Spoken Language Processing for Audio Archives of Lectures and Panel Discussions

Tatsuya Kawahara
Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp
http://www.ar.media.kyoto-u.ac.jp/

Abstract

Intelligent archiving systems of lectures and panel discussions based on automatic transcription and indexing are introduced. Our speech recognition system Julius *has been improved to deal with these kinds of spontaneous speech. Transcriptions are automatically edited for improving readability, and then key sentences are indexed for each segment based on statistically-derived discourse markers and topic words. For panel discussions consisting of multiple speakers, unsupervised speaker indexing is applied beforehand to segment audio into speaker turns. Thus, we realize efficient browsing of these audio archives.*

1. Introduction

Recent progress of large-volume storage devices and high-speed networks has enabled digital archiving and streaming of audio and video materials. In academic societies and universities, multi-media archives of lectures and panel discussions will be technically feasible. Such archives would help students audit lectures at their convenient time and places with their own paces. Similar digital archives can be considered for discussions and debates including those of the national assembly and courtrooms as well as for general business meetings. In these kinds of audio archives, appropriate indices are necessary for efficient browsing and searching portions of specific topics or speakers. Conventionally, the indexing and annotation are manually done and require a lot of time and cost. Spoken language technologies will be useful for (semi-)automating the indexing process; they are applied to generate 'rough & ready' indices, which can be corrected by a manual post-processing if necessary.

The paper gives an overview of our studies on various aspects of spoken language processing toward the intelligent archiving of speech materials. While most of the previous studies dealt with broadcast news[1][2] that are of interest to many people and rather technically easy, we focus on lecture presentations and panel discussions, which are major real-world materials in academic communities. In these materials, speech information seems more dominant than visual information, and thus spoken language technologies will have important roles.

2. Overview of Lecture Archiving System

Lecture presentations are typically long monologue of dozens of minutes, while the broadcast news consists of short clips, each containing a few minutes of speech. Moreover, topics of news clips are completely different even in a successive sequence, and the broad categories of topics can be determined a priori. These characteristics make possible the conventional topic classification and segmentation approach that relies on keywords. On the other hand, a different approach is needed for indexing lecture materials, during which one broad topic remains unchanged while closely-related small sub-topics succeed each other. For this kind of material, a browsing function is essential[3][4]. Specifically, exact time indices for boundaries of sub-topics or 'sections' are required, since such indices can be used to locate segments to be replayed.

We approach the problem of indexing lecture audio archives by assuming a discourse structure of 'sections' and automatically detecting their boundaries. We focus on 'discourse markers', which are rather topic independent and defined as expressions characteristic of the beginning of new sections. Then, from each section we extract key sentences that can be used as content-based tags for the corresponding audio segments. The alignment of audio segments and transcriptions is also obtained as the result of automatic speech recognition.

Based on the approach, we are developing an intelligent lecture archiving system. An overview of the system is depicted in Figure 1. First, whole speech is automatically transcribed by an automatic speech recognition (ASR) system. The transcriptions are automatically transformed to document-style sentences for improved readability. Then, the discourse segmentation into section units is performed and key sentences are indexed for each section. Collection of these sentences might also suffice a summary of the talk. In the generated archive, the index sentences are hyper-linked with the segmented audio for easy browsing.

The processes involve the following issues of spoken language processing.

(1) Automatic transcription of spontaneous speech

While automatic speech recognition of read speech has achieved accuracy exceeding 90%, spontaneous speech recognition faces difficult problems of acoustic and linguistic variations which are yet to be solved.

(2) Automatic segmentation of a lecture into sentences and sections

Baseline indexing for quick browsing of lecture audio is done by sentence segmentation. We realize the process in a framework of transforming (or translating) the raw transcription into document style. Moreover, we conduct segmentation into sections by assuming a lecture-style discourse structure[5]. The method is based on presumed discourse markers that are derived in an unsupervised manner.

(3) Automatic indexing of key sentences

More elaborate indexing for efficient browsing is realized by extracting key sentences, as they concisely express topics of the segments. We introduce a statistical measure of importance of sentences by focusing on both discourse markers and topic words.

These are described more in detail in the following three sections.

3. Automatic Transcription of Spontaneous Speech

Oral presentations and panel discussions are regarded as in-between of broadcast news and telephone conversation, both of which are widely dealt with so far in speech transcription projects. The speaker is not professional, nor reading a draft material as in broadcast news. But the speaking style is not so casual as in telephone conversation.

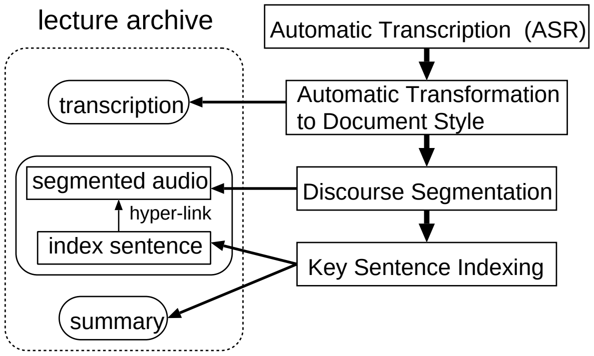


Figure 1. System overview of lecture archiving

We have taken part in the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan[6][7]. The *Corpus of Spontaneous Speech (CSJ)*[8] developed by the project consists of roughly 7M words or 500 hours, which is the largest in scale and provided us with an infrastructure for our automatic speech recognition (ASR) research.

As many previous studies point out, various factors in spontaneous speech affect ASR performance. They include acoustic variation caused by fast speaking and imperfect articulation, and linguistic variation such as colloquial expressions and disfluencies. Thus, the problems should be addressed from the viewpoint of acoustic, pronunciation and language modeling.

We also revised our open-source speech recognition software Julius [1] so that very long speech can be handled without prior segmentation[9].

3.1. Acoustic Model

We have set up a variety of baseline acoustic models[10]. The training data consist of 781 presentations that amount to 106 hours of speech.

Acoustic models are based on diagonal-covariance Gaussian-mixture HMM. The number of phones used is 43. We trained a PTM (phonetic tied-mixture) triphone model[11]. As a whole, there are 25K Gaussian components and 576K mixture weights.

Increase of training data thanks to the increased size of the CSJ consistently, though modestly, improved the word accuracy. For reference, the standard read speech model[12] obtained lower accuracy by about 10% absolute.

[1] downloadable at `http://julius.sourceforge.jp`

3.2. Language and Pronunciation Model

A baseline language model is constructed using the transcriptions of 2592 talks. The total text size is about 6.67 million words including fillers and word fragments. Word segmentation was automatically done using a morphological analyzer that was trained with the maximum entropy criterion[13].

In spontaneously spoken Japanese, pronunciation variation is so large that multiple pronunciation entries are needed for a lexical item. We found that statistical modeling of pronunciation variations integrated with the language modeling was effective in suppressing false matching of less frequent entries[14, 15]. Here, we adopt a simple trigram model of word-pronunciation entries.

The increase of training data size had even stronger impact than the case of the acoustic model. WER (Word Error Rate) is significantly reduced according to the increase of the data. The result strongly demonstrates that the corpus of this scale is meaningful in modeling spoken language.

3.3. Model Adaptation and Speaking Rate Dependent Decoding

Next, we incorporate speaker adaptation of acoustic and language models. Since lecture speech has long duration (large data) per speaker, the unsupervised adaptation scheme works very well.

First, we generate transcriptions for the test utterances using the baseline speaker-independent model. For acoustic model, MLLR (Maximum Likelihood Linear Regression) adaptation of Gaussian means is performed using the phone labels of the initial recognition result, and a speaker-adapted model is generated.

We have also studied unsupervised methods of language model adaptation to a specific speaker and a topic[14, 15], which are based on a model trained with the initial transcription. The first method is to select similar texts using the word perplexity and TF-IDF measure and weight them in re-training. The second method makes direct use of the model generated from the initial recognition result by linear interpolation with the baseline model.

We also proposed a decoding strategy adapted to the speaking rate[16, 15]. In spontaneous speech, speaking rate is generally fast and may vary a lot within a presentation. We also observe different error tendencies for portions of presentations where speech is fast or slow. The proposed speaking rate dependent decoding strategy applies the most appropriate acoustic analysis, phone models, and decoding parameters according to the speaking rate.

method	WER
baseline	30.9
+ acoustic model adaptation	26.0
+ language model adaptation	23.9
+ speaking rate adaptation	22.0

Table 1. Word Error Rate (WER) in automatic lecture transcription

The effect of these methods for the task of transcription of 15 academic presentations is summarized in Table 1. The unsupervised acoustic model adaptation reduced WER by 4.9% absolute from 30.9% to 26.0%, and the combination with the language model adaptation methods reduced WER further by 2.1% absolute. The speaking rate dependent decoding strategy gained additional improvement of 1.9% absolute. Finally, WER of 22.0% is achieved.

4. Automatic Transformation of Transcription into Document Style

Transcriptions of spontaneous speech include many colloquial expressions peculiar to spoken language. The Japanese spoken language in particular is quite different from the written language, and is not suitable for documents in terms of readability. Thus, it is necessary to transform transcriptions or speech recognition results into document style for practical archives. This process is also important as a pre-process of automatic summarization.

We approach the problem by using a statistical framework that has become popular in machine translation. We regard the spoken and written Japanese languages as different languages and apply the translation methodology to transform the former into the latter. Within this framework, correction of colloquial expressions, deletion of fillers, insertion of periods (end-of-sentence symbols), and insertion of particles are performed in an integrated manner[17].

The statistical machine translation framework is formulated by finding the best output sequence Y for an input sequence X, such that a posteriori probability $P(Y|X)$ is maximum. According to Bayes rule, maximization of $P(Y|X)$ is equivalent to the maximization of the product (sum in log scale) 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 correspondence of input and output word sequences.

In the task of style conversion, the input X is a word sequence of spoken language transcription that does not have periods but includes pause duration. The output Y is a word sequence of the written language. For $P(Y)$ calculation, we use a word 3-gram model trained with a written language corpus. Since the conversion of one word affects neighbor words in the N-gram model, decoding is performed for a whole input word sequence with beam pruning.

5. Automatic Indexing of Lecture Presentations using Discourse Markers

Next, we address automatic segmentation of lecture audio and then extraction of key sentences, which will be useful indices. The framework extracts a set of natural sentences to be scanned, which are aligned with audio segments to be replayed.

5.1. Discourse Modeling of Lecture Presentations

There is a relatively clear prototype in the flow of presentation, which is similarly observed in technical papers[18]. When using slides for presentation, one or a couple of slides constitute a topic discourse unit we call 'section' in this paper. The unit in turn usually corresponds to the (sub-)sections in the proceedings paper.

It is also observed that there is a typical pattern in the first utterances of the units. Speakers try to briefly tell what comes next and attract audiences' attention; for example, "Next, I will explain how it works." and "Now, let's move on to experimental evaluation". We define such characteristic expressions that appear at the beginning of section units as discourse markers. We have proposed a method to automatically train a set of discourse markers without any manual tags, and shown the effectiveness in segmentation of lecture audio[19, 5]. Then, we apply the discourse segmentation to extraction of key sentences from lectures[20, 5] for generating more informative tags of the indices.

5.2. Segmentation using Discourse Markers

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. Here, we set a threshold on pause duration to pick up boundary candidates. We use the average pause length during a talk as the threshold.

From the candidate first sentences picked up by the pause information, we extract characteristic expressions, namely select discourse markers useful for indexing. Discourse markers should frequently appear in the first utterances, but should not appear in other 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 its occurrence count in the set of first sentences. The sentence frequency sf_j is the number of sentences in all lectures that contain the word. We adopt the following evaluation function.

$$S_{DM}(w_j) = wf_j * \log(\frac{N_s}{sf_j}) \qquad (1)$$

Here, N_s is the total number of sentences in all lectures. A set of discourse markers are selected by the order of $S_{DM}(w_j)$. For a given new lecture, a candidate section boundary is indexed if the summed score over all markers appearing in the following sentence s_i, i.e., $\sum_{w_j \in s_i} S_{DM}(w_j)$ is larger than a certain threshold.

5.3. Measure of Importance based on Discourse Markers

In the text-based natural language processing, a well-known heuristic for key sentence extraction is to pick up initial sentences of the articles or paragraphs. Using the automatically-derived discourse markers that characterize the beginning of sections, the heuristic is now applicable to speech materials.

The importance of sentences is evaluated using the same function (equation (1)) that was used as appropriateness of discourse markers. For each sentence s_i, we compute a sum score $S_{DM}(s_i) = \sum_{w_j \in s_i} S_{DM}(w_j)$.

Then, key sentences are selected based on the score up to the specified number (or ratio) of sentences from the whole lecture.

The other approach to extraction of key sentences is to focus on keywords that are characteristic to the lecture. The most orthodox statistical measure to define and extract such keywords is the following TF-IDF criterion.

$$S_{KW}(w_j) = tf_j * \log(\frac{N_d}{df_j}) \qquad (2)$$

Here, term frequency tf_j is the occurrence count 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. For each sentence s_i, we compute $S_{KW}(s_i) = \sum_{w_j \in s_i} S_{KW}(w_j)$.

method	recall	precision	F-measure
DM	69.9%	51.7%	0.594
KW	70.4%	52.0%	0.598
DM+KW	72.4%	53.5%	0.615
human	81.5%	60.1%	0.692

DM: discourse marker, KW: keyword

Table 2. Performance of key sentence indexing (text)

Then, we introduce a new measure of importance that combines the two measures by taking a geometric mean with a weight α.

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

5.4. Experimental Evaluation of Key Sentence Indexing

For part of the CSJ, key sentences labeled by human subjects are included. In this work, we made use of 21 academic presentations that are also used for evaluation of automatic speech recognition (ASR)[10]. A set of key sentences were labeled by three human subjects. They were instructed to select sentences which seemed important by 50% of all, and then 10% from those 50%.

We set up experiments based on the agreed portion of the 50% extraction data. Specifically, we picked up sets of sentences agreed upon by two subjects. Since three combinations exist for picking up two subjects out of three, we derived three answer sets. The performance is evaluated by averaging for these three sets. 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.

The proposed method based on the discourse markers (DM) and its combination with the keyword-based method (KW) were evaluated. Indexing performance of the key sentences for manual transcriptions is listed in Table 2. The method using the discourse marker (DM) was comparable to the keyword-based method (KW), and the synergetic effect of their combination (DM+KW) was clearly confirmed. When we compare the system performance against the human judgment, the accuracy by the system is lower by about 10%. The proposed method performs reasonably, but it still has room for improvement.

transcript	segment	recall	precision	F-measure
manual	manual	72.4%	53.5%	0.615
manual	automatic	72.7%	46.5%	0.567
automatic	automatic	76.1%	45.5%	0.569

Table 3. Performance of key sentence indexing (ASR results)

Then, we made evaluation using the transcriptions generated by the ASR system. Since ASR results do not include periods, we incorporate the automatic period insertion procedure presented in Section 4 in order to segment the lecture into sentences. The indexing method is based on the discourse marker and keyword combination (DM+KW). Table 3 lists the recall, precision and F-measure in comparison with the case of manual transcription. Here, we also tested the case where the sentence segmentation or period insertion is done automatically on the manual transcriptions to see individual effects. It is observed that the automatic segmentation lowered the accuracy, especially the precision rate. On the other hand, no degradation is observed by adopting automatic speech recognition even with the word error rate of around 30%. These results demonstrate that the statistical evaluation of the importance of sentences is robust.

6. Overview of Panel Discussion Archiving System

The archiving scheme for panel discussions differs from that for lecture presentations because there are multiple speakers in discussions and speaker information is critical in segmenting and indexing the audio materials. Namely, the most appropriate segmentation will be given by speaker turns. The indexing scheme should consider the relationship between speaker turns in terms of discourse. In typical panel discussions, moreover, a chairperson plays a distinct role in controlling the agendas and speaker turns.

When browsing panel discussion archives, we assume that users are interested in the topic of the discussion and opinions of the individual panelists in regard to the topic. Thus, when developing such an archive, this kind of hierarchy must be considered for topic and speaker segmentation. Appropriate indices will enable users to access to the topics being discussed directly, and then allow browsing the opinions of each panelist based on speaker indices, while more detailed information can be gained by accessing to the transcription and audio data.

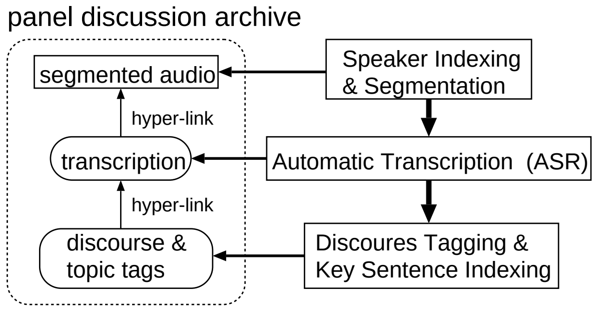

panel discussion archive
segmented audio
Speaker Indexing & Segmentation
hyper-link
transcription
Automatic Transcription (ASR)
hyper-link
discourse & topic tags
Discoures Tagging & Key Sentence Indexing

Figure 2. System overview of panel discussion archiving

For this application, we address the following issues of spoken language processing in addition to those mentioned for the lecture presentations.

(1) Automatic speaker indexing

Previous speaker indexing approaches often used a supervised training scheme that requires sufficient training data for speaker models beforehand. They are not practical for panel discussions where new speakers participate at every occasion. Therefore, we have studied an unsupervised approach[21, 22][23] which does not require any prior information on speakers and automatically clusters audio segments into participating speakers.

(2) Automatic discourse tagging

Rather than only performing topic segmentation, we explore to generate discourse tags to each speaker's turn such as questions, assertions and opinions. These kinds of discourse information provide effective keys for information retrieval from panel discussions.

An overview of the proposed automatic archiving system is shown in Figure 2. The system consists of three stages: speaker indexing, automatic transcription (ASR), and discourse tagging. First, unsupervised speaker indexing is conducted and the resulting speaker indices are used to train speaker-adaptive acoustic models, which are used for speech recognition. Discourse tagging is then applied to the generated transcription using both rule-based and statistical methods. Finally, the resulting archival data consisting of the original audio, transcriptions, speaker indices and discourse tags are integrated within a framework of MPEG-7 encoding.

7. Automatic Speaker Indexing

In the first stage of the archiving process, each utterance is indexed by an individual speaker. We have proposed two different approaches of unsupervised indexing oriented for meetings and panel discussions.

One approach[23] is based on *anchor models* or a large set of speaker models. Approximately 300 anchor GMMs (Gaussian Mixture Models) are initially created, each trained on a specific speaker from a large-scale speech database. First, likelihood vectors are generated for each input utterance by calculating the likelihood against all anchor models. These likelihood vectors are then automatically clustered using LBG algorithm and the resulting classes are used to train speaker classification models. Finally, the generated models are used to perform speaker indexing. On the evaluation test-set of panel discussions collected from a TV program “Sunday Discussion”, an average indexing accuracy of 97% was achieved using this approach.

The other approach[21, 22] incrementally clusters audio segments with a model-based distance measure while automatically refining the speaker models. Specifically, we have proposed a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the Bayesian Information Criterion (BIC) according to the amount of training data. The framework makes it possible to use a discrete model (VQ) when the data is sparse, and to seamlessly switch to a continuous model (GMM) after a large amount of data is obtained. For the same evaluation test-set, the method also achieved indexing accuracy of 97%. Moreover, the method works even without prior information of the number of speakers.

The speaker indices are also used to adapt a speaker-independent acoustic model to each participant for automatic transcription of the discussions. We demonstrated that the automatic speaker indexing is sufficiently accurate for adaptation of the acoustic model; the adapted model improved the word accuracy from 51% to 57%, which is comparable to the case of supervised model adaptation using manual speaker labels.

8. Automatic Discourse Tagging of Panel Discussions

For each speaker turn, we generate more informative tags based on discourse information.

A panel discussion is composed of a chairperson who presides over the discussion, and several panelists who typically have conflicting opinions on the given topic. A chairperson introduces (sub-)agendas and gives a brief overview, and then prompts panelists for their opinions, while panelists state their own opinions and may also query other panelists.

We analyzed typical panel discussions and observed that key sentences consisting of characteristic discourse

type	description
Suggestion	Expediting proceedings
Confirmation	Confirmation by chairperson
Question	Initial question
Opinion	Giving one's opinion
Answer	Answer to *Question*
Agenda	(Sub-)topic of discussion

Table 4. Proposed set of discourse tags

Chairperson	recall	precision
Agenda	92.6%	96.3%
Question	99.1%	96.3%
Suggestion	27.8%	15.2%
Confirmation	100.0%	42.9%
Panelists	recall	precision
Opinion (key sentences)	74.7%	51.1%

Table 5. Performance of discourse tagging

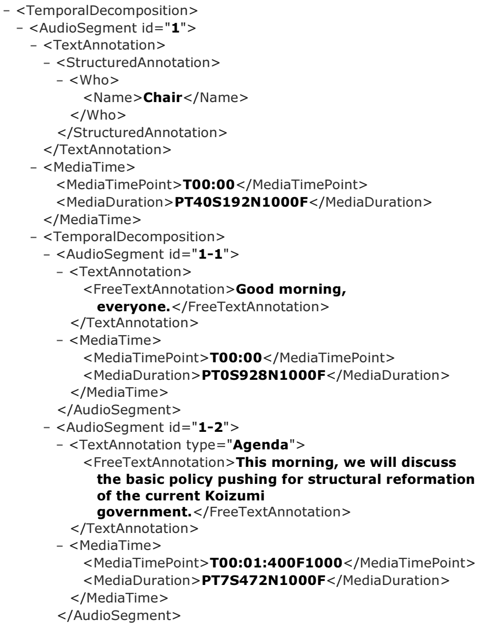
markers exist in each turn and that they will provide effective indexing. Based on this analysis, a set of discourse tags were defined, as listed in Table 4. Besides indexing purposes, these discourse tags might also be useful for automatic summarization.

For effective discourse tagging, the speaker's role within the discussion is considered. Thus, in panel discussions the chairperson and panelists should be treated differently. For the chairperson, rule-based discourse tagging is applied. Rules for *Agenda, Question, Confirmation* and *Suggestion* tags were manually crafted.

For panelists' utterances, we set up *Question*, *Answer* and *Opinion* tags. *Question* and *Answer* tags are given by heuristic rules. An *Opinion* tag is attached to key sentences of panelists' utterances based on the statistically-derived discourse markers as adopted for lecture presentations, although a different set of markers is estimated with evaluation function (1) using the discussion corpus.

The effectiveness of the proposed discourse-based indexing method is investigated on the transcriptions of the evaluation test-set. The recall and precision rates for various discourse types are listed in Table 5. High recall rates were achieved for all discourse types except *Suggestion*. These tagged key sentences are useful for efficient browsing or retrieval.

In the final stage, an archive is constructed by incorporating the original audio, speaker indices, transcriptions and discourse tags generated in these processes. They are combined using an MPEG-7 framework to generate an archive as shown in Figure 3. Since MPEG-7 is based on an XML format, popular XML-based software including parsers, browsers and editors can be used to access or edit the generated data. The style-sheet framework (XSL) also allows the visual presentation of the archive to be altered easily.



```
- <TemporalDecomposition>
  - <AudioSegment id="1">
    - <TextAnnotation>
      - <StructuredAnnotation>
        - <Who>
            <Name>Chair</Name>
          </Who>
        </StructuredAnnotation>
      </TextAnnotation>
    - <MediaTime>
        <MediaTimePoint>T00:00</MediaTimePoint>
        <MediaDuration>PT40S192N1000F</MediaDuration>
      </MediaTime>
    - <TemporalDecomposition>
      - <AudioSegment id="1-1">
        - <TextAnnotation>
            <FreeTextAnnotation>Good morning,
              everyone.</FreeTextAnnotation>
          </TextAnnotation>
        - <MediaTime>
            <MediaTimePoint>T00:00</MediaTimePoint>
            <MediaDuration>PT0S928N1000F</MediaDuration>
          </MediaTime>
        </AudioSegment>
      - <AudioSegment id="1-2">
        - <TextAnnotation type="Agenda">
            <FreeTextAnnotation>This morning, we will discuss
              the basic policy pushing for structural reformation
              of the current Koizumi
              government.</FreeTextAnnotation>
          </TextAnnotation>
        - <MediaTime>
            <MediaTimePoint>T00:01:400F1000</MediaTimePoint>
            <MediaDuration>PT7S472N1000F</MediaDuration>
          </MediaTime>
        </AudioSegment>
```

Figure 3. Sample annotation in MPEG-7 format

9. Conclusions and Ongoing Works

The paper has described fully automatic archiving systems focusing on lecture presentations and panel discussions, and their underlying technologies of spoken language processing: spontaneous speech recognition, unsupervised speaker indexing, automatic style conversion, and automatic indexing based on discourse markers presumed for these applications.

We are now developing a prototype system, which will be evaluated with respect to not only the effects of the component techniques but also user interfaces of the total system. We expect that such an evaluation will give us directions on refining a set of indices and tags for more efficient browsing of these kinds of audio archives.

Acknowledgments: The work was conducted with the contribution of many graduate students including Mr. H.Nanjo, Mr. Y.Akita, Mr. K.Shitaoka and Mr. T.Kitade.

References

[1] F.Kubala, S.Colbath, D.Liu, A.Srivastava, and J.Makhoul. Integrated technologies for indexing spoken language. *Communications of the ACM*, 43(2), 2000.

[2] Y.Hayashi, K.Ohtsuki, K.Bessho, O.Mizuno, and Y.Matsuo. Speech-based and video-supported indexing of multimedia broadcast news. In *Proc. ACM SIG-IR*, 2003.

[3] 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*, pages 2355–2358, 1998.

[4] 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*, volume 1, pages 597–600, 2001.

[5] T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, page (accepted for publication), 2004.

[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*, volume 3, pages 518–521, 2000.

[7] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–6, 2003.

[8] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[9] T.Kawahara, H.Nanjo, and S.Furui. Automatic transcription of spontaneous lecture speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[10] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 135–138, 2003.

[11] A.Lee, T.Kawahara, K.Takeda, and K.Shikano. A new phonetic tied-mixture model for efficient decoding. In *Proc. IEEE-ICASSP*, pages 1269–1272, 2000.

[12] T.Kawahara, A.Lee, T.Kobayashi, K.Takeda, N.Minematsu, S.Sagayama, K.Itou, A.Ito, M.Yamamoto, A.Yamada, T.Utsuro, and K.Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, volume 4, pages 476–479, 2000.

[13] K.Uchimoto, C.Nobata, A.Yamada, S.Sekine, and H.Isahara. Morphological analysis of Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 159–162, 2003.

[14] H.Nanjo and T.Kawahara. Unsupervised language model adaptation for lecture speech recognition. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 75–78, 2003.

[15] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, page (accepted for publication), 2004.

[16] H.Nanjo and T.Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. IEEE-ICASSP*, pages 725–728, 2002.

[17] H.Nanjo, K.Shitaoka, and T.Kawahara. Automatic transformation of lecture transcription into document style using statistical framework. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 215–218, 2003.

[18] S.Teufel and M.Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[19] T.Kawahara and M.Hasegawa. Automatic indexing of lecture speech by extracting topic-independent discourse markers. In *Proc. IEEE-ICASSP*, pages 1–4, 2002.

[20] H.Nanjo, T.Kitade, and T.Kawahara. Automatic indexing of key sentences for lecture archives using statistics of presumed discourse markers. In *Proc. IEEE-ICASSP*, page (to appear), 2004.

[21] M.Nishida and T.Kawahara. Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion. In *Proc. IEEE-ICASSP*, volume 1, pages 172–175, 2003.

[22] M.Nishida and T.Kawahara. Speaker model selection based on Bayesian information criterion applied to unsupervised speaker indexing. *IEEE Trans. Speech & Audio Process.*, page (accepted for publication), 2004.

[23] Y.Akita and T.Kawahara. Unsupervised speaker indexing using anchor models and automatic transcription of discussions. In *Proc. EUROSPEECH*, pages 2985–2988, 2003.