Speech Recognition with a Seamlessly Updated Language Model for Real-Time Closed-Captioning

Toru Imai, Shinichi Homma, Akio Kobayashi, Takahiro Oku, and Shoei Sato

NHK (Japan Broadcasting Corporation)
Science & Technology Research Laboratories, Tokyo, Japan
{imai.t-mq, homma.s-fc, kobayashi.a-fs, oku.t-le, satou.s-gu}@nhk.or.jp

Abstract

It is desirable to consistently and seamlessly update a language model of speech recognition without stopping it for online applications such as real-time closed-captioning. This paper proposes a novel speech recognition system that enables the model to be updated at any time even while it is running. It can run the second decoder with the latest model in parallel, and their priority that must be accessed is controlled at a non-speech portion by an additional job process, which sends acoustic features only to an active target decoder with the latest model and sends recognized words to the backend manual error correction for closed-captioning. The system seamlessly updates the model and ensures endless speech recognition with the latest model at any time. Our new practical real-time closed-captioning system reduced word errors by two thirds with the proposed language model update mechanism in the speech recognition and captioning experiments for Japanese broadcast news programs.

Index Terms: speech recognition, closed-captioning, language model, model updates

1. Introduction

Simultaneous captioning of live broadcast programs is of great value to the hearing impaired and elderly. All non-live TV programs of NHK (Nippon Hoso Kyokai; Japan Broadcasting Corp.) General TV shows are already closed-captioned, but only 20% of all live broadcasts are. Although Japanese stenographic keyboards can be used for real-time captioning, they require several highly skilled operators working at the same time to deal with the great number of homonyms in Japanese.

To provide text from speech more efficiently, NHK has extensively researched automatic speech recognition aimed at providing closed-captioned live TV programs [1].

NHK started to operate a speech recognition system with an internally developed recognition engine and a manual error correction system for closed-captioning broadcast news in 2000 [2]. However, because of the difficulties of speech recognition at that time, captions of this sort were limited to program parts where an anchorperson read manuscripts, which were revised from the original electronic news manuscripts. Later on, other portions such as field reports and interviews were manually captioned by using stenographic keyboards. For practical reasons, these keyboards have been used to closed-caption the entire news program these past few years, but we have been developing a more accurate and efficient speech recognition system than the original one, as described in this paper.

For captioning other live programs, such as sports programs, with a sufficient degree of accuracy, we use the "re-speak" method, in which another speaker listening to the original speech of the programs rephrases the commentary so that it can be recognized [3][4]. The re-speaker works in a quiet studio, not the arena or stadium from where the broadcast originates. This method not only improves recognition accuracy, but also makes captions easier to read since it allows summarizing and paraphrasing. A speech recognition system with the re-speak method has been practically used since 2001 in NHK’s live sports and information programs.

To more efficiently expand the range of closed-captioned programs, especially for news programs, we have developed a hybrid speech recognition system [5] that switches input speech between that of the original program and a re-speak network in a decoder without stopping it. However, we would like to also change the vocabulary and its n-gram probabilities, and it is not easy to dynamically change the whole language model and the dictionary while captioning. Thus, this paper proposes a novel speech recognition mechanism that enables the model to be updated any time even while it is running. The system can run the second decoder with the latest model in parallel, and their priority that must be accessed is controlled at a non-speech portion by an additional job process. It seamlessly updates the model and ensures endless speech recognition with the latest model at any time. Since it activates a set of another new speech recognizer, it enables not only a language model to change into a totally different domain of a successive TV program, but also acoustic models, search conditions, and a decoder to change into different types.

Section 2 describes our previous hybrid closed-captioning system that had only one decoder to be restarted. Section 3 details the new closed-captioning system that has multiple decoders to enable seamless language model updates and also describes speech recognition and closed-captioning experiments for Japanese broadcast news.
2. Closed-captioning system

Our closed-captioning system consists of an analog to digital (A/D) converter, which inputs target speech, and speech recognition with an acoustic analysis and a decoder using an acoustic model, a language model, and a dictionary, followed by an operator manually confirming and correcting recognition results to produce texts for live closed-captions (Figure 1). It should work in real-time online with the manual error correction, have fewer miss-recognized words, and produce accurate closed-captions with a very low latency in just a few seconds before they are shown on the TV screen for the hearing impaired.

2.1. Hybrid of direct and re-speak methods

Our previous closed-captioning system works with speech recognition in a hybrid way, combining the "direct method", which directly recognizes the program sound, and the "re-speak method", which recognizes rephrased utterances by a different re-speaker. In news programs, the direct method is used for not only speech read by an anchorperson in a studio, but also field reports by a reporter, and the re-speak method is used for other parts, such as conversations and interviews. The program sound and the re-phrased utterances come in two separate input channels of the A/D converter. They can be switched exclusively with a space bar of a keyboard of the speech recognition PC by the re-speaker or the correction operator in accordance with each news item in a program. As the switching is done manually while listening to the sound, a speech buffer of about one second is used to avoid losing any speech beginnings of the direct program sound for recognition.

2.2. Online speech detection

The selected channel of the program sound or the re-phrased speech is A/D-converted in 16 kHz and acoustically analyzed into 39-dimensional parameters (12 Mel frequency cepstral coefficients with the log-energy and their first- and second-order regression coefficients) every 10 ms with a 25-ms-wide Hamming window. To obtain only speech segments online with a low latency, we use a dual-gender phoneme recognizer [8], where male and female monophone acoustic models of hidden Markov models (HMMs) endlessly run in parallel. Utilizing the cumulative phoneme likelihood compared with the cumulative non-speech likelihood, the dual-gender phoneme recognizer detects the start-point and the end-point in the speech segment with only a small delay from the audio input.

2.3. Dual-gender speech recognition

As soon as the start-point of an utterance is detected, the subsequent continuous speech recognizer (decoder) starts decoding the speech using gender-dependent triphone acoustic models for the male and the female in parallel in a single lexical phonetic network [8]. The HMMs, which were trained with NHK’s news database of hundreds of hours, search for the best parallel word sequence, but with the common beam threshold. Then the word hypotheses for only the matched gender naturally remains, and the other one stops the search in accordance with their score difference. Conversations between male and female speakers sometimes give a gender-mixed utterance due to the difficulty in chopping it up in accordance with a gender-change without any pauses in between. The decoder utilizes the gender-change information on the basis of the gender attributes identified in the simultaneously running speech detection. This enables the search to restart for the deactivated gender’s acoustic models.

The decoding process composes a word lattice along with the search as the first pass with a bigram language model trained from Japanese broadcast transcriptions. To make a quick decision on speech recognition results, the partial word lattice is rescored with a trigram language model as the second pass, without waiting for the end-point to be detected [9].

2.4. Language models

A general language model trained with NHK’s 17-year long news manuscripts is adapted with a much higher weight [10] given to the latest news manuscripts. It is carried out just once ten minutes before the show starts in practical use so as not to break the captioning service. Although NHK’s anchorpersons and reporters usually revise the original electronic manuscripts with printed paper and a pen, such adaptation is effective to reduce the perplexity and an out of vocabulary rate. However, the adaptation preceding the show cannot cover delayed or new manuscripts, which may cause lack of vocabulary or word n-gram statistics and produce recognition errors.

2.5. Manual error correction

The closed-captioning system uses manual confirmation and correction with a touch panel and an ordinary keyboard for any speech recognition errors. It requires only one or two flexible correction operators depending on the TV programs and the difficulties of the speech recognition [5]. The number of correction operators was reduced from four needed in our first operated system for news programs (two sets of an error pointer and an error corrector) [2] due to recent improvements of our previous speech recognition system. Therefore, we expect this more efficient system with fewer operators will help to enable expansion of closed-captioned program coverage, especially not only for national but also local news programs produced at different small local TV stations.

![Figure 1: Previous hybrid closed-captioning system with only one decoder and a static language model.](image-url)
3. Updating a model while running

3.1. Why online model updates needed

In broadcasting stations with some TV or radio channels, reporters write electronic manuscripts and upload them to news archives asynchronously with broadcast news programs. Directors in charge of a specific news program make a news order, select some of the manuscripts, and revise them on printed paper with a red pen by hand in accordance with the assigned length to each news item at the last minute or even in the middle of the live program. Sometimes the manuscripts are not ready to be uploaded for some news items in the beginning of the news show, and news may break in the middle.

To adapt the language model to the latest news, the previous closed-captioning system with only one decoder used the electronic manuscripts available at ten minutes before the show opening, because it initially took about eight minutes to complete the model update. Besides, the decoder could not be stopped and restarted for the new language model once the show opening, because it initially took about eight minutes to complete the model update. Therefore, manuscripts of any delayed or breaking news uploaded after the language model training could never be used in the show despite their effectiveness to reduce recognition errors especially for new proper nouns and topics.

A cache language model [6] was developed for dynamic updates during recognition, but such self-training without any additional data cannot cover new words or topics. It may be possible to add only new words to a lexical phonetic network in a decoder with a fixed probability of unknown words [7] without stopping the running decoder. However, the whole vocabulary and its n-gram probabilities must be replaced for more accurate and updated recognition, and it is not easy to dynamically update the model while it is running.

3.2. New closed-captioning system with two decoders

We have changed the closed-captioning system so that the dictionary can be consistently and seamlessly adapted to the latest manuscripts and loaded at any time without stopping closed-captioning even while the news show is on the air. Figure 2 shows the block diagram of the proposed new closed-captioning system. The system can run the second decoder with the latest language model in parallel, and their priority that must be accessed is controlled by an additional job process, which sends acoustic features only to an alive decoder with the latest model, receives recognized words, and sends them to the backend manual error correction for closed-captioning. The switching to the newly started decoder is carried out only at a non-speech portion to avoid breaking the running speech recognition in the middle. This seamlessly updates models and ensures endless speech recognition with the latest models at any time.

The detailed procedure of the seamless model update, which is controlled by the controller job process, is as follows.

1) Once a new electronic manuscript written by a reporter is uploaded in the news manuscript archives, a previous language model is automatically adapted to it with a much higher weight [10]. Its vocabulary (60k in our case) is also updated on the basis of word frequencies, and the pronunciations for the newly added words in the dictionary are automatically estimated using our much larger dictionary on the basis of a morphological analyzer and some web resources.

2) Notified that the new adapted language model is ready, the controller job process lets the non-active second decoder run with the latest model and the dictionary in parallel.

3) If the new decoder is ready to recognize speech, the priority that must be accessed is switched to the new one only at a non-speech portion. Then the older decoder is deactivated.

4) The above process is repeated every time a new manuscript is uploaded even during closed-captioning.

Figure 3 illustrates all timings of the model and decoder update process. Suppose only decoder-A is active in the beginning of the speech recognition and a new manuscript is uploaded at time T1. The language model adaptation takes only two and a half minutes in the new system till time T2.
Then decoder-B is activated, and its initialization reading the new adapted language model takes about 20 seconds to be ready for speech recognition at time T3. However, in this example, time T3 is in a speech portion of Speech 4, which is being recognized by decoder-A. Then at time T4 in a non-speech portion, the priority that must be accessed is switched to the new decoder-B, which recognizes Speech 5 and later speeches until decoder-A becomes active next time. All these timings are controlled by the controller job process.

### 3.3. Experiments

Through speech recognition and captioning experiments for Japanese broadcast news, we examined the effectiveness of the proposed language model updates. We set up two kinds of Japanese broadcast news, we examined the effectiveness of the speech recognition and captioning experiments for Japanese broadcast news. The general-LM was trained with NHK’s 17-year long news manuscripts, which did not include news manuscripts directly related to news programs with higher weights to the word frequencies. The vocabulary size of the both LMs was 60K. The general-LM simulates the case where no directly related news manuscripts are used for training because of delayed uploads of manuscripts or breaking news. The updated-LM simulates the proposed updated case where news manuscripts directly related to all the news episodes in evaluation data are used for adaptation, although they do not perfectly match the actual spoken words. Since how many directly related news manuscripts are ready ten minutes before actual news programs is on a case-by-case basis, these experiments simulate the worst case and the best case in terms of the adaptation data. The evaluation data were NHK’s two Japanese news programs presented by a male anchorperson: news-A was 9 episodes of a 12-min.-long show of national news and local news and was 1,765 words in total; news-B was 5 episodes of a 5-min.-long show of national news, and was 781 words. Table 1 gives the perplexity rate, the out of vocabulary (OOV) rate, and the trigram hit rate for each language model.

<table>
<thead>
<tr>
<th>News programs</th>
<th>General-LM</th>
<th>Updated-LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perplexity, an OOV rate, and a trigram hit rate</td>
<td>Perplexity, an OOV rate, and a trigram hit rate</td>
</tr>
<tr>
<td>News A</td>
<td>Perplexity =17.8</td>
<td>Perplexity =12.0</td>
</tr>
<tr>
<td></td>
<td>OOV rate =0.74%</td>
<td>OOV rate =0.16%</td>
</tr>
<tr>
<td></td>
<td>Trigram hit rate =86.4%</td>
<td>Trigram hit rate =91.5%</td>
</tr>
<tr>
<td>News B</td>
<td>Perplexity =19.2</td>
<td>Perplexity =12.3</td>
</tr>
<tr>
<td></td>
<td>OOV rate =0.81%</td>
<td>OOV rate =0.00%</td>
</tr>
<tr>
<td></td>
<td>Trigram hit rate =84.2%</td>
<td>Trigram hit rate =89.3%</td>
</tr>
</tbody>
</table>

In the results, the word error rates of the speech recognition for the longer news-A were 3.1% for the general-LM and 1.2% for the updated-LM. As for the shorter news-B, they were 2.9% for the general-LM and 0.7% for the updated-LM. On average, the updated-LM could significantly reduce the word error of 3.0% by two-thirds to 1.0%. The updated-LM largely reduced the perplexity and the OOV rate and increased the trigram hit rates. In these experiments, one operator manually corrected errors, and all the miss-recognized words could be corrected when the updated-LM was used. With the general-LM, two words in news-A remained wrong even after the manual correction, although they were very trivial function words and so did not change the meaning of the news contents. The general-LM produced three times more word errors than the updated-LM, which caused the correction operator to work harder and captions to take longer to appear on the TV screen. We may, therefore, conclude that the speech recognition with the consistently and seamlessly updated LM is very effective and useful in the real-time closed-captioning application.

### 4. Conclusions

This paper proposed a novel speech recognition mechanism that enables the model to be updated at any time even while it is running. The system can run the second decoder with the latest model in parallel, and their priority that must be accessed is controlled at a non-speech portion by an additional job process. It seamlessly updates models and ensures endless speech recognition with the latest model at any time for more accurate speech recognition. Our new real-time closed-captioning system reduced word errors by two-thirds with the proposed language model update mechanism in the speech recognition and captioning experiments for Japanese broadcast news. We are aiming to soon release this closed-captioning system for news programs and will further improve its performance for more captioned TV programs through research on spontaneous and conversational speech recognition.

### 5. References


