New Transcription System using Automatic Speech Recognition (ASR) in the Japanese Parliament (Diet) -- The House of Representatives --

Tatsuya Kawahara
Professor of Kyoto University
Technical Consultant of the House

The Japanese Parliament (Diet) was founded in 1890. Since the very first session, verbatim records had been made by manual shorthand over a hundred years. However, early in this century, the government terminated recruiting stenographers, and investigated alternative methods including ASR technologies. The House of Representatives has chosen ASR for the new system. The system was deployed and tested in 2010, and it has been in official operation from April 2011.

The new system handles all plenary sessions and committee meetings. Speech is captured by the stand microphones in meeting rooms. Separate channels are used for interpellators and ministers. The speaker-independent ASR system generates an initial draft, which is corrected by reporters. Roughly speaking, the system's recognition error rate is around 10%, and disfluencies and colloquial expressions to be corrected also account for 10%. Thus, reporters still play an important role.

There are Japanese language-specific issues. First, we need to convert *kana* phonetic symbols to *kanji* or Chinese characters. This conversion often involves ambiguity because of many homonyms. Therefore, it is very hard to type in real time. Only limited stenographers using a special keyboard can perform. Moreover, there are differences between the spoken-style and the transcript style. So, we need to rephrase in many cases, but re-speaking or shadow speaking is not so simple.

Requirements for the ASR system are as follows. The first is high accuracy; over 90% is preferred. This can be easily achieved in plenary sessions, but is difficult in committee meetings, which are interactive, spontaneous, and often excited. The second requirement is fast turn-around. In the House, each reporter is assigned every 5-minute segment of a meeting session. ASR should be performed almost in real-time, so reporters can start working promptly even during the session. The third issue is compliance to the orthodox transcript guideline of the House. The electric dictionary of 60K lexical entries used in the system was proofed. In summary, the compliance issue is solved by hard work, fast turn-around is feasible by current computers, and high accuracy is technically most challenging.

The ASR system is realized by integrating the models or dictionaries developed by Kyoto University into the software engine of NTT Corporation, which made a successful bid for the entire system. The acoustic model stores sound patterns for phonemes, and the language model keeps frequent word sequence patterns. In order to achieve high performance, these models must be customized to Parliamentary speech. This means the system is independent of individual speakers, but dependent on Parliamentary speech. Therefore, they need to be trained with a large amount of speech and transcript data, which is called corpus.

As you know, there is a large amount of data of Parliamentary meetings. There is a huge archive of official meeting records in text; actually it amounts to 15M words per year, which is comparable to newspapers. There is also a huge archive of meeting speech, which amounts to 1200 hours per year. However, official meeting records are different from actual utterances due to the editing process by reporters. There are several reasons for this: differences between spoken-style and written-style, disfluency phenomena such as fillers and repairs, redundancy such as discourse markers, and grammatical corrections. In our analysis, Japanese has more disfluency and redundancy, but less grammatical corrections, because the Japanese language has a relatively free grammatical structure.

From these reasons, we need to build a corpus of Parliamentary meetings, which consists of faithful transcripts of utterances including fillers, aligned with official records. We prepared this kind of corpus in the size of 200 hours in speech or 2.4M words in text. The corpus is vital for satisfactory performance, but very costly. Moreover, it needs to be updated; otherwise, the performance would degrade in time.

In order to exploit the huge archive of Parliamentary meetings in a more efficient manner, we have investigated the differences between the official meeting record and the faithful transcript that is the target of ASR. Although there are differences by 13% in words, 93% of them are simple edits such as deletion of fillers and correction of a word. These can be computationally modeled by a statistical framework.

This leads to an innovative approach for semi-automated corpus generation and ASR model training. With the statistical model of the difference, we can predict what is uttered from the official records. By counting the possible word sequences, we can make the language model. Moreover, by referring to the audio data of each utterance, we can reconstruct what was actually uttered. By memorizing the sound pattern for each phoneme, we can make the acoustic model. As a result, we can build precise models of spontaneous speech in Parliament, and this model will evolve in time, reflecting the change of MPs and topics discussed.

Evaluations of the ASR system have been conducted since the system was deployed in last year. The accuracy defined by the character correctness compared against the official record is 89.4% for 108 meetings done in 2010 and 2011. When limited to plenary sessions, it is over 95%. No meetings got accuracy of less than 85%. The processing speed is 0.5 in real-time factor, which means it takes about 2.5 minutes for a 5-miniute segment. The system can also automatically annotate and remove fillers, but automation of other edits is difficult.

The post-editor used by reporters is vital for efficient correction of ASR errors and cleaning transcripts. We adopted a screen editor, which is similar to the word-processor interface, not a line editor, so that reporters can concentrate on making correct sentences. Note that it was designed by reporters, not by engineers. The editor provides easy reference to original speech and video, by time, by utterance, and by character. It can speed up and down the replay of speech.

There are several issues in system operation and reliability. First, we adopted the dual system configuration to backup for possible system troubles. Second, portable IC recorders are used in each meeting room for another backup. Except for plenary sessions and budget committee meetings, reporters do not attend the session, but a staff (=logger) attends to monitor what is going on.

A side effect of the ASR-based system is all of text, speech, and video are digitized and aligned (hyperlinked) by speakers and by utterance. This feature provides a good platform for reporters even if ASR result is not usable. It also allows for efficient search and retrieval of the multi-media archive.

For system maintenance, we continuously monitor the ASR accuracy, and update ASR models. Specifically, the lexicon and language model are updated once a year to incorporate new words and topics. Note that new words can be added temporarily at any time. The acoustic model will be updated after the change of the Cabinet or MPs, which takes place after the general election.

In summary, we believe we have realized one of the highest-standard ASR systems dedicated to Parliamentary meetings. Specifically, the accuracy is 89% in character correctness. We expect the system will improve or evolve with more data accumulated. There was a drastic change from the manual short-hand to this fully ICT-based system. Thus, it will need some time for reporters to be accustomed and we also need to develop a new training methodology. But the most important point is reporters still play a central role in the new system.