

New Real-Time Closed-Captioning System for Japanese Broadcast News Programs

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

NHK (Nippon Hoso Kyokai; Japan Broadcasting Corp.)
Science and Technical Research Laboratories,
1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510, Japan
{homma.s-fc, kobayashi.a-fs,
oku.t-le, satou.s-gu, imai.t-mq,
takagi.t-fo}@nhk.or.jp
<http://www.nhk.or.jp/str1>

Abstract. A new real-time closed-captioning system for Japanese broadcast news programs is described. The system is based on a hybrid automatic speech recognition system that switches input speech between the original program sound and the rephrased speech by a "re-speaker". It minimises the number of correction operators, generally to one or two, depending on the difficulties of the speech recognition, although four correction operators were needed in our previous news system. Experiments show that the system could be used for captioning typical news programs at local stations, which have fewer staff and where simple operation is required.

Key words: real-time closed captioning, speech recognition

1 Introduction

There is a great need for more TV programs to be closed captioned to enable hearing impaired and elderly people to watch TV. In 2007, the Japanese Ministry of Internal Affairs and Communications published an administrative guideline to promote closed-captioned TV programs; it requires Japanese broadcasters to caption 100% of TV programs, including live ones—except for a few special programs—by 2017. All non-live TV programs of NHK General TV are already closed captioned, but when live broadcasts were included, its closed-captioned percentage was only 43.1% in 2006. We strongly need to expand the number of closed-captioned live broadcasts.

Although Japanese stenographic keyboards can be used for real-time captioning, they require six highly skilled operators working at the same time to deal with the great number of homonyms present in ideograms (Kanji). To enable speech to be transcribed more efficiently, NHK has done extensive research on automatic speech recognition (ASR) aimed at providing closed-captioned TV

programs in real time. NHK started to operate an ASR system with a manual error correction system for closed-captioning broadcast news in 2000 [1]. However, because of the difficulties of speech recognition, captions of this sort were limited to program sections where an anchorperson read manuscripts. Moreover, the running costs were relatively high because the system always required four correction operators.

NHK started to operate another ASR system for closed-captioned TV programs besides news in 2001 [2]. The system is based on the "re-speak" method, where another speaker listening to the original speech of the programs rephrases the commentary so that it can be recognised for captioning. The method is mainly used for sports programs because it helps prevent a noisy environment and spontaneous or emotional commentary. The BBC also uses the re-speak method, not only for sports programs but also news programs [3]. However, this method is less efficient than direct recognition of program sound.

To expand the range of closed-captioned programs, we have developed a new hybrid ASR system that switches input speech between the original program sound and rephrased speech and that requires fewer correction operators. The overview and experiments using the system are described next.

2 System Overview

The new hybrid ASR system consists of a speech recogniser and an error correction system, as illustrated in Fig. 1. The system will be located in a broadcasting station and generate closed-caption data, which will be combined with video signals and transmitted to viewers. Because closed captioning is standard in the Japanese digital TV system, which becomes widespread completely by 2011, all hearing impaired and elderly people will be able to watch TV programs with closed captions in Japan.

NHK has an internally developed speech recogniser that runs on a Linux or Windows PC. We focused on developing our recogniser and applying it to real-time closed-captioned TV programs. We developed the recogniser with a method of progressively outputting the latest available words because the system needs not only high accuracy but also low latency from the speech input to the text output [4]. The method enabled outputting words continuously with very small delays and with a negligible increase in the number of word errors.

A speech recogniser typically consists of an acoustic model, a language model, a dictionary, and a recognition engine. Our recogniser uses two acoustic models. They are speaker independent but gender dependent, and the recogniser automatically detects the gender of the speaker, allowing for the use of more accurate gender-dependent acoustic models [5]. The language model of our recogniser is domain specific and is adaptable to the latest news or training texts [6]. At NHK, manuscripts¹ for news programs written by reporters are uploaded to an online system named the "News Information System". Just before broadcasting

¹ The manuscripts do not always correspond to the speech read in a news program.

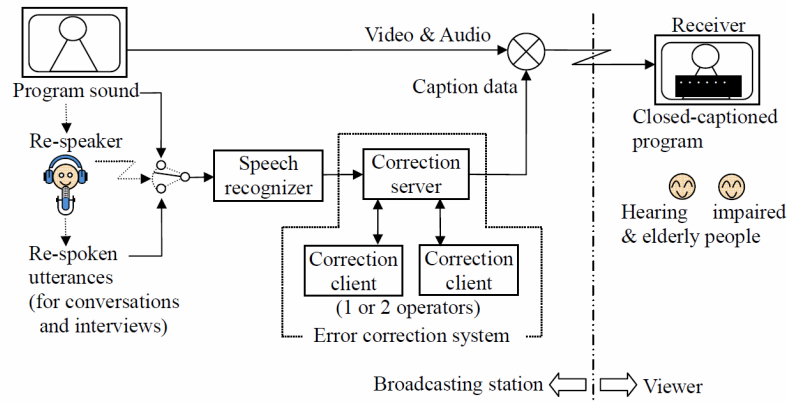


Fig. 1. New hybrid ASR system.

a news program, our ASR system makes a request to the online system and downloads the manuscripts written for the news program, and it automatically constructs the language model adapted to the program. The vocabulary size of the dictionary is 60K words, and the pronunciations of new words are generated automatically in the manuscripts. An operator checks the pronunciations, and if pronunciation errors are detected, the errors are corrected manually.

Our latest speech recogniser for news programs can directly recognise not only speech read by an anchorperson in a studio, but also field reports by a journalist with a word accuracy of more than 95%. However, we use the re-speak method for other parts of news programs because the recogniser cannot adequately decipher spontaneous speech such as conversations and interviews with a sufficient degree of accuracy. This method involves having another speaker rephrase the content after switching the input speech to his or her own voice. We call this speaker a "re-speaker", and the switching operation is also his or her assignment. The new system uses manual correction that minimises the number of correction operators, generally to one or two, depending on the difficulties of the speech recognition. However, four correction operators were needed in the previous news system (two sets for an error finder and an error corrector). The modified system allows closed captioning of an entire news program and fewer correction operators than before.

The error correction system consists of a server and one or two correction clients with touch-panel screens and keyboards. Speech recognition results are displayed on the touch screens progressively, and immediately after each operator detects speech recognition errors, he or she touches the errors on the screen and manually corrects them by the keyboard. While an operator requires many months to acquire special skills for using a Japanese stenographic keyboard, our system only requires about 1 week of training for a re-speaker and about 1 month of training for a correction operator.

3 Experiments

We conducted experiments on nationwide and local regular short news programs with one anchorperson. The new system achieved a caption accuracy of 99.9% without any fatal errors when two correction operators were used. When only one correction operator was used, the accuracy was 99.8%. The average delay of captioning was the same as that achieved by six people operating stenographic keyboards. We found that the new system will be especially suitable for local stations, which have fewer staff members and where easy operation is required. Also, the style of presenting the news at these stations tends to be comparatively simpler, with only one anchorperson reporting.

However, our system is not yet good enough for large-scale news shows with spontaneous and conversational speaking styles performed by a couple of anchorpersons, reporters, and guests. We intend to improve the speech recognition accuracy for such speaking styles in the future.

4 Conclusion

We described our new system, which is based on a hybrid method of switching between a direct program sound and a re-speaker's voice for simple news programs. The new system achieved very high caption accuracy without long delays during our experiment using the news programs with one anchorperson, and we found that the new system is especially suitable for local stations. We are now testing the system for practical use. To expand the closed-captioned coverage of live programs efficiently, we intend to further refine the speech recognition system so that it can caption a wide variety of live programs in the future.

References

1. Ando, A., Imai, T., Kobayashi, A., Isono, H.: Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Programs. *IEEE Trans. on Broadcasting*, 46(3), 189–196 (2000)
2. Imai, T., Matsui, A., Homma, S., Kobayakawa, K., Onoe, S., Sato, S., Ando, A.: Speech Recognition with a Re-Speak Method for Subtitling Live Broadcasts. *International Conference on Spoken Language Processing*, pp. 1757–1760, Denver (2002)
3. Marks, M.: A distributed live subtitling system. *BBC R&D White Paper*, WHP070 (2003)
4. Imai, T., Kobayashi, A., Sato, S., Tanaka, H., Ando, A.: Progressive 2-Pass Decoder for Real-Time Broadcast News Captioning. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1937–1940, Istanbul (2000)
5. Imai, T., Sato, S., Kobayashi, A., Onoe, K., Homma, S.: Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News. *International Conference on Spoken Language Processing*, pp. 1–4, Pittsburgh (2006)
6. Kobayashi, A., Onoe, K., Imai, T., Ando, A.: Time Dependent Language Model for Broadcast News Transcription and Its Post-Correction. *International Conference on Spoken Language Processing*, pp. 2435–2438, Sydney (1998)