Automatic Audio Archiving System for Panel Discussions

Yuya Akita Masahiro Hasegawa* Tatsuya Kawahara
School of Informatics, Kyoto University, Kyoto 606-8501, JAPAN
akita@ar.media.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

We present an automatic audio archiving system suitable for panel discussions. In our archive framework, audio data, transcription of speech, speaker and content based indices are integrated in order to realize efficient archive browsing. Speaker indexing is performed in a totally unsupervised manner. The speaker information is also used for enhancing the automatic speech recognition system. These results are aligned with audio segments. Moreover, we also introduce a novel indexing of utterances based on discourse tags that represent intentions and importance of utterances. A discourse tagger combining rule based and statistical methods is developed to automatically generate high-level indices. Finally, these results are combined and encoded using an MPEG-7 framework, resulting in highly portable archives.

1. Introduction

With recent improvements in storage and computing technologies, it has become possible to archive vast amounts of information including that generated from lectures, meetings, and all forms of broadcast media. However, the archiving of such material is of no use if there are no methods to efficiently browse these vast information sources. Traditional approaches have tended to archive minutes from important meetings such as the National Congress, however the generation of such data is extremely expensive and the transcription processes tend to lose the nuances that exist in the original speech audio. Improved archiving can be achieved by the original speech directly.

Various speech archiving systems have been studied, including those for broadcast news[1, 2], lectures[3], meetings[4], and monologue[5]. As well as transcription of the speech material, effective archive indexing is essential. Previous works have typically focused on topic indices which are not necessarily effective for rapid understanding of long speech material such as panel discussions. Rather than focusing just on topic-based segmentation, the importance and discourse of utterances must also be considered.

Most of these systems use proprietary formats for encoding. Propriety formats offer low portability and typically limit the user to a proprietary browser or search interface. Proprietary formats also make the transfer of information between systems difficult. For improved portability standards such as MPEG-7[6] have recently been proposed which provide a common framework to handle multimedia information. However, the number of archives based on this framework is still extremely limited.

In this paper, we propose an automatic archiving system for speech-based material such as panel discussions. The archiving system combines speech recognition, speaker indexing, discourse tagging and MPEG-7 encoding. To improve information retrieval efficiency, we introduce a novel indexing method based on discourse tagging. First we analyze typical panel discussions to determine appropriate discourse structures for information retrieval, then we develop a discourse tagger that is used to automatically generate discourse-based indices which can be used for information retrieval.

2. Panel discussion archiving

The advantage of archiving speech material is to maintain the nuances of the content to be archived. The archiving schemas for broadcast news and lecture material will differ from that used for panel discussions as these sessions typically have a monologue based discourse structure with limited changes of topic.

2.1. Discourse of panel discussion

A panel discussion is composed of a chairperson who presides over the discussion and provides discussion themes and background information, and several

*Currently with Asahi Broadcasting Corporation.

0-7803-8603-5/04/$20.00

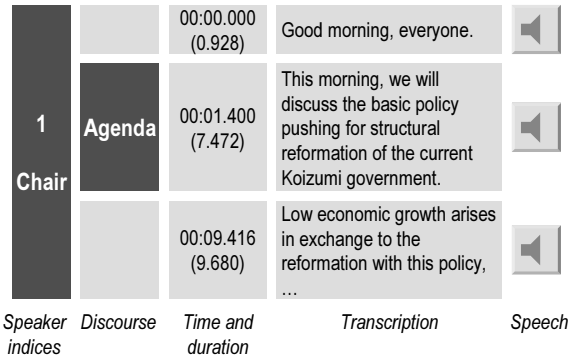

Figure 1. An example of archive

panelists who typically have conflicting opinions on the given topic. Panel discussions generally involve the chairperson providing a question or statement, and then prompting each panelist for their opinion. Panelists state their own opinion and may also query other panelists. A topic change is usually announced by the chairperson.

When browsing panel discussion archives, we assume users are interested in the topic of discussion and opinions of the individual panelist in regard to this topic. Thus, when developing such an archive, discourse structure must be considered. Generating appropriate indices will enable users to access to the topics being discussed directly, and then allow browsing the opinions of each panelist in regards to these topics while more detailed information can be gained by browsing the transcriptions or audio data.

For improved archiving effectiveness, we propose an archive schema as illustrated in Figure 1. The archive will consist of the original audio, speech transcriptions and a set of indices. These indices must allow for quick retrieval of important information while the vast amount of unnecessary data is retained in the automatic transcriptions.

2.2. Archive indices

In the proposed archiving system two sets of indices are created; speaker indices which are generated by unsupervised speaker indexing, and discourse-based indices generated from automatic discourse tagging.

For long speech material with multiple speakers, speaker indices are necessary to quickly obtain information on a person of interest. Previous approaches[1, 4] often used supervised based methods which provide high indexing accuracy but require sufficient training data beforehand. For panel discussions this approach is unpractical as participants frequently change, making the collection of such data extremely difficult. To overcome this problem, an unsupervised approach is adopted

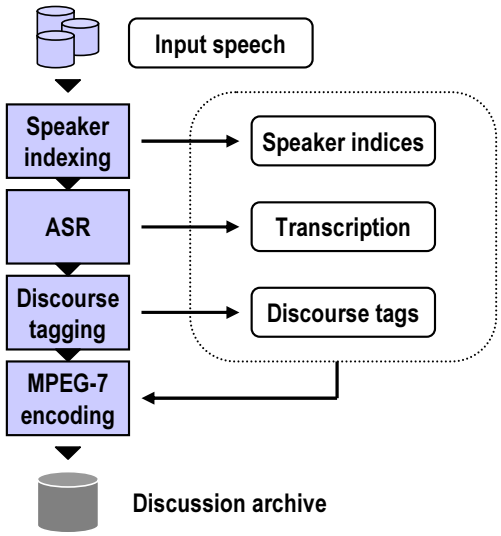

Figure 2. Process flow of proposed system

based on our previous work [7].

The second set of indices relates to content based information, and is based on discourse tagging. Rather than only performing topic segmentation, we use discourse information to provide indices to discussion themes, and panelists' questions, assertions, and opinions. These discourse structures are necessary for effective information retrieval for the panel discussions.

2.3. Overview of archive system

An overview of the proposed automatic archiving system is shown in Figure 2. The system consists of 4 main stages; speaker indexing, ASR (Automatic Speech Recognition), discourse tagging, and archive encoding. The proposed system is used to automatically index and encode audio data from panel discussions. First, unsupervised speaker indexing is conducted and the resulting speaker indices are used to train speaker-adaptive acoustic models, which are then used for speech recognition. Discourse tagging is then applied to the recognition transcription using both rule based and statistical methods. Finally the resulting archival data, consisting of the original audio, transcribed speech, speaker indices and discourse tags are integrated using MPEG-7 based encoding.

3. Components and their evaluation

In this section, each stage within the archiving system is described in detail.

3.1. Evaluation test-set

For evaluation, we use televised panel discussions from "Sunday Discussion", NHK (Japan Broadcasting Corporation). This program presents one hour discussions on political and economic topics. Panelists consist of politicians, economists and experts from relevant fields. The input audio is initially segmented into individual utterances by segmenting where short pauses longer than 400 milliseconds exist. The evaluation test-set consists of ten panel discussions and approximately 5500 utterances.

3.2. Speaker indexing

In the first stage of the archiving process each utterance is indexed by an individual speaker. We have proposed an unsupervised approach based on *anchor models*[7]. Approximately 300 GMM anchor 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 clustering and the resulting classes are used to train speaker classification models. Finally, the models generated during clustering are used to perform speaker indexing. On the evaluation test-set, an average indexing accuracy of 97% was achieved using this approach.

3.3. Automatic speech recognition

The second stage involves generating a transcription of the input audio using speaker-adaptive speech recognition. Our Julius 3.4 recognition engine [8] is used to perform recognition, and sequential decoding is applied to handle long speech segments.

The recognition language model is constructed by linearly combining two independent language models: a "minutes" model trained from the minutes of the National Diet of Japan, and a "lecture" model based on the Corpus of Spontaneous Japanese (CSJ)[9]. These models cover different linguistic features common in discussions. The "minutes" model provides coverage over political and economic topic words, while the "lecture" model provides coverage of spontaneous speech phenomena, such as filled pauses and colloquial phrases.

An initial speaker-independent acoustic model is trained using the CSJ corpus. Speaker-adapted models are then generated using unsupervised MLLR adaptation based on the indices from speaker indexing. The resulting recognition system improved an average word accuracy from 51% to 57%.

Table 1. Proposed discourse tags

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

3.4. Discourse tagging

The third stage in the archival process involves discourse tagging. When accessing panel discussion archives, users typically require information on the opinions of the panelists in regard to various topics or statements. To allow quick and efficient access to this information discourse-based indices are required.

For this purpose, we introduce indices based on discourse tags. We analyzed typical panel discussions and observed that key-sentences exist in each turn and they will provide effective indexing. Based on this analysis a set of discourse tags were defined, as described in Table 1. Besides indexing purposes, these discourse tags may 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 handled differently. For the chairperson, rule based discourse tagging is applied. Rules for *Agenda*, *Question, Confirmation* and *Suggestion* discourse tags were manually defined from transcriptions of discussions.

For panelists' utterances, we set up *Question, Answer* and *Opinion* tags. *Question* and *Answer* tags are given by heuristic rules. *Opinion* tag is attached to key-sentences of panelists' utterances and tagged using the statistical based method as described in [10]. First, relevant discourse markers are selected for the current discussion. These discourse markers are selected using a $wf \cdot isf$ criteria. The relevance score (S_{w_i}) is calculated as the product of word frequency wf_i and inverse of sentence frequency sf_i, as shown in equation (1).

$$S_{w_i} = wf_i \cdot \log\left(\frac{N}{sf_i}\right) \quad (1)$$

Word frequency is defined occurrence of words in initial and final utterances in respective turns, because these utterances tend to be important. Sentence frequency is the occurrence counts of sentences that contain the word. Utterances are then tagged based on the occurrence of these discourse markers.

The effectiveness of the proposed discourse-based indexing method is investigated on the transcriptions of

Table 2. Recall and precision of opinion indices

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

the evaluation test-set. The recall and precision measures for various discourse types are shown in Table 2. High recall rates were achieved for all discourse types except *Suggestion*. These tagged key-sentences are useful for efficient browsing or retrieval.

3.5. MPEG-7 encoding

In the final stage, an archive is constructed for the current panel discussion incorporating the original audio, speaker indices, discussion transcriptions and discourse tags generated in the previous three stages. These are combined using an MPEG-7[6] framework to generate an XML format archive as shown in Figure 3.

As MPEG-7 is based on an XML framework, popular XML-based software including parsers, browsers, and editors can be used to access or edit this information data. The style-sheet framework (XSL) also allows the visual presentation of the archive to be altered simply.

4. Conclusions

One significant problem for archiving multimedia content such as public panel discussions is to extract effective indices that can be used for efficient information retrieval. In this work we introduce a novel indexing method based on discourse tagging. An appropriate set of role-dependent discourse tags was defined by analyzing the typical panel discussions. A discourse tagger was then constructed by rule based and statistical tagging techniques. Based on this tagging method, we develop a full automatic speech archiving system combining ASR, speaker indexing, discourse tagging within MPEG-7 framework.

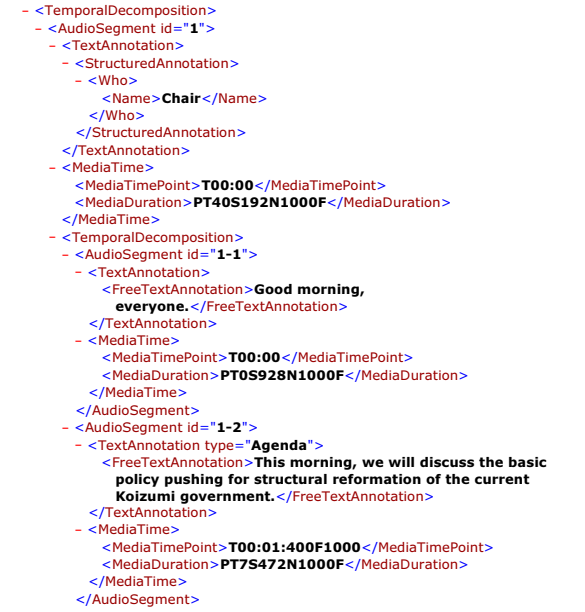

Figure 3. Sample annotation in MPEG-7 format

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] N. Yamamoto, J. Ogata, and Y. Ariki. Topic Segmentation and Retrieval System for Lecture Videos based on Spontaneous Speech Recognition. In *Proc. Eurospeech*, 2003.

[4] F. Metze, A. Waibel, M. Bett, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in Automatic Meeting Record Creation and Access. In *Proc. ICASSP*, 2001.

[5] M. Franz, B. Ramabhadran, T. Ward, and M. Picheny. Automated Transcription and Topic Segmentation of Large Spoken Archives. In *Proc. Eurospeech*, 2003.

[6] Moving picture experts group. http://www.cselt.it/mpeg/.

[7] Y. Akita and T. Kawahara. Unsupervised Speaker Indexing using Anchor Models and Automatic Transcription of Discussions. In *Proc. Eurospeech*, 2003.

[8] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, , and K. Shikano. Continuous Speech Recognition Consortium — An Open Repository for CSR Tools and Models —. In *Proc. IEEE Int'l Conf. on Language Resources and Evaluation*, 2002.

[9] 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*, 2000.

[10] T. Kawahara and M. Hasegawa. Automatic Indexing of Lecture Speech by Extracting Topic-independent Discourse Markers. In *Proc. ICASSP*, 2002.