Video Retrieval System Using Automatic Speech Recognition for the Japanese Diet

Mikitaka Masuyama[1], Tatsuya Kawahara[2], Kenjiro Matsuda[3]
National Graduate Institute for Policy Studies[1], Kyoto University[2], Kobe Shoin Women's University[3]
Minato-ku, Tokyo[1], Sakyo-ku, Kyoto[2], Nada-ku, Kobe[3], Japan
mmasuyama@grips.ac.jp[1], kawahara@i.kyoto-u.ac.jp[2], kenjiro@shoin.ac.jp[3]

Abstract

The Japanese House of Representatives, one of the two houses of the Diet, has adopted an Automatic Speech Recognition (ASR) system, which directly transcribes parliamentary speech with an accuracy of 95 percent. The ASR system also provides a timestamp for every word, which enables retrieval of the video segments of the Parliamentary meetings. The video retrieval system we have developed allows one to pinpoint and play the parliamentary video clips corresponding to the meeting minutes by keyword search. In this paper, we provide its overview and suggest various ways we can utilize the system. The system is currently extended to cover meetings of local governments, which will allow us to investigate dialectal linguistic variations.

Keywords: speech recognition, video retrieval, keyword search

1. Introduction

In recent times, there has been a surge in the development of analytical tools and techniques for analyzing the textual data of parliamentary proceedings. However, with the growing trend of parliamentary video streaming, there is a pressing need for similar tools to be developed for audio-visual data. While visual data offers a clear advantage over textual data for a more comprehensive analysis of parliamentary debates, it can be challenging to pinpoint the exact scene of a particular utterance by a specific speaker in lengthy video recordings that can span for hours.

To remedy this situation, we have launched an Internet video retrieval system for the Japanese Diet. Using the speech recognition system dedicated to Parliamentary speech which creates timestamp data to match parliamentary video feeds and the minutes of proceedings, it can pinpoint and play the parliamentary video clips corresponding to the minutes of proceedings through keyword search.

2. Video Retrieval System for Diet Deliberations

One of the authors has developed automatic speech recognition (ASR) technology, which the Japanese House of Representatives has deployed in the transcription system since 2011. The ASR system was trained with a large amount of parliamentary speech data, which covers terms and expressions used in the Parliament (Kawahara 2012, Kawahara 2017). It introduced an efficient lightly-supervised training based on statistical language model transformation, which fills the gap between faithful transcripts of spoken utterances and final texts for documentation. Once the mapping is trained, faithful transcripts for training acoustic and language models are no longer needed. The ASR system has consistently achieved character accuracy of over 90% since 2011, which helps streamline the transcription process. The accuracy rate currently has improved to 95 percent.

The Diet Library currently provides digitized minutes of parliamentary meetings via the Internet. Although these are not "official" records, they are amenable to keyword searching. On the other hand, we can watch the online live streaming at each house's secretariat website. We can also search the video library and watch videos of parliamentary meetings. The House of Representatives has made the parliamentary videos available since 2010, while the House of Councillors, the other house of the Diet, makes the videos available only one year after the meetings.

https://www.shugiintv.go.jp/index.php

https://www.webtv.sangiin.go.jp/webtv/index.php

Diet deliberation videos can be searched by meeting date, meeting title, subject, and speaker, although the English interface only offers the first two search options. Even if we successfully retrieve the desired deliberation video, we must watch the video from the beginning to the speech or debate segment we are interested in. It is not uncommon for a committee meeting to last more than 7 hours. While the video breakdown by questioner is available in the Japanese interface, video segmentation is usually 30 to 60 minutes long. No such breakdown is available in the English interface. Moreover, replies to parliamentary questions are arranged by the questioner, and we cannot search prime and cabinet ministers' deliberation videos answering parliamentary questions.

By linking the Diet Library's proceedings database and the Diet secretariats' deliberation video libraries, our "Video Retrieval System for Diet Deliberations (VRS)" makes it possible to retrieve the video clips corresponding to the minutes of the parliamentary meetings through keyword searching:

https://gclip1.grips.ac.jp/video/

With our system, we can directly retrieve the portion of the video feed we are interested in. We can instantly gain a visual understanding of the flow of parliamentary debate and check the facial expressions and body language of the speaker, all of which are not possible from a simple reading of the minutes of parliamentary meetings.

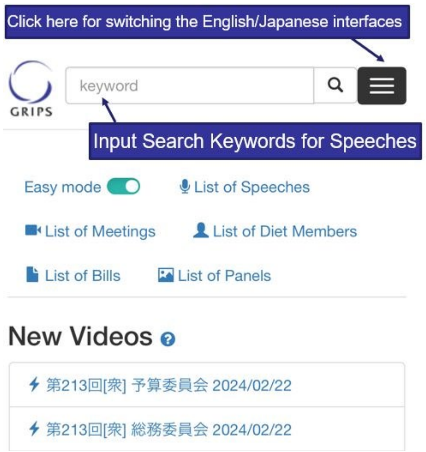
Our video retrieval system consists of two sub-systems. One uses the latest speech recognition techniques to create timestamp data to match the Diet Library's proceedings database and the Diet secretariats' deliberation video databases. The second sub-system uses the timestamp data to search the parliamentary minutes stored in our system and retrieve the Diet deliberation videos corresponding to the minute in question by keyword search. The results of keyword searches are deliberation video links, and the portion of the video we are interested in can be played partially by clicking the URL link for the deliberation video available in the Diet secretariats' databases (not stored in our system).

The system has been in operation and publicly available since November 2012. It is possible to keyword search all the plenary and committee meetings in the House of Representatives since January 2010 and those in the House of Councillors since December 2012[1].

Below, we briefly describe how our video retrieval system works. Figure 1 shows the top page of our web-based search interface, allowing us to search for deliberation video segments by typing keywords. The Japanese interface will appear when the user clicks "Japanese" in the upper right-hand corner.

One can type English keywords separated by spaces in the search field, and they will be translated automatically into Japanese and used in keyword searching. For instance, if one types "Kishida Fumio" (the name of the current Prime Minister of Japan) and "tax increase" in the search field and hits the search button, a list of the search results will appear in ascending order of date (Figure 2). As the default setting, our system searches the database for the past year, although it can be extended or shortened by calendar and filtered by other factors in the search results interface. Then, one can click one of the video links, and our system will instantly play the portion of the video corresponding to the speech, including the keywords (Figure 3).

The video-playing interface shows subtitles under the video and the speeches at the meeting on the right side, highlighting the current speech (not shown in Figure 3). By default, the video will play for one minute or three speeches. Alternatively, one can keep playing the video by clicking the play button in the toolbar under the video. Double-clicking any speech in the speech list allows one to instantly watch the video portion of the speeches before and after the speech found by keyword search. Once the user has moved on to another speech, the original speech found by keyword search remains highlighted in yellow.



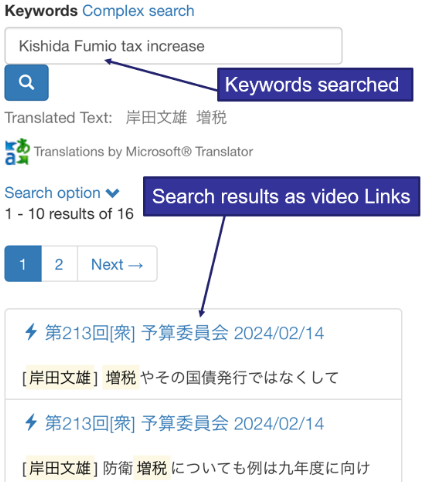
keyword

New Videos

第213回[衆] 予算委員会 2024/02/22

第213回[衆] 総務委員会 2024/02/22

Figure 1: Keyword Search Interface in English



Kishida Fumio tax increase

Translated Text: 岸田文雄 増税

Translations by Microsoft® Translator

1 - 10 results of 16

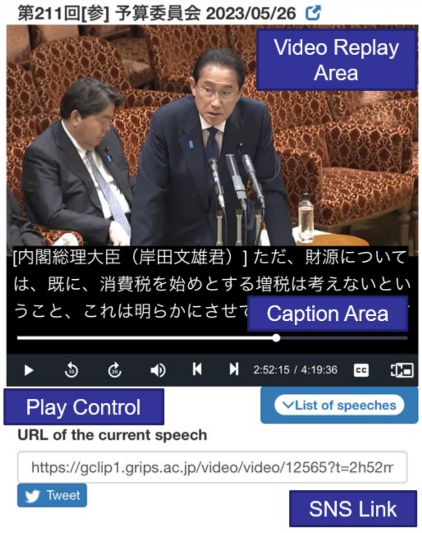
第213回[衆] 予算委員会 2024/02/14

[岸田文雄] 増税 やその国債発行ではなくして

第213回[衆] 予算委員会 2024/02/14

[岸田文雄] 防衛 増税 についても例は九年度に向け

Figure 2: Search Results Interface



第211回[参] 予算委員会 2023/05/26

[内閣総理大臣（岸田文雄君）] ただ、財源については、既に、消費税を始めとする増税は考えないということ、これは明らかにさせて

URL of the current speech

https://gclip1.grips.ac.jp/video/video/12565?t=2h52m

Figure 3: Video Replay Interface

[1] At the time of Feb. 22, 2024, our database covers the time period since Jan. 18, 2010, which includes a total of 14,156 hours, 12,685 meetings, 8,245,621 speeches, 13,282 speakers, and 554,634,003 Japanese characters.

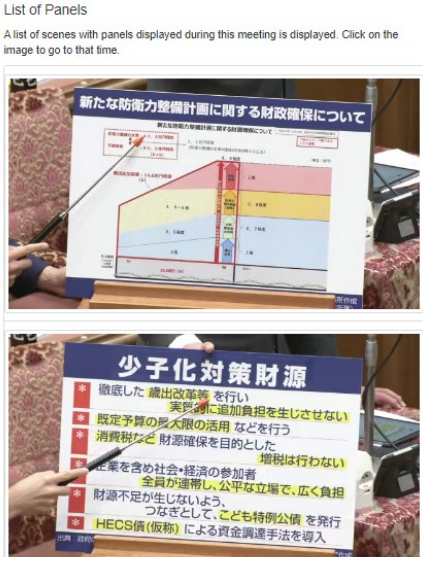
The video-playing interface shows subtitles under the video, highlighting the current speech. By default, the video will play for one minute or three utterances. Alternatively, one can keep playing the video by clicking the play button in the toolbar under the video. Double-clicking any speech in the speech list (not shown in Figure 3) allows one to instantly watch the video portion of the speeches before and after the speech found by keyword search. Once the user has moved on to another speech, the original speech found by keyword search remains highlighted in yellow.

Moreover, the video-playing interface shows the URL for the corresponding video portion, and one can easily share the URL via SNS by clicking the tweet button while the video stream is playing. The text of the speech and the URL will immediately appear in the tweet box. Moreover, the bottom of the page offers information about the speaker, followed by a list of agendas and the Diet members attending the meeting (not shown in Figure 3).

3. Usage beyond Keyword Search

We can utilize our video retrieval system in various ways. For instance, it allows us to obtain the URL for a moment of video streaming and to create and share a list of video links without downloading and editing the video files. Another way of utilizing the interfaces for keyword searching and partial replay is to post deliberation video links to internet news.

The minutes are essential for parliamentary discussion but do not tell the whole story. For instance, supplementary materials often used in committee meetings are graphic materials such as figures and tables, which concisely summarize the discussion points but are not usually included in the minutes. Thus, we combined speech and pattern recognition techniques to distinguish between the portions of videos that focus on the speaker and automatically extract video clips, including the moments focusing on supplementary materials used in committee meetings. Furthermore, we have developed an automatic text recognition system for these clips to extract and store text information in the database to be amenable to keyword search so that our system searches the video portion, focusing on the supplementary materials by keyword search through their content (Figure 4). The minutes are silent regarding non-verbal communication, and we are developing a web-based program to automatically extract and analyze the speaker's facial expressions and body language[2].



List of Panels

A list of scenes with panels displayed during this meeting is displayed. Click on the image to go to that time.

新たな防衛力整備計画に関する財政確保について
少子化対策財源
徹底した歳出改革等を行い
実質的に追加負担を生じさせない
既定予算の最大限の活用などを行う
消費税など財源確保を目的とした
増税は行わない
企業を含め社会・経済の参加者
全員が連帯し、公平な立場で、広く負担
財源不足が生じないよう、
つなぎとして、こども特例公債を発行
HECS債(仮称)による資金調達手法を導入

Figure 4: Supplementary Materials

The speech recognition output may contain irrelevant fillers and incorrectly recognized errors. The minutes become "official" by eliminating fillers, correcting inappropriate wording, and adding commas and periods. With our web-based program, we can systematically analyze the correspondence between the official minutes and speech recognition results. We can also check how we pronounce heteronyms, common in Japanese. While it is impossible to detect by reading the minutes, we can utilize our video retrieval system to analyze how parliamentary members pronounce heteronyms through keyword searching[3].

4. Conclusion

Our video retrieval system has great potential to boost the usage of parliamentary information. The speech recognition techniques for creating timestamp data for matching video and text information can apply to various meetings, including local assemblies, international conferences, and other less formal public and private meetings. For instance, many local assemblies in Japan increasingly use YouTube to disseminate deliberation videos. By extending the video retrieval system to such local assemblies, we can expect to improve speech recognition for dialectal diversity. Also, since some parliaments use multiple languages, we can develop multi-linguistic speech recognition by extending our system to such parliaments. Furthermore, international conferences like the United Nations Commission on Human Rights have stopped producing conference proceedings and recently disseminated meeting videos. A video retrieval system like ours may become the only way to search the content of such meetings.

[2] There are studies extracting emotions from minutes and videos (Rheault et al. 2016, Werlen et al. 2021, Rheault & Borwein 2019) and comparing verbal and non-verbal emotions (Werlen et al. 2018).

[3] Linguistic scholars focus on how politicians pronounce Iraq and figure out their diplomatic stance (Hall-Lew et al. 2010). Political scientists try to unravel politicians' gender differences in discussing women's issues by analyzing pitch (Dietrich 2019).

5. Bibliographical References

Kawahara. T. (2012). Transcription system using automatic speech recognition for the Japanese Parliament (Diet). In Proc. AAAI/IAAI, pp.2224—2228.

Kawahara T. (2017). Automatic meeting transcription system for the Japanese Parliament (Diet). In Proc. APSIPA ASC.

Dietrich et al. (2019). “Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech” *American Political Science Review* 113(4) 941-962.

Hall-Lew. (2010). “Indexing Political Persuasion: Variation in the Iraq Vowels” *American Speech*. 85: 91-102.

Rheault et al. (2016). “Measuring Emotion in Parliamentary Debates with Automated Textual Analysis” *PLOS ONE* 11(12) 1-18.

Rheault & Borwein. (2019). “Multimodal Techniques for the Study of Affect in Political Videos” *Prepared for the PolMeth Conference*, MIT, Cambridge, MA, July 18-20, 2019.

Werlen et al. (2018). “Is reading mirrored in the face? A comparison of linguistic parameters and emotional facial expressions” *CEUR-WS.org*, 1-2226, paper2.

Werlen et al. (2021). “Emotions in the parliament: Lexical emotion analysis of parliamentarian speech transcriptions” *CEUR-WS.org*, 1-2957, paper10.