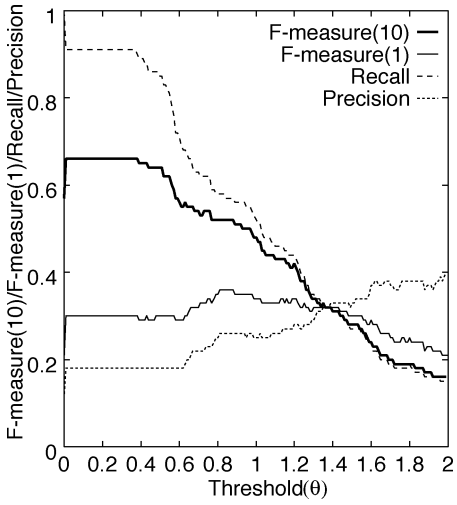

Fig. 3. Indexing performance of section boundaries using DMs.

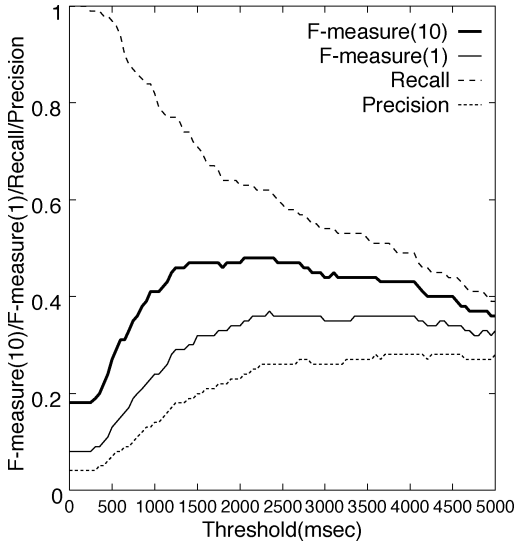

Fig. 4. Indexing performance of section boundaries using pause duration only.

B. Characteristics of DMs

Next, we investigate the effect of DMs by changing their number between 25, 75, and 125. The F-measure(10) is plotted

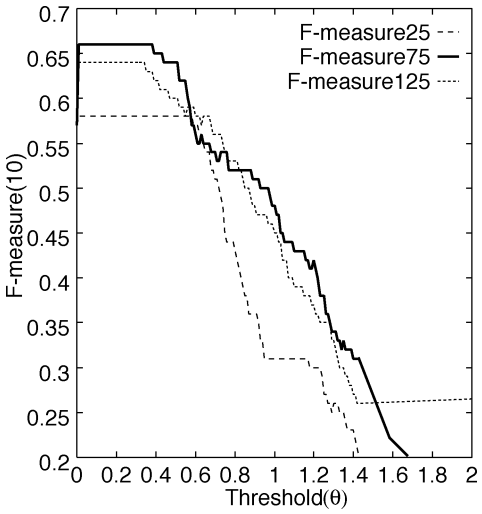

Fig. 5. Indexing performance of section boundaries by changing the number of DMs.

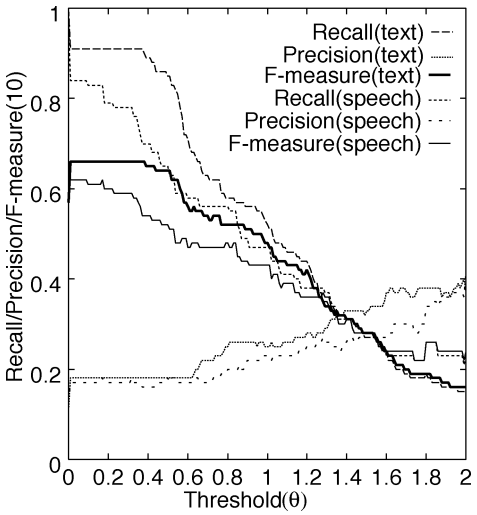

Fig. 6. Indexing performance of section boundaries for speech recognition results in comparison with manual transcription (text) case.

TABLE IV
EXAMPLE OF DERIVED DMS

tsugi (next), *sakihodo* (As I said), *ima* (now), *jiQsai* (actually), *koNkai* (this time), *saigo* (lastly), *saisho* (firstly)
keNkyuu (study), *setsumei* (to explain), *haQpyou* (to present)
keQka (result), *jiQkeN* (experiment), *moNdai* (problem), *hyouka* (evaluation), *houhou* (method), *mokuteki* (purpose)

listed in Japanese with corresponding English words

in Fig. 5. It is clear that we need to use a large enough number of DMs to get most of the section boundaries, but using too many markers increases false alarms, and thus degrades the precision and F-measure. Based on this comparison, we chose 75 markers.

Table IV lists examples of the presumed DMs. The first category corresponds to conventional DMs such as "now" and "actually". The members of the second category, exemplified by "to explain" and "to present", are somewhat more dependent on lecture-style speech, but still can be regarded as general DMs, because these words suggest a new topic. The third category is totally dependent on the prototypical discourse structure of the lecture presentations: "purpose", "problem", "method" and "evaluation". The top three entries in order of the evaluation measure [function (2)] are "*tsugi* (next)", "*keQka* (result)" and "*keNkyuu* (study)". Although the set also contains several technical terms such as "model" and "data", which are rather specific to the engineering field, the results show that the proposed method works essentially as we expected.

C. Evaluation With ASR Results

Then, the segmentation method was applied to transcriptions produced by the ASR system. The ASR word accuracy for the test-set lectures is around 60%–70%.[4]

Fig. 6 plots the F-measure(10) as well as the recall and precision rates, overlaid on the results for the manual transcriptions from Fig. 3. Although the recall rate gets lower because of speech recognition errors, the degradation is relatively small considering the word error rates. The F-measure is still definitely better than with the baseline method shown in Fig. 4. These results show that the statistical evaluation of section boundaries with the presumed DMs is robust.

The overall precision rate is around 20%, which means we can find the correct section indices one out of five times, and must eliminate the others (80%) for more accurate indexing by a manual post-processing. However, even experts need 5–6 h to manually attach these indices (including precise time-alignment) for each lecture presentation. Therefore, the candidate time indices automatically given by our method will be useful in reducing time and labor.

VI. INDEXING OF KEY SENTENCES

Next, we apply the proposed methodology to the problem of extracting key sentences, which can be used as tags to indicate the content of the sections that are indexed by the segmentation method described in the previous section. Collection of these sentences might also suffice as a summary of the talk [24]. The framework extracts a set of natural sentences for output. Even erroneous transcriptions generated by the ASR system can be aligned with the audio segments in order to provide an alternative summary output. This is considered to be a more practical solution for spontaneous speech, for which the ASR accuracy is around 70%–80%, as opposed to the approach of generating a summary by shortening, fusing, and otherwise modifying the text output by the ASR system [25].

Teufel and Moens [26] proposed a strategy for extracting key sentences from technical papers. They described a typical rhetorical structure of papers in scientific journals, which is similar to our discourse model for lecture presentations described in Section II-B. This structure defines principles for extracting sentences based on their content, along with each sentence's rhetorical status. Since their target is to provide a set of sentences for a short summary highlighting the original contribution of the paper, they extracted only a dozen sentences or so for a compression ratio of 2.5%. In contrast, our goal is to extract index sentences that indicate the content of each of the sections

[4]As the ASR system used in this section was not the current version, the accuracy is worse than those reported in Sections II and VII.

in order to provide efficient access to the audio lecture archive. For this purpose, a less drastic compression (20%–40%) is more appropriate, because it should be easier to skim over several sentences extracted from a segment than to deduce the content of segments where no key sentence was extracted. Our purpose also requires a strategy that is less sensitive to ASR errors, and in this task, where the compression ratio is similar to summarizing news stories, we expect that the rhetorical structuring of sections provides an effective heuristic of identifying key sentences from section beginnings.

A. Measure of Importance Based on DMs

Using the sections that are defined and autonomously derived in this paper, the heuristic is now applicable to speech materials in which there are no obvious section boundaries.

The importance of sentences is evaluated using the same function that was used for detection of section boundaries. For each sentence s_i, the summed DM potential score $S_{\mathrm{DM}}(s_i)$ is computed using function (3). While we used pause information for pre-selecting candidate section boundaries, we do not impose any assumption on pauses in this key sentence extraction. Then, key sentences are selected based on the score, up to a specified number (or ratio) of sentences from the whole lecture.

B. Combination With the Keyword-Based Method

The other approach to extraction of key sentences focuses on keywords (KWs) that are characteristic to the lecture. The most orthodox statistical measure to define and extract such KWs is the following tf-idf criterion

$$S_{\mathrm{KW}}(w_j) = tf_j * \log\left(\frac{N_d}{df_j}\right). \tag{5}$$

Here, term frequency tf_j is the number of occurrences of a word w_j in the lecture, and document frequency df_j is the number of lectures (=documents) in which the word w_j appears. N_d is the number of lectures used for normalization of the inverse document frequency. Together, a larger value for tf_j and a smaller value for df_j suggest a greater importance for the word in the lecture. Here, we treat compound nouns that appear more than three times in a talk as individual entries.

For each sentence s_i, the following score of importance is computed over the KWs in the sentence:

$$S_{\mathrm{KW}}(s_i) = \sum_{w_j \in s_i} S_{\mathrm{KW}}(w_j). \tag{6}$$

Moreover, we introduce a new hybrid measure of importance that combines this KW-based importance score $S_{\mathrm{KW}}(s_i)$ with $S_{\mathrm{DM}}(s_i)$, the importance score based on the DMs. The two scores are combined by taking a geometric mean with a weight α.

$$S_{\mathrm{final}}(s_i) = S_{\mathrm{DM}}(s_i)^{\alpha} \cdot S_{\mathrm{KW}}(s_i)^{(1-\alpha)}. \tag{7}$$

Although the value of the weight is chosen empirically, actual performance is fairly insensitive unless extreme values are used.

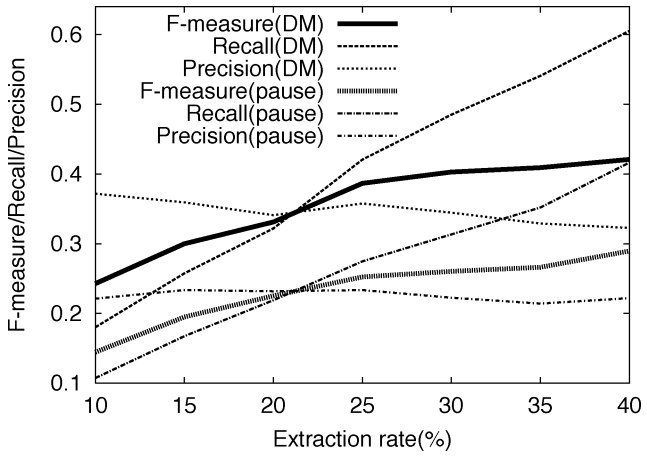

Fig. 7. Indexing performance of key sentences comparing the use of DM and pause duration.

VII. Evaluation of Key Sentence Extraction

A. Preliminary Evaluation

For a first evaluation, we used the same test set as in the previous section, and had a human subject select key sentences. The ratio of selected key sentences to the number of overall sentences is 21.6% $(= 233/1077)$. Therefore, we tried to extract 30% of the sentences in order to measure the performance. The evaluation measures we used are the recall, precision rates, and F-measure.

First, we evaluated the effect of the heuristic on the section structure and its automatic detection. The proposed method using the statistics of the DMs [function (3)] was evaluated when 30% of the sentences were extracted based on the score. The recall rate of correct key sentences was 48.5%.

For reference, when the same number of sentences were extracted from the beginning and end of the whole lecture, which corresponds to the introduction and conclusion, respectively, the recall rate was only 27.5%. When the section structure was segmented by a human expert and the same number of initial sentences of sections were extracted, the recall rate was 54.2%. These results show that identifying key sentences based on the section structure is a useful heuristic and that automatic detection of section boundaries provides comparable performance with only a little degradation relative to human segmentation.

For further comparison (=baseline), we also tested a method that detects key sentences based on the distribution and lengths of pauses only. The method takes advantage of the fact that putting a pause after an important sentence, as well as before it, is known to be an effective tactic in Japanese public speaking, as verified in studies by Nakamura *et al.* [27]. Here, all pause durations are converted to z-scores via $N(0,1)$ normalization. Then, for each sentence, the following pause is compared to the preceding pause, and the score of the longer one is used as a measure of importance for the sentence. The recall rate for this method was only 31.3%. Thus, the proposed method using the DMs is shown to be more effective in detecting section boundaries and extracting key sentences. Fig. 7 plots the recall, precision and F-measure for the proposed method based on the DM and the baseline method based on pause information (pause) for

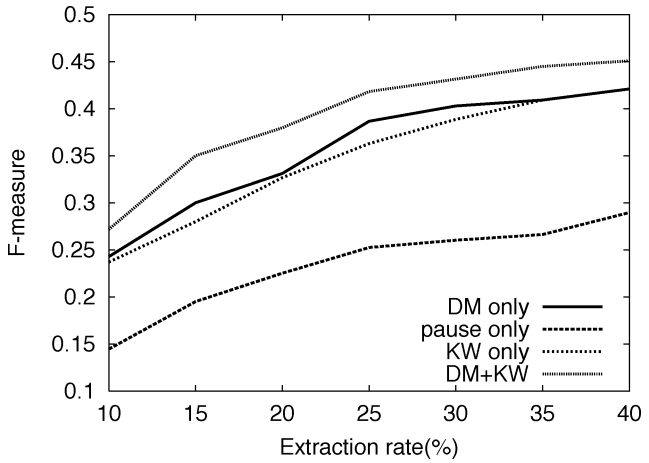

Fig. 8. Indexing performance of key sentences using DMs and KWs.

TABLE V
AGREEMENT AMONG SUBJECTS IN LABELING KEY SENTENCES

	by 2 persons	by 3 persons
agreement in 50% extraction	75%	60%
(ratio in total sentences)	(37.5%)	(30.0%)
agreement in 10% extraction	45%	30%
(ratio in total sentences)	(4.5%)	(3.0%)

sentence extraction rates of 10% to 40%. The superiority of the proposed method (DM) is confirmed by this graph.

Next, the method based on the DMs is compared and combined with the conventional method that focuses on topic-dependent KWs. Fig. 8 plots the F-measure for each method alone, as well as the combined method. The proposed method (DM) achieved slightly better performance than the KW-based method. Moreover, their combination significantly improved the performance. While the DM-based method tries to extract key sentences by focusing on the section boundaries, the key sentences that appear at other positions are picked up by the KW-based method. Thus, the features that the two methods capture are quite different and have a synergetic effect when combined.

B. Evaluation With the CSJ Key Sentence Set

We did a more thorough evaluation using another set of data. For part of the CSJ, key sentences labeled by human subjects will be included in the final published corpus. In this work, we used the key sentence labels that were available as of August 2003. Specifically, key sentences were labeled by three human subjects for nineteen academic presentation speeches. The subjects were researchers in linguistics, thus they were familiar with the academic presentation style, but were not professionals in the area of most of the test-set lectures. They were instructed to select sentences which seemed important, first choosing a set of 50% of the sentences in a lecture, and then choosing a subset of that 50% to make an overall 10% compression factor.

Table V shows the agreement rates among the three subjects in indexing. The agreement between two persons is the average of all combinations of the three. While a relatively high agreement is observed in the 50% extraction, it is much harder to get agreement in the 10% extraction. This is partly to be expected, since agreement in the 10% extraction is intrinsically more difficult (e.g. chance agreement between two persons is only 1% as compared to 25% in the 50% extraction). In any case, the very low number of agreed-upon sentences (3%–5%) is too small for our purpose of indexing key sentences in all sections, as previously described. By contrast, the number of agreed-upon sentences in the 50% extraction (30%–38%) matches the targeted compression ratio of 20%–40% very well. Therefore, we set up experiments based on the agreed portion of the 50% extraction data for reliable and meaningful evaluation. Specifically, we choose sets of sentences agreed upon by two subjects as our 'answer' sets. Since three combinations exist for taking two subjects out of three, we derived three answer sets. The performance is evaluated by averaging for these three sets. As shown in the table, they amount to 37.5% of all sentences on the average. Using this scheme, we can also estimate the human performance by matching one subject's selection with the answer set derived from the other two. The recall, precision and F-measure are 83.2%, 62.7%, and 0.715, respectively. These figures are regarded as a target for the proposed system.

TABLE VI
RESULTS OF KEY SENTENCE EXTRACTION FROM MANUAL TRANSCRIPTIONS (TEXT)

method	recall	precision	F-measure
DM	71.0%	53.3%	0.609
KW	71.7%	53.8%	0.614
DM+KW	74.0%	55.5%	0.635
human	83.2%	62.7%	0.715

DM: discourse marker (proposed), KW: keyword

The proposed method based on the DM and its combination with the KW-based method were evaluated on this test-set. Indexing performance of the key sentences (average for three sets) for the manual transcriptions is listed in Table VI. We confirmed much the same tendency as in Fig. 8. Although the method using the DMs alone was comparable to the KW-based method, the synergetic effect of using the two in combination was clearly verified.

When we compare the system performance against the human judgments, the accuracy by the system is lower by about 10%. The proposed method performs reasonably, but it still has room for improvement.

C. Evaluation With ASR Results

Finally, we made an evaluation using the transcriptions generated by the ASR system for the new test-set. Since the ASR results do not include periods, we first applied the period insertion procedure presented in Section III-A in order to segment each lecture into sentences. The indexing method is based on the combined importance score using both DMs and KWs (DM+KW).

Table VII lists the recall, precision and F-measure in comparison with the case applied to the manual transcriptions. Here, we also tested the case where the sentence segmentation or period insertion is done automatically on the manual transcriptions to tease apart the effects of the sentence segmentation from the

TABLE VII
RESULTS OF KEY SENTENCE EXTRACTION FROM ASR RESULTS

	transcription	segmentation	recall	precision	F-measure
(1)	manual	manual	74.0%	55.5%	0.635
(2)	manual	automatic	73.1%	45.8%	0.563
(3)	automatic	automatic	72.7%	45.6%	0.561

TABLE VIII
LIST OF TEST-SET LECTURES WITH SPEECH RECOGNITION ACCURACY, SEGMENTATION PERFORMANCE, AND INDEXING PERFORMANCE

lecture ID	recognition accuracy	segment accuracy	indexing accuracy
A01M0056	85.15%	0.821	0.458
A01M0096	91.21%	0.812	0.567
A01M0151	92.21%	0.920	0.656
A01M0035	64.95%	0.505	0.529
A01M0007	78.32%	0.613	0.533
A01F0001	77.56%	0.851	0.559
A01M0025	92.18%	0.878	0.671
A01M0110	86.15%	0.915	0.598
A01F0132	87.15%	0.794	0.495
A01M0083	91.35%	0.822	0.580
A01M0137	72.74%	0.740	0.561
A01M0074	80.54%	0.745	0.484
A01M0097	84.76%	0.844	0.536
A03M0112	81.41%	0.912	0.630
A03M0106	61.37%	0.720	0.489
A03F0072	71.31%	0.735	0.591
A05M0031	74.68%	0.783	0.629
A06M0134	68.58%	0.643	0.606
YG99JUN001	69.17%	0.512	0.501
total	76.99%	0.740	0.561

effects of using the ASR output instead of the manual transcription. Since the derived sets of sentences for automatic and manual segmentations are different, we automatically align the hypothesized sentences with the correct ones, and calculate the accuracy based on the alignment.

Comparing cases (1) and (2) in Table VII shows that automatic segmentation substantially lowers the precision rate. As shown in Table I, our period insertion algorithm has a much higher precision than recall, which means that correct sentences tend to be concatenated in the hypotheses. Therefore, the correct key sentences (=recall rate) are kept, while neighboring sentences are incorrectly indexed together, resulting in lower precision for the key sentence indexing. On the other hand, no further degradation is observed by adopting ASR even with this word error rate of 23%. These results demonstrate that the statistical evaluation of the importance of sentences is robust.

The detailed results for the individual lectures in the test-set are listed in Table VIII. Here, the indexing accuracy of the key sentences is shown with the word recognition accuracy and the sentence segmentation accuracy (=F-measure of period insertion). While recognition accuracy varies considerably across the speakers, and segmentation accuracy depends to some extent on recognition accuracy ($R^2 = 0.53$), indexing accuracy is fairly stable and independent of recognition accuracy ($R^2 = 0.09$).

VIII. CONCLUSIONS

We have presented automatic indexing methods for lecture audio materials. The methods assume a slide-based discourse structure and focus on DMs, defined as expressions characteristic to the initial utterances of section units. The set of DMs is chosen on the basis of statistics of the distribution of words relative to segment boundaries, in a completely unsupervised manner which does not need any manual tags. For detection of section boundaries, the method based on these DMs achieved a recall rate of 85% and a precision rate of 20%, which should be a sufficiently practical indexing for browsing long speech materials. It is also robust against speech recognition errors.

The method was extended to the task of extracting key sentences for more informative indexing. A measure of importance is defined based on the statistics that were used for deriving the DMs. This method achieved comparable performance to the conventional KW-based method. Moreover, the combination of the two methods significantly improves the accuracy because the two statistics focus on different characteristics of a lecture.

We have also described a statistical framework for transforming raw transcriptions of spontaneous lectures into the document style, and showed that incorporation of the transformation model into the conventional N-gram model is effective for the tasks of period and particle insertion. The methodology is useful for improving the readability of sentences of the digital archives. Experimental evaluation showed that the accuracy of sentence segmentation (period insertion) is more vital than the speech recognition accuracy in indexing key sentences.

Ongoing work includes application of the method to other domains such as lectures at universities and automatic annotation of more specific tags based on the "slide-based discourse structure" for development of a comprehensive digital archiving system.

ACKNOWLEDGMENT

This work was conducted as part of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology." The authors are grateful to Prof. S. Furui and other members of the project for their fruitful collaboration. The authors are also thankful to the Associate Editor, Prof. M. Beckman, for her extensive suggestions for polishing the argument and English of this paper.

REFERENCES

[1] T. Imai, R. Schwartz, F. Kubala, and L. Nguyen, "Improved topic discrimination of broadcast news using a model of multiple simultaneous topics," in *Proc. IEEE ICASSP*, 1997, pp. 727–730.

[2] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Video mail retrieval: the effect of word spotting accuracy on precision," in *Proc. IEEE ICASSP*, 1995, pp. 309–312.

[3] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek, "Approaches to topic identification on the SWITCHBOARD corpus," in *Proc. IEEE ICASSP*, vol. 1, 1994, pp. 385–388.

[4] S. Whittaker, J. Choi, J. Hirschberg, and C. H. Nakatani, "What you see is (almost) what you hear: design principles for user interfaces for accessing speech archives," in *Proc. ICSLP*, 1998, pp. 2355–2358.
[5] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE ICASSP*, vol. 1, 2001, pp. 597–600.
[6] S. Furui, K. Maekawa, and H. Isahara, "Toward the realization of spontaneous speech recognition—introduction of a Japanese priority program and preliminary results—," in *Proc. ICSLP*, vol. 3, 2000, pp. 518–521.
[7] T. Kawahara, H. Nanjo, and S. Furui, "Automatic transcription of spontaneous lecture speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2001, pp. 186–189.
[8] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1–6.
[9] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 135–138.
[10] K. Maekawa, "Corpus of Spontaneous Japanese: its design and evaluation," in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.
[11] C.-Y. Lin and E. H. Hovy, "Identifying topics by position," in *Proc. Applied Natural Language Conf.*, 1997, pp. 283–290.
[12] R. J. Passonneau and D. J. Litman, "Discourse segmentation by human and automated means," *Computat. Ling.*, vol. 23, no. 1, pp. 103–139, 1997.
[13] M. Haase, W. Kriechbaum, G. Mohler, and G. Stenzel, "Deriving document structure from prosodic cues," in *Proc. EUROSPEECH*, 2001, pp. 2157–2160.
[14] D. Schiffrin, *Discouse Markers*. Cambridge, U.K.: Cambridge Univ. Press, 1987.
[15] J. Hirschberg and D. Litman, "Empirical studies on the disambiguation of cue phrases," *Computat. Ling.*, vol. 19, no. 3, pp. 501–530, 1993.
[16] M. Kawamori, T. Kawabata, and A. Shimazu, "Discourse markers in spontaneous dialogue: a corpus based study of Japanese and English," in *Proc. Workshop Discourse Relations & Discourse Markers (ACL-COLING)*, 1998, pp. 93–99.
[17] F. M. Quimbo, T. Kawahara, and S. Doshita, "Prosodic analysis of fillers and self-repair in Japanese speech," in *Proc. ICSLP*, 1998, pp. 3313–3316.
[18] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.
[19] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. IEEE ICASSP*, 2002, pp. 725–728.
[20] H. Nanjo, K. Shitaoka, and T. Kawahara, "Automatic transformation of lecture transcription into document style using statistical framework," in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 215–218.
[21] M. Murata and H. Isahara, "Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language," in *Proc. Int. Conf. Language Resources and Evaluation*, Las Palmas, Spain, 2002.
[22] I. Garcia-Varea, F. Casacuberta, and H. Ney, "An iterative, DP-based search algorithm for statistical machine translation," in *Proc. ICSLP*, 1998, pp. 1135–1138.
[23] K. Kato, H. Nanjo, and T. Kawahara, "Automatic transcription of lecture speech using topic-independent language modeling," in *Proc. ICSLP*, vol. 1, 2000, pp. 162–165.
[24] I. Mani and M. Maybury, Eds., *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.
[25] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 9–12.
[26] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computat. Ling.*, vol. 28, no. 4, pp. 409–445, 2002.
[27] M. R. Draguna, T. Nakamura, C. Nagaoka, and M. Komori, "Vocal strategies in expressing importance," in *Proc. Human Interface Symp.*, Sapporo, Japan, 2002, pp. 459–462.

Tatsuya Kawahara (M'91) received the B.E. degree in 1987, the M.E. degree in 1989, and the Ph.D. degree in 1995, all in information science, from Kyoto University, Kyoto, Japan.

In 1990, he became a Research Associate with Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor with the Academic Center for Computing and Media Studies, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories. He has published more than 100 technical papers covering speech recognition, confidence measures, and spoken dialogue systems. He has been managing several speech-related projects in Japan, including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. Since 2003, he has been a member of the IEEE Signal Processing Society's Speech Technical Committee.

Masahiro Hasegawa received the B.E. degree in 2001 and the M.E. degree in 2003 from Kyoto University, Kyoto, Japan. Currently, he is with Asahi Broadcasting Corporation, Osaka, Japan.

Kazuya Shitaoka received the B.E. degree in 2002 from Kyoto University, Kyoto, Japan. Currently, he is a Master Course student at School of Informatics, Kyoto University.

Tasuku Kitade received the B.E. degree in 2003 from Kyoto University, Kyoto, Japan. Currently, he is a Master Course student at School of Informatics, Kyoto University.

Hiroaki Nanjo received the B.E. degree in 1999 and the M.E. degree in 2001 from Kyoto University, Kyoto, Japan, where he is currently pursuing the Ph.D. degree in the School of Informatics. He has been working on speech recognition and understanding.

Mr. Nanjo is a member of Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Information Processing Society of Japan (IPSJ).