



REAL-TIME CLOSED-CAPTIONING USING SPEECH RECOGNITION

**TORU IMAI, SHINICHI HOMMA, AKIO KOBAYASHI, SHOEI SATO, TOHRU TAKAGI,
KYOUICHI SAITOU, AND SATOSHI HARA**

NHK (NIPPON HOSO KYOKAI; JAPAN BROADCASTING CORP.)

Abstract

There is a great need for more TV programs to be closed-captioned to help hearing impaired and elderly people watch TV. For that purpose, automatic speech recognition is expected to contribute to providing text from speech in real-time. NHK has been using speech recognition for closed-captioning of some of its news, sports and other live TV programs. In news programs, automatic speech recognition applied to anchorpersons' speech in a studio has been used with a manual error correction system from 2000 to 2006. Live TV programs, such as music shows, baseball games, and the Olympic Games, have been closed-captioned since 2001 by using a re-speak method in which another speaker listens to the program contents and rephrases them for speech recognition. To efficiently expand closed-captioning, a new hybrid speech recognition system that switches input speech between the original program sound and the rephrased speech with fewer correction operators is under study.

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 when live broadcasts are included only 43.1% of them were closed-captioned in 2006 [1]. 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 in Japanese. To provide text from speech more efficiently, NHK has done extensive research on automatic speech

recognition aimed at providing closed-captioned live TV programs in real-time.

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 March 2000 [2]. However, because of the difficulties of speech recognition, captions of this sort were limited to program parts where an anchorperson read manuscripts, which were revised from original electronic news scripts. Later on, other portions such as field reports and interviews have been manually captioned by using stenographic keyboards. Since 2006, these keyboards have been applied to the entire news program for economic reasons.

Captioning of other live programs, such as sports programs, in addition to news programs, would also benefit our viewers. However, current speech recognition technology cannot adequately recognize spontaneous and emotional commentary in such a program with a sufficient degree of accuracy. Therefore, we use the "re-speak" method, where another speaker listening to the original speech of the programs rephrases the commentary so that it can be recognized for captioning [3][4]. This speaker works in a quiet studio, not in the field, stadium, or hall 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 used since 2001 in live programs, such as music shows, baseball games, the Grand Sumo Tournaments, the Olympic Games, and World Cup Football Games.

To expand the range of closed-captioned programs efficiently, we are developing a new hybrid speech recognition system that will

switch input speech between the original program sound and the rephrased speech with fewer correction operators. Our latest speech recognizer for news programs can directly recognize not only speech read by an anchorperson in a studio, but also field reports by a reporter with sufficient word accuracy of more than 95% [5]. Other parts of news programs, such as conversations and interviews, can be captioned with the re-speak method where another speaker rephrases the contents after switching the input speech to his or her voice. This allows closed-captioning of an entire news program using only the automatic speech recognition and fewer correction operators than before. One of our research goals is to enable closed-captioning of nationwide regular short news programs and local news programs at an acceptable operation cost [6].

We describe our automatic speech recognizer in Section 2, the current captioning system with the re-speak method in Section 3, and the hybrid system now being developed in Section 4.

2. Automatic Speech Recognizer

Automatic speech recognition is a technique to obtain text from speech by using a computer. Speech recognition has greatly advanced over the last few decades along with progress made in statistical methods and computers. Large-vocabulary continuous speech recognition can now be found in several applications, though it does not work as well as human perception and its target domain in each application is still limited. We have focused on developing a better speech recognizer and applying it to closed-captioned TV programs.

A speech recognizer typically consists of an acoustic model, a language model, a dictionary, and a recognition engine (Fig. 1). The acoustic model statistically represents

the characteristics of human voices; i.e., the spectra and lengths of vowels and consonants. It is trained beforehand with a speech database recorded from NHK broadcasts. The language model statistically represents the frequencies of words and phrases used in the individual target domain; e.g., news, baseball or soccer. It is also trained beforehand with a text database collected from manuscripts and transcriptions of previous broadcasts. The dictionary provides phonetic pronunciation of the words in the language model. As the recognition engine searches for the word sequence that most closely matches the input speech based on the models and the dictionary, it cannot recognize words not included in them. Training databases are therefore very important to obtain satisfactory speech recognizer performance.

Notable features of our speech recognizer are the speaker-independent acoustic model, the domain-specific language model which is adaptable to the latest news or training texts, and the very low latency from the speech input to the text output, which makes this recognizer suitable for real-time closed-captioning [7].

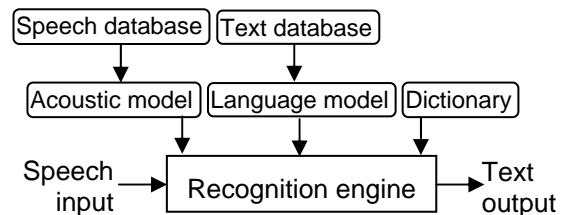


Fig. 1 Automatic speech recognizer.

3. Re-Speak Method

The commentaries and conversations in live TV programs such as sports are usually spontaneous and emotional, and a number of speakers sometimes speak at the same time. If such utterances are directly fed into a

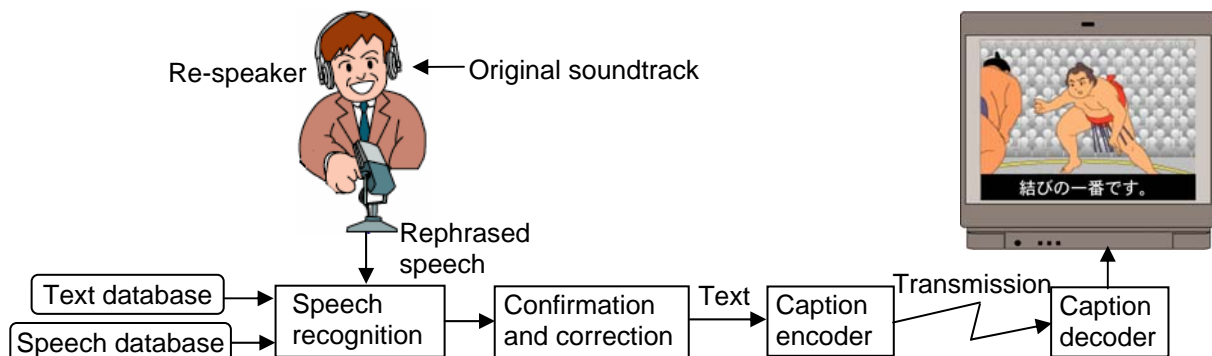


Fig. 2 Closed-captioning system with a re-speak method.

speech recognizer, its output will not be accurate enough for captioning because of background noise, unspecified speakers, or speaking styles that do not match acoustic models and language models. It is difficult to collect enough training data (audio and text) in the same domain as the target program. Therefore, we employ the re-speak method to eliminate such problems.

In the re-speak method, a different speaker from the original speakers of the target program carefully rephrases what he or she hears [3][4]. We call this person the re-speaker. The re-speaker listens to the original soundtrack of live TV programs through headphones, and repeats the contents, rephrasing if necessary, so that its meaning will be clearer or more acceptable than the original and the expression will be more easily recognized (Figs. 2 and 3). This method provides several advantages for speech recognition.

3.1. Advantages

Re-spoken utterances have no background noise. As only one re-speaker rephrases the speech of all the speakers in a program, the speech does not overlap. The re-speaker is known in advance, and acoustic models of the speech recognizer can be adapted prior to the program with a relatively large amount of adaptation data. The re-speaker speaks clearly and calmly, rather than emotionally, without repeating filled pauses and hesitations in the original sounds. If a recognition error occurs, the re-speaker repeats the same phrase or tries a different phrase. The re-speaker can also supplement the speech by mentioning audience sounds, such as applause, even if no mention is made in the original narration. These advantages improve the recognition accuracy and make closed-captions easier to understand for hearing impaired viewers.

This method enables summarization or rephrasing of the original narrations. Conversational speech is rephrased into a planned speech style. The mismatch between the language model of the speech recognizer and the speech is reduced, and this makes the closed-captions more accurate and more understandable.

Since the quality of re-speaking affects the speech recognition performance, though, skillful re-speakers are needed to ensure the final captions are as good as possible.

3.2. Operation

Since December 2001, NHK has been using the re-speak method for automatic speech recognition and closed-captioning of sports events and other live shows (Fig. 3). For example, this method of captioning was used in NHK's coverage of the Olympic Games, the World Cup Football Games, the Grand Sumo Tournaments, and Japanese professional baseball games. For Major League Baseball games, a commentary directly from NHK's broadcasting studio is recognized, instead of using a re-speaker, because it includes no background noise. The language models are adapted to each program and the acoustic models are adapted to each re-speaker. The recognition accuracy is approximately 95% [4], and any recognition error is promptly corrected manually by an operator using a touch-panel and a keyboard (Fig. 4). The texts of closed-captions can be colored differently to indicate who has made a comment. The height of the caption display on the screen can be flexibly controlled online by an operator to avoid overlapping with an open-caption. Closed-captions can be presented within 5 to 8 seconds of the original speech. We received a large number of positive responses from viewers about the simultaneous captioning. Hearing-impaired viewers expressed delight at finally being able to enjoy programs together with their families.



Fig. 3 Re-speaker.

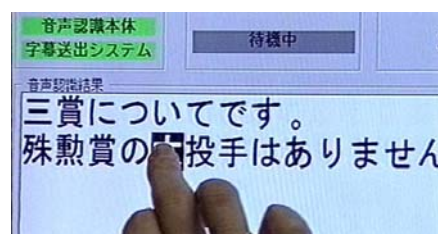


Fig. 4 Manual error correction.

4. Hybrid system

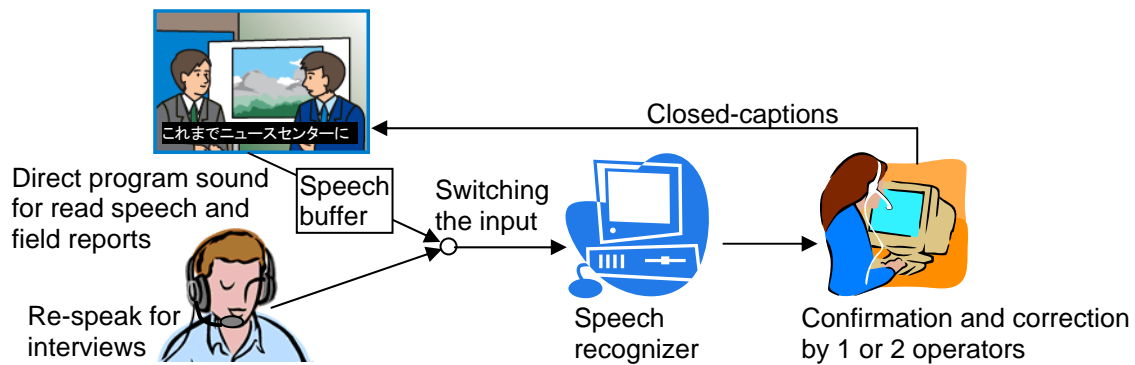


Fig. 5 New hybrid speech recognition system.

4.1. Overview

The progress made in our speech recognition algorithms has enabled our latest speech recognizer for news programs to directly recognize not only speech read by an anchorperson in a studio, but also field reports by a reporter, with sufficient word recognition accuracy of more than 95% [5][6]. However, as the recognition accuracy for other parts, such as conversations and interviews, can still be insufficient, we rely on the re-speak method for those parts. Therefore, the system we are currently developing is a hybrid which allows switching of the input speech for recognition between the program sound and the re-speaker's voice according to each news item. This allows an entire news program to be covered using only the automatic speech recognizer.

The new speech recognizer runs on a Linux server or a PC. It automatically detects the gender of a speaker, which allows use of more accurate gender-dependent acoustic models [8]. As the switching of the speech input is done manually with a small delay by the re-speaker, a speech buffer of about one second is used to avoid losing any speech beginnings of the direct program sound. Moreover, the new system employs a manual correction method that requires only one or two flexible correction operators depending on the difficulties of the speech recognition [6]. Four correction operators were needed in the previous news system (two sets of an error pointer and an error corrector) [2]. Therefore, we expect the new system will help to enable expansion of closed-captioned program coverage, especially for nationwide regular short news and local news programs since their news styles are based on comparatively simple direction with only one anchorperson.

4.2. Performance

In our experiment on such simple news programs with one anchorperson, the new system with two correction operators achieved caption accuracy of 99.9% without any fatal errors [6]. However, it is not yet good enough for large-scale news shows with more than one anchorperson and spontaneous and conversational speaking styles. We intend to improve the speech recognition accuracy for such speaking styles in the future.

5. Conclusion

NHK's current simultaneous-captioning systems for live TV programs with speech recognition technologies are based on the re-speak method which is suitable for sports programs. The system we are developing is based on a hybrid method of switching between the direct program sound and the re-speaker's voice for simple news programs. To expand the closed-captioned coverage of live programs efficiently, we intend to further refine speech recognition systems so that they will be able to cover a wide variety of live programs in the future.

6. References

- [1] Ministry of Internal Affairs and Communications, "Achievements of closed-captions," http://www.soumu.go.jp/s-news/2007/070629_9.html(in Japanese), 2007.
- [2] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Programs", *IEEE Transactions on Broadcasting*, 46(3): 189-196, 2000.
- [3] M. Marks, "A distributed live subtitling system," *BBC R&D White Paper*, WHP070, 2003.

- [4] T. Imai, A. Matsui, S. Homma, T. Kobayakawa, K. Onoe, S. Sato, and A. Ando, "Speech Recognition with a Re-Speak Method for Subtitling Live Broadcasts," Proceedings of International Conference on Spoken Language Processing, pp.1757-1760, 2002.
- [5] T. Imai, K. Onoe, S. Homma, S. Sato, and A. Kobayashi, "Study of Real-Time Captioning by Using Speech Recognition of Program Sound with a Re-Speak Method," Proceedings of Annual Meeting of The Institute of Image Information and Television Engineers (in Japanese), 2007.
- [6] S. Homma, K. Onoe, A. Kobayashi, S. Sato, T. Imai, and T. Takagi, "Experiment of Real-Time Captioning for Broadcast News Using Speech Recognition of Direct Program Sound and Re-Spoken Utterances," Proceedings of Annual Meeting of The Institute of Image Information and Television Engineers (in Japanese), 2007.
- [7] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe, and T. Kobayakawa, "Speech Recognition for Subtitling Japanese Live Broadcasts, Proceedings of The 18th International Congress on Acoustics (ICA), Vol. I, pp.165-168, 2004.
- [8] T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma, "Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News," Proceedings of Interspeech, pp.1602-1605, 2006.