Speech Recognition for Real-time Closed Captioning

Closed captioning to convey the speech of TV programs by text is becoming a useful means of providing information for elderly people and the hearing impaired, and real-time captioning of live programs is expanding yearly thanks to the use of speech recognition technology and special keyboards for high-speed input. This paper describes the current state of closed captioning, provides an overview of speech recognition technology, and introduces real-time closed-captioning systems using speech recognition. It also describes recent research at NHK STRL on speech recognition for improving recognition performance and operability and discusses technical issues for further research.

1. Introduction

Closed captioning is a form of broadcasting by which speech such as narration and dialog in a TV program are conveyed by text. It is becoming an important means of providing information for people such as the elderly and hearing impaired who have trouble hearing the audio portion of TV programs. Today, viewers may select whether to display captions according to personal preferences. Previously in analog broadcasting, viewers needed a special receiver for viewing captions, but in digital broadcasting, a closed captioning function is included as standard in TV receivers, and viewers need only to push a "closed captioning" button on their remote controls to display captions on their screens. One-Seg mobile terminals are also capable of closed captioning, and even people with normal hearing ability are using closed captioning in places such as trains where common courtesy dictates that sound be turned off. In short, closed captioning is becoming a universal means of providing information.

Closed captioning comes in two forms: offline captioning that captions prerecorded programs and real-time captioning that captions live programs. Although offline captioning can be completed before the program is broadcast and does not present any particular technical difficulties, real-time captioning has been technically difficult to achieve. Nevertheless, recent progress in speech recognition technology and the use of special keyboards for high-speed input have been driving the expansion of real-time closed captioning.

At NHK Science and Technology Research Laboratories (STRL), we are researching real-time speech recognition to efficiently produce the captions for a variety of live programs. To date, we have developed and implemented real-time closed captioning production systems using a direct method and a re-speak method. The former method automatically recognizes that portion of a news programs read by an anchorperson from a manuscript. The latter method recognizes the speech of a "re-speaker" who, while sitting in a room with no background noise, rephrases the speech of another speaker, e.g., a play-by-play announcer of a sports program. This paper begins by describing the current state of closed captioning and presenting an overview of speech recognition technology. It then presents speech recognition technology for real-time closed captioning focusing on the above systems implemented by NHK and future systems now under study. It also describes recent research on speech recognition technology for improving recognition performance and operability and discusses future technical issues.

2. Current State of Closed Captioning

As mentioned above, closed-captioned programs are expanding year by year, but that is not to say that all programs are being captioned. According to Japan's Ministry of Internal Affairs and Communications (MIC), the ratio of total broadcast time in fiscal year 2010 occupied by programs with closed captioning was 56.2% for NHK General TV (digital) and on average 43.8% for the five main commercial TV broadcasters in Tokyo (digital)[1]. Administrative guidelines issued by MIC in 2007 to promote the expansion of closed captioned programs as a goal closed captioning of all programs including live programs broadcast between the hours of 7:00 and 24:00 by FY2017 except for programs deemed difficult to caption such as talk shows and other live programs in which several persons tend to talk at the same time[2]. The ratio of captioned programs to those targeted for captioning by this guideline was 62.2% for NHK General TV in FY2010[1]. Researchers at NHK STRL are looking at ways to achieve efficient production of closed captioning to meet the goal established by MIC.

Broadcast programs include prerecorded programs like dramas and documentaries that are obviously completed before broadcasting and live programs like news and sports. The technique used for producing captions differs between these two types of programs. In the case of NHK General TV, all prerecorded programs in the time period from 7:00 to 24:00—which make up about 40% of all programs in that period—are captioned. Here, closed captioning is achieved by manually inputting the text using a personal computer and manually making adjustments to caption display position, caption timing,

and text color.

Closed captioning for live programs, on the other hand, requires real-time input of text. Broadcasters in the United States typically perform real-time closed captioning by having a single skilled operator input text using a stenographic keyboard developed for rapid input in the courtroom. The large number of homonyms in the Japanese language, however, requires that *kana* (Japanese syllabic characters) be converted to the appropriate *kanji* (Chinese characters) in real time. This makes it difficult for a single operator to input text using a standard keyboard on a personal computer. To deal with this situation, NHK presently uses four real-time closed-captioning methods using keyboards and speech recognition according to program type as summarized below.

(1) The coordinated input method[3] using standard keyboards in which multiple input operators take turns entering text for just a short interval of time in a relay manner [for programs with singing in them and information programs]
(2) A method using a special, high-speed stenographic keyboard (Stenoword[4]) allowing multiple keys to be pressed at the same time. In this method, several pairs of input operators and checkers take turns entering text for just a short interval of time in a relay manner [for news programs]
(3) The direct method[5] in which program speech is recognized and converted into text directly [for Major League Baseball broadcasts and (2000—2006) news programs]
(4) The re-speak method[6] in which the speech of an announcer specialized in rephrasing original program speech for closed captioning is recognized and converted into text [for sumo, Japanese professional baseball, and other sports events and information programs]

A keyboard-based input method allows an operator to enter text freely regardless of the program's topic, speaking style, background noise, etc., but it requires several skilled operators. A method based on speech recognition, meanwhile, requires pre-training of a dictionary for each program genre and may not be able to achieve sufficient recognition accuracy due to the effects of speaking style and background noise. Such a method is therefore limited in terms of the programs it can be used on, but it has the advantage of being able to train personnel for re-speaking and error correction relatively easily.

3. Overview of Speech Recognition Technology

Automatic speech recognition technology recognizes human speech in order to carry out tasks such as converting that speech into text or determining what the speaker wants to do with a certain device and controlling that device accordingly. The most popular speech recognition technique at present is the "stochastic method" in which recognition system is trained beforehand with a large quantity of text and speech data and the word closest to the spoken one is selected from among the words learned in this way. Although this technique is limited to certain applications such as car navigation, robot control, and closed captioning, it is capable of real-time continuous speech recognition of tens of thousands of words uttered by an unspecified number of speakers requiring no preregistration of voices. The following describes the standard procedure for achieving continuous speech recognition.

3.1 Acoustic Model

The human ear analyzes and uses the frequency components making up sound. In a similar manner, automatic speech recognition analyzes the ever-changing frequency components of sound through signal processing called "acoustic analysis" and extracts "acoustic features" (coefficients that express the peaks and valleys of the sound spectrum), which are deeply related to the content of human utterances. In human speech, frequency components differ greatly among different vowels and consonants, and they are therefore used as acoustic features in speech recognition. However, as with sound picked up by the human ear, the physical value of an acoustic feature may not be the same and the temporal length of a sound may differ even for the same word spoken by the same speaker. The frequency components of sound can be greatly affected by preceding and following vowels and consonants and by background noise.

The approach taken in light of the above variation is to use a large quantity of speech data to learn the range up to which the spread of an acoustic feature can be treated as the same vowel or consonant. In speech recognition for closed captioning, the process involves the training of an acoustic (stochastic) model called the hidden Markov model (HMM) using the acoustic features of several hundred hours of broadcast speech in order to express the spread of features in the voices of diverse speakers. To improve recognition performance, it is important to collect large quantities of speech data approximating that of actual broadcasts and to train the acoustic model by using good learning algorithms.

3.2 Language Model

Continuous speech recognition, which attempts to recognize phrases or sentences consisting of a sequence of uttered words, makes use of the fact that learned words making up a vocabulary often appear together (i.e., they tend to form connections). That is to say, there is a propensity for a certain word to be followed by another certain word as in "global"→ "warming." This statistical property can be expressed in terms of a probability (such as 0.8 or 0.3), and a collection of such probabilities can be used to create a "language model (dictionary)" (*N*-gram language model expressing the probability that a certain N-word sequence will appear). Such a language model is trained using words and phrases that tend to appear in programs targeted for closed captioning as determined by analyzing large quantities of text data beforehand. Here, as well, recognition performance can

be improved by collecting large quantities of expressions that appear in actual broadcasts and of text data related to topics covered in those broadcasts and then to train the language model accordingly. Words not learned as vocabulary cannot, in principle, be recognized, so it is important here that the language model be carefully trained

3.3 Word Search

The acoustic model is a set of data that expresses voice features, and the language model is a set of data that expresses the ease at which words connect with certain other words. The first step in automatic recognition of input speech is thus to analyze the acoustic features (frequency components) of the input speech and to narrow down candidate words by using the acoustic model to calculate which words registered in the dictionary (consisting, in general, of tens of thousands of words) come closest in terms of pronunciation to the acoustic features obtained from analysis (Figure 1). The next step is to refer to the language model to determine which words easily connect to which other words from among the words deemed acoustically possible in the first step to narrow down candidate words even further. At this time, homonyms in the input speech are determined by the relationship between the preceding and following words. Given the speech input, a computer can repeat this process of narrowing down words in terms of acoustics and linguistics to test a large number of word combinations. The combination of words having the highest probability of being correct is output as the recognition results. In real-time closed captioning, the time delay from speech input to caption display must be made as short as possible. The word-search algorithm used at NHK STRL quickly determines recognition results one after another without waiting for speaker utterances to conclude[7].

In this way, the procedure developed at NHK STRL for performing speech recognition is not a bottom-up process that recognizes vowels and consonants, converts them into *kana* (Japanese syllabic characters), and converts the *kana* into *kanji* (Chinese characters). Rather, it registers words beforehand in a dictionary together with phonetic-symbol sequences in mixed *kanji/kana* notation and outputs recognition results in a top-down manner as a combination of the most probable word/phonetic sequences based on acoustics and linguistic probabilities.

4. Real-time Closed-captioning System Using Speech Recognition

To date, NHK has implemented two types of real-time closed-captioning systems using speech recognition: the direct method, which, as the name implies, directly recognizes human speech in a program, and the re-speak method, which recognizes the speech of a specially trained announcer who rephrases the original program speech for the purpose of closed captioning. This section describes these two methods and introduces a new hybrid

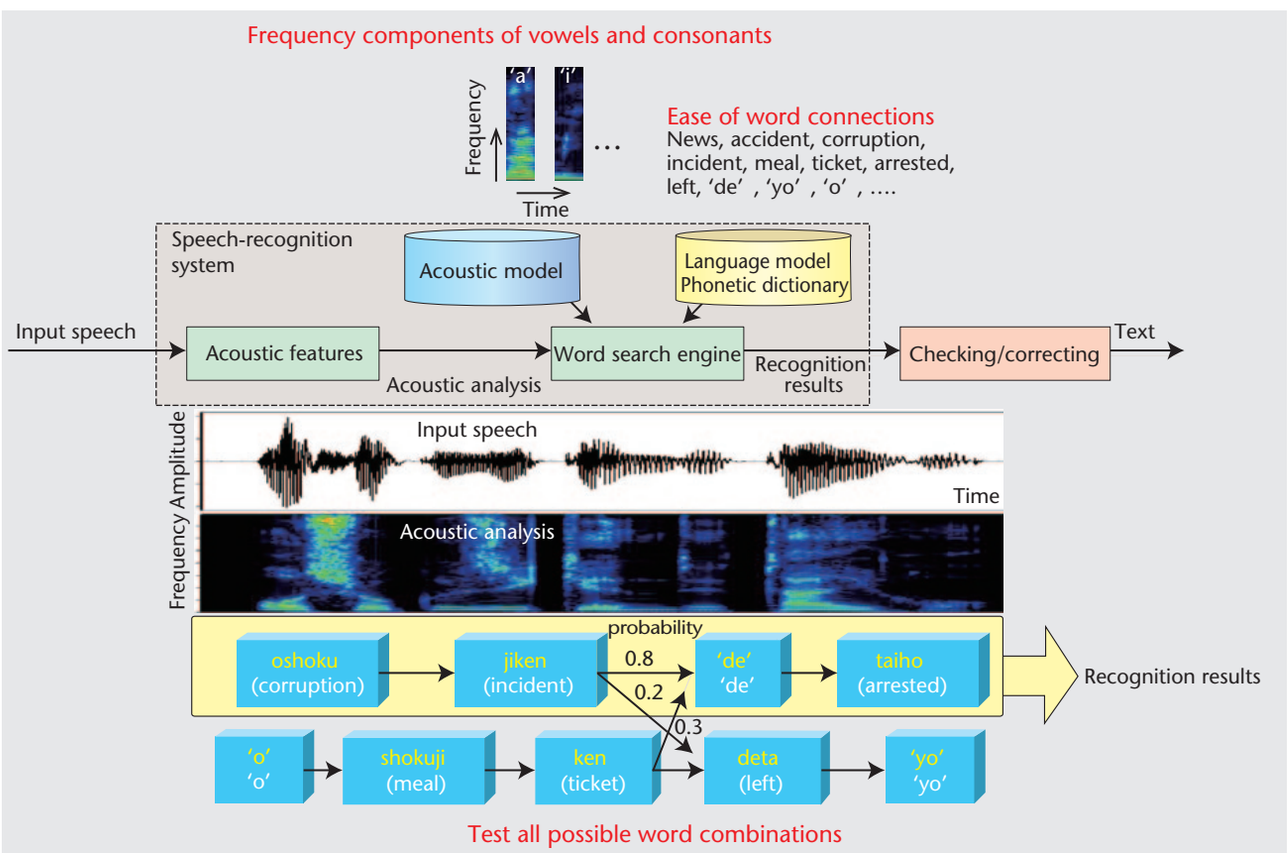

Figure 1: Processing procedure for speech recognition

method that combines those two methods for use in news programs.

4.1 Direct Method

In 2000, NHK implemented the world's first real-time closed-captioning system for broadcast news programs using the direct method of speech recognition (Figure 2)[5]. At that time, a recognition rate of 95%—a level appropriate for practical use—could only be achieved for an anchorperson's speech when reading from a manuscript, which meant that the system could not be used for a whole program. Nevertheless, the use of this system marked the first time in Japan that captions were generated for a live broadcast program by direct speech recognition, which had repercussions throughout the industry. As shown in the Figure 2, this system consists of separate speech-recognition equipment for male and female speech and employs four operators for checking and correcting recognition results (two pairs of operators each consisting of an error detector and error corrector). In this checking/correcting process, detected errors are quickly corrected using a touch panel or keyboard so that captions are broadcast as fast as possible. The speech-recognition system, moreover, can automatically obtain the latest news manuscript (anchorperson's manuscript in digital form prior to any handwritten corrections) for that program before broadcasting for use in weighted training of the language model[8].

It was later decided to use a specialized stenographic keyboard for high-speed input (Stenoword[4]) for program portions other than those in which the anchorperson is reading from a manuscript. Such portions include field reports or discussion segments. The production of closed captioning in a news program thus used this method for some program segments and the direct method for other segments. In short, there was a time at NHK in which automatic speech recognition was used for the closed captioning of news programs though only for those portions of the program in which an anchorperson read from a manuscript. In 2006, it was decided that closed captioning for all news programs would be performed using high-speed stenographic keyboards. At present, the direct method is being used in broadcasts of Major League Baseball games to recognize the speech of a play-by-play announcer situated inside a studio in Japan.

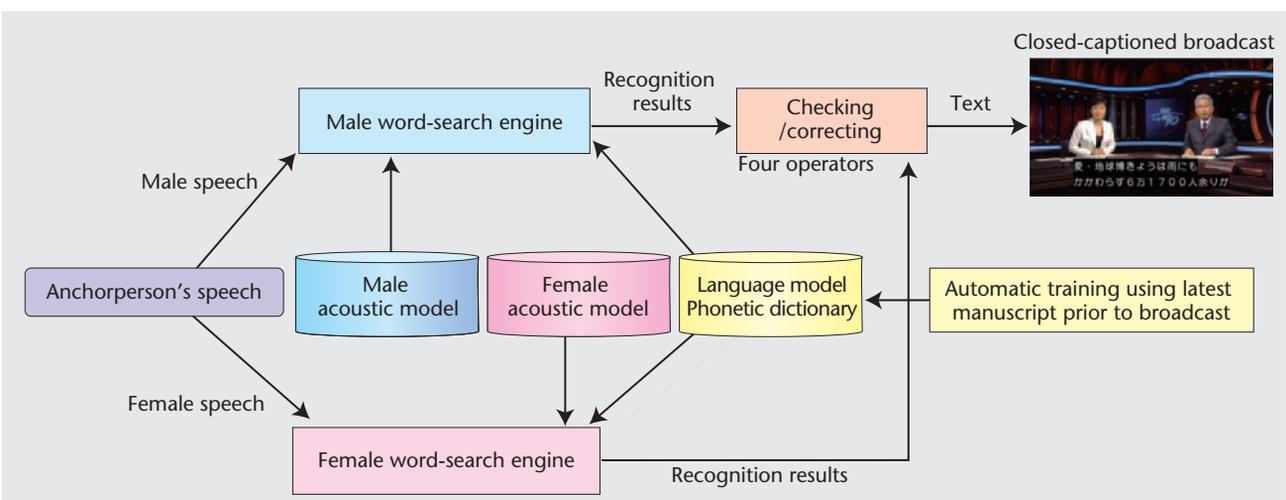

Figure 2: Closed-captioning system for news programs by the direct method

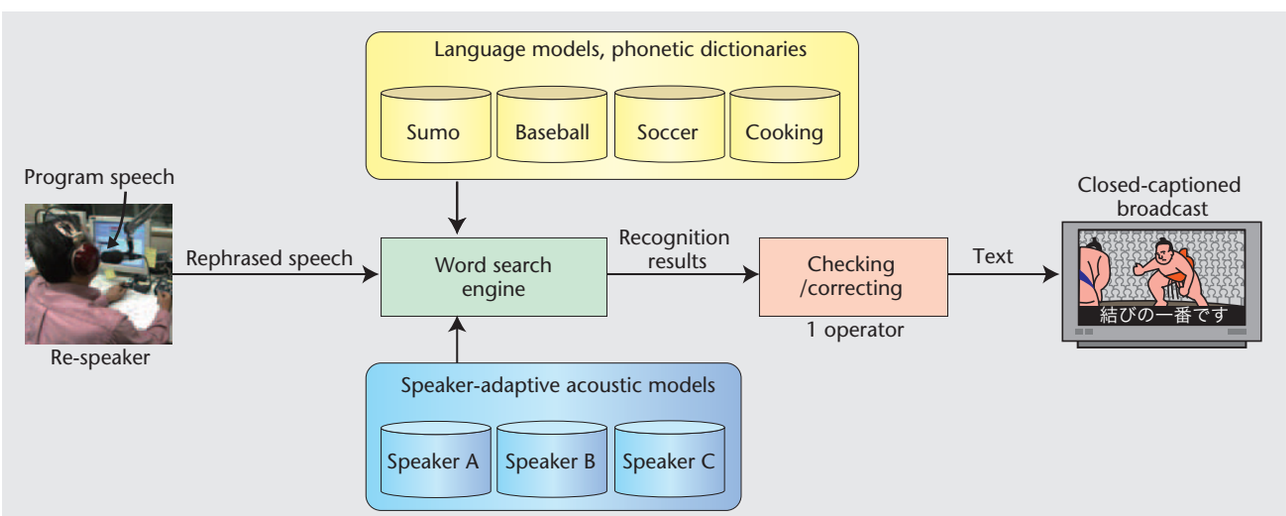

Figure 3: Real-time closed-captioning system by the re-speak method

4.2 Re-speak Method

NHK developed a real-time closed-captioning system using the re-speak method in collaboration with a major electronics manufacturer to enable captioning of live broadcasts for non-news programs such as sports events and information programs (Figure 3)[9]. Beginning in 2001 with the popular "Red and White" year-end singing contest, this system has been used to caption a variety of programs including coverage of the Olympic Games, world-cup soccer, grand sumo tournaments, and Japanese professional baseball[10]. In the re-speak method, a specially trained re-speaker listens to live program speech of the actual play-by-play announcer or commentator through headphones and repeats and/or summarizes that speech in a quiet booth. The re-speaker can rearrange the content of original program speech to achieve a sufficiently high rate of speech recognition for closed captioning even for programs with considerable background noise, multiple speakers, etc. For sports programs, moreover, with the aim of shortening the caption-display delay (5—10 sec.) as much as possible and making captions easier to read and understand, the re-speaker can omit what is visually understood and simplify or rephrase the content of the original speech[6], or even supplement the original speech with descriptions of on-site clapping, cheering, and other happenings not described by the play-by-play announcer. In addition, sometimes, the re-speaker may be able to re-vocalize speech for which erroneous recognition results were obtained in time for presentation.

As shown in Figure 3, this system makes use of multiple language models trained for different programs types by using genre-specific training texts. Adaptive acoustic models trained for different re-speakers can also be used to increase the recognition rate. This system enables re-speakers to take turns during a live broadcast by instantly switching from one acoustic model to another without interrupting online processing.

4.3 Hybrid Method

Current speech-recognition technology provides a practical level of speech-recognition performance (an average word-recognition rate of 95% or better) for those portions of news programs in which anchorpersons read from a manuscript, reporters make reports from the field, or anchorpersons converse with a reporter. On the other hand, speech-recognition performance is low for interviews, edited video, and discussions with lots of spontaneous speech due to the effects of background noise and music, indistinct utterances, rapid speaking, casual expressions, etc.

For the above reason, NHK has developed a real-time closed-captioning system that uses a hybrid method[11]: the direct method is used to recognize speech in program segments corresponding to a high recognition rate, and the re-speak method is used for other program segments such as interviews (Figure 4). This hybrid system allows the re-speaker to manually switch between inputting actual program speech or the re-speaker's speech to the speech-recognition system, and it uses a speech-buffering mechanism to prevent leading utterances from being lost at the time of switching. The system also has an improved user interface to enable multiple operators to simultaneously check and correct recognition results[11][12] and the number of operators can be increased or decreased (1—2 persons) depending on the difficulty of caption production. This hybrid approach makes it possible to caption an entire news program through speech recognition, which was previously difficult to do. It also incorporates recent improvements to elemental speech-recognition technologies such as automatic gender detection. NHK expects the hybrid method to be part of an efficient, low-cost closed-captioning system, and it plans to use it for news programs in the near future. A prototype version of this hybrid closed-captioning system, constructed by NHK STRL, was used by the NHK Broadcasting Center after the Great East Japan Earthquake in 2011 to caption news programs covering the disaster.

5. Current Research on Speech Recognition

Speech-recognition technology for real-time closed captioning must satisfy a variety of conditions, the foremost of which are 1) accuracy, 2) real-time processing,

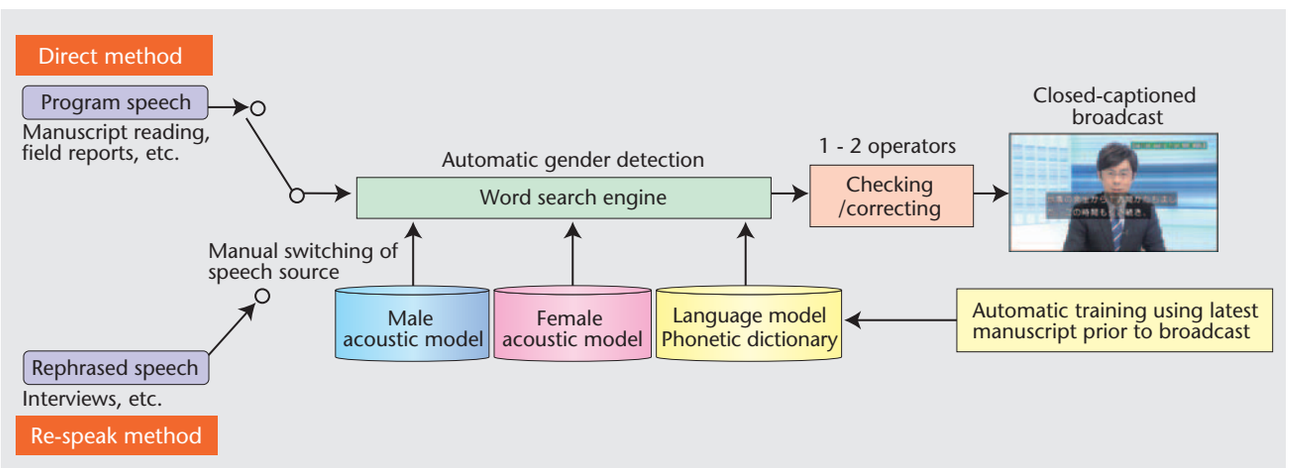

Figure 4: Real-time closed-captioning system using the hybrid method

3) efficient checking and correction, and 4) ease of preparation. For this reason, we made the improvements described below to facilitate the use of speech recognition using the hybrid method for closed captioning of news programs. First, to improve accuracy, we used online speech detection and dual-gender speech recognition[13], selective discriminative training of an acoustic model taking speech deformation into account[14)15)], and automatic updating of the language model[16]. Second, to improve the real-time characteristics of speech recognition, we decreased the amount of computational processing and shortened the time taken to display captions. Third, to make checking and correcting more efficient, we improved the user interface[11)12)] and improved the operability of equipment while decreasing the number of required operators. Finally, to make advance preparations easier, we undertook the measures described in section 5.3.

Furthermore, to improve the recognition performance of spontaneous speech as in program personalities engage in casual conversation, we are also researching ways of dealing with equivalent expressions in spontaneous speech[17], speaker adaptation using speaker identification[18], and discriminative rescoring[*1] for learning trends in recognition errors and improving the recognition rate[19]. In the following, we describe automatic updating of the language model and a method for dealing with equivalent expressions in spontaneous speech.

5.1 Automatic Updating of the Language Model[16]

In closed captioning for news programs, the appearance of new proper nouns and topics must be dealt with promptly to reduce recognition errors. To this end, we use digital manuscripts that serve as a basis for the manuscripts to be read by anchorpersons as adaptive training data and continuously subject the language model and phonetic dictionary to adaptive training using the latest digital manuscripts. Additionally, we have modified part of the hybrid closed-captioning system to enable digital manuscripts to be read in and used for training at any time even during the broadcast of a news program (Figure 5). In the past, it was necessary to restart the word search engine whenever updating the language model with a news manuscript, but in the new system, the speech-recognition controller starts up a second word search engine at the time of a language-model update and switches to the new model without interrupting the speech-recognition process. In this way, closed captioning can likewise continue uninterrupted using the most up-to-date language model even during a broadcast and recognition errors can be decreased to about one-third that of previous levels. Operability can be improved as well.

5.2 Method of Handling Equivalent Expressions in Spontaneous Speech[17]

Talk shows often have guests who use casual expressions or who speak in a rapid, indistinct manner. This speech is significantly different both linguistically and acoustically from the training speech that uses standard, clearly spoken language. This mismatch must be resolved to improve the recognition rate. To improve speech recognition in such situations, we conducted research using NHK's popular "Close-up Gendai" news-related talk show as representative programs in which guests speak freely on a certain topic. In general, training data for a language model draws heavily upon the written language, and training using colloquial expressions unique to the spoken language is insufficient.

[*1] "Rescoring" is a method that first calculates an acoustic probability (score) and linguistic probability (score) using a simple model to narrow down candidate words and a more detailed model to recalculate the score and to determine the final recognition results.

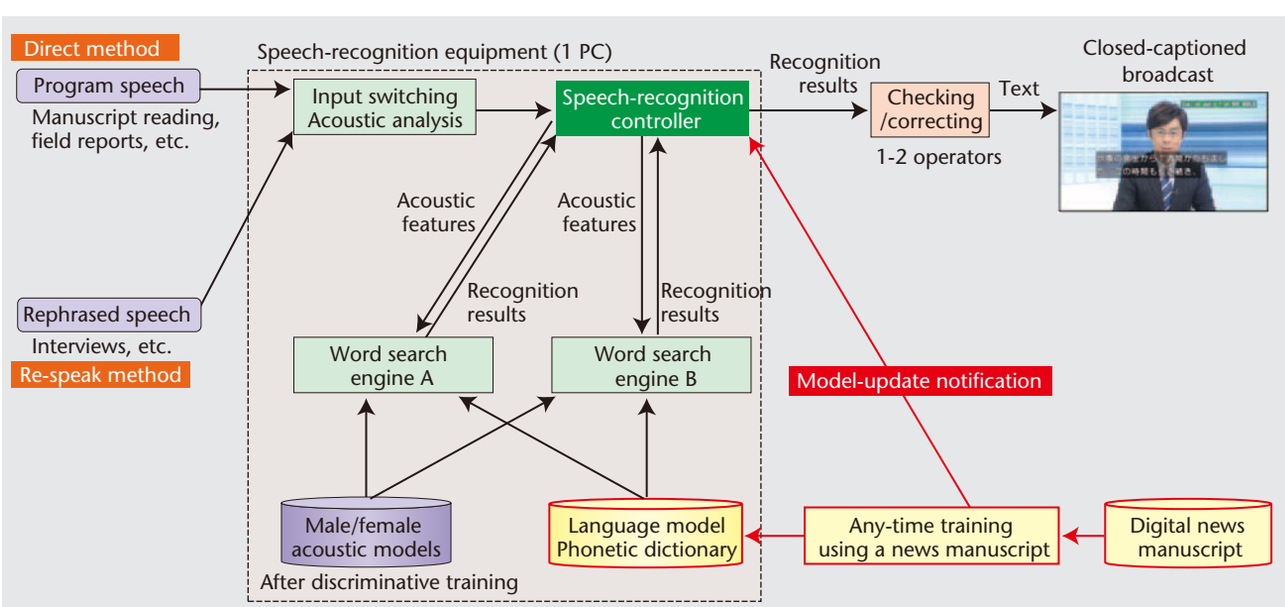

Figure 5: Automatic updating of language model

To resolve this linguistic mismatch, we investigated a technique for correcting the occurrence probability of words in the language model by extracting expressions having the same meaning from among expressions used in the spoken language. For example, in the *N*-gram language model, the occurrence probability of the spoken word "te-iu" (meaning "what is called…") can be averaged out with the occurrence probability of the written word "to-iu" (having the same meaning) and the two expressions can be treated as being equivalent in word searches. In addition, acoustic mismatches can be resolved by training the acoustic model to reflect trends in recognition errors unique to the spoken language.

As a result of combining the various measures described above, we were able to reduce the word-recognition error rate in the above evaluation task from 24.5% to 11.3% (error-reduction rate of 54%).

5.3 Future Issues

Large-vocabulary continuous speech-recognition technology for real-time closed captioning has reached a practical level for news, sports, and information programs, but it has not, unfortunately, reached a level suitable for all programs. The performance of speech recognition is considered to be affected somewhat by the "3W1H" conditions, that is, by "who" (speaker), "what" (subject), "where" (acoustic environment), and "how" (speaking style) as summarized in Table 1. Speech recognition using the direct method currently has a recognition rate of about 90% or better, but this is limited to the conditions to the left of center in the table. We are presently researching ways of improving the recognition rate targeting information programs like "Close-up Gendai" and tabloid shows that cover diverse topics.

For real-time closed captioning, we can expect online adaptive techniques for handling different speakers, acoustic environments, and topics to become increasingly important. There are also many problems related to speech recognition that have yet to be solved such as overlapping speech from multiple speakers, verbal mistakes, hesitations, laughter, foreign speech, and background noise.

In the case of a news program, a digital manuscript submitted by a reporter can be used for training the language model. However, for other live programs (especially information programs such as tabloid shows), the general content of the program can be obtained beforehand only from the several pages of information contained in the program configuration table. The collection of textual material that can be used for training purposes requires considerable effort at present. As a result, closed captioning using speech recognition can only be performed for programs for which textual materials for training of the language model can be adequately obtained and regularly scheduled programs for which the cost of collecting training data can be lowered. To expand closed captioning to a wide variety of programs, it is important that advance preparations such as vocabulary selection, registration of new words, and collection of training texts be easy to perform in an efficient manner even by non-specialists in speech recognition. Looking forward, we seek to construct a speech recognition system that enables the language model and acoustic model to adaptively and autonomously grow to keep up with ever-changing topics and the diverse voices of re-speakers without having to expend considerable effort in pre-training and with only a limited amount of information from the program configuration table. Additionally, while the re-speak method is presently a practical and effective, it will be necessary to improve the direct method in the future for the sake of reducing the cost of closed captioning even further.

As long as even a small possibility exists of errors in speech-recognition results in the closed-captioning process for broadcast programs, manual checking and correcting of recognition results will be essential. A closed-captioning system using speech recognition must be able to handle errors in a short period of time. This goes for the entire system including the human component consisting of re-speakers, checkers, and correctors.

6. Conclusion

This paper described the current state of closed captioning in broadcast programs and provided an overview of speech-recognition technology. It also introduced the current speech-recognition systems for real-time closed captioning and systems currently under study, presented recent research results, and pointed out technical issues for future research. Since the beginning of real-time closed captioning using speech recognition, NHK has received praise from elderly and hearing impaired people for the increasing number of captioned programs it has offered them to enjoy together with their families. Unfortunately, the current level of speech-recognition technology is not capable of captioning all programs. Researchers at NHK STRL will thus continue to work to solve the various acoustic and linguistic problems described in this paper. They will also make better use of what we already know about non-linguistic information including context, meaning, and intonation that people use in recognizing speech together with auditory psychological characteristics,

Table 1: Relationship between broadcast-speech conditions and recognition difficulty

3W1H Conditions	Low ← Difficulty → High				
Who (speaker)	Anchorperson	Reporter	Guest	Entertainer	General
What (subject)	Manuscript	Restricted topic	In-depth topic	Highly diverse	Undetermined topics
Where (acoustic environment)	Studio	Low noise	Background noise	High noise	Noisy conditions
How (speaking style)	Reading	Clear	Spontaneous	Unclear	Friendly chat	Emotional

all with the aim of developing speech-recognition technology approximating human capabilities. To meet the targets set forth by the Ministry of Internal Affairs and Communications for the expansion of closed captioning, NHK seeks to develop a more efficient closed-captioning system with lower operating costs that even local broadcasting stations can implement. It will also study a variety of new developments in the field of broadcasting such as the use of speech-recognition technology for the creation of metadata[20].

(Toru Imai)

References

1) Ministry of Internal Affairs and Communications: "Status of Closed Captioning for FY2010," Press Release (2011) (in Japanese) http://www.soumu.go.jp/menu_news/s-news/01ryutsu05_01000012.html
2) Ministry of Internal Affairs and Communications: "Administrative Guidelines on the Spread of Broadcasting for the Visually and Hearing Impaired," Press Release (2007) (in Japanese) http://www.soumu.go.jp/menu_news/s-news/2007/071030_2.html
3) Ministry of Internal Affairs and Communications and Mitsubishi Research Institute: "Survey of Domestic and Overseas Broadcasting on Broadcasting for the Visually and Hearing Impaired" (2006) (in Japanese) http://www.soumu.go.jp/main_sosiki/joho_tsusin/b_free/pdf/060810_1.pdf
4) S. Nishikawa, H. Takahashi, M. Kobayashi, Y. Ishihara and K. Shibata: "Real-Time Japanese Captioning System for the Hearing Impaired Persons," IEICE Transactions, Vol. J78-D-II, No. 11, pp. 1589-1597 (1995) (in Japanese)
5) A. Ando, T. Imai, A. Kobayashi, S. Homma, J. Goto, N. Seiyama, T. Mishima, T. Kobayakawa, S. Sato, K. Onoe, H. Segi, A. Imai, A. Matsui, A. Nakamura, H. Tanaka, T. Takagi, E. Miyasaka, and H. Isono: "Simultaneous Subtitling System for Broadcast News Programs with a Speech Recognizer," IEICE Transactions on Information and Systems, Vol.E86-D, No.1, pp. 15-25 (2003)
6) A. Matsui, S. Homma, T. Kobayakawa, K. Onoe, S. Sato, T. Imai and A. Ando: "Simultaneous Subtitling for Live Sports Using a "Re-speak" Method with Paraphrasing," IEICE Transactions, Vol. J87-D-II, No. 2, pp. 427-435 (2004) (in Japanese)
7) T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando: "Progressive 2-Pass Decoder for Real-Time Broadcast News Captioning," Proc. IEEE ICASSP, pp. 1559-1562 (2000)
8) A. Kobayashi, T. Imai, A. Ando and K. Nakabayashi: "Time Dependent Language Model for Broadcast News Transcription (Special Issue on Spoken Language Processing," IPSJ Journal, Vol. 40, No. 4, pp. 1421-1429 (1999) (in Japanese)
9) NHK Technical Information: "Real-time Closed-captioning Production System by Automatic Speech Recognition," (2003) (in Japanese) http://www3.nhk.or.jp/pr/marukaji/m-giju093.html
10) T. Hattori, T. Shiina and H. Domen: "Closed Captioning Service for Broadcasting Programs: Systems and Services," ITE Technical Report, BCT2004-24, Vol. 28, No. 5, pp. 17-20 (2004) (in Japanese)
11) S. Homma, A. Kobayashi, T. Oku, S. Sato, T. Imai and T. Takagi: "Real-Time Closed-Captioning System Using Speech Recognition of Direct Program Sound and Re-Spoken Utterances," The Journal of the Institute of Image Information and Television Engineers," Vol. 63, No. 3, pp. 331-338 (2009) (in Japanese)
12) S. Oode, T. Mishima, M. Emoto, A. Imai and T. Takagi: "An Efficient Real-Time Correction System of Speech Recognition Errors for TV News Closed Captions," Proceedings of the ITE Convention, 6-3 (2003) (in Japanese)
13) T. Imai, S. Sato, S. Homma, K. Onoe and A. Kobayashi: "Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News," IEICE Transactions on Information and Systems, Vol. E90-D, No. 8, pp. 1286-1291 (2007)
14) D. Povey and P.C. Woodland: "Minimum Phone Error and I-smoothing for Improved Discriminative Training," Proc. IEEE ICASSP, pp. I-105-108 (2002)
15) T. Imai, A. Kobayashi, S. Sato, S. Homma, T. Oku and T. Takagi: "Improvements of the Speech Recognition Technology for Real-Time Broadcast Closed-Captioning," Proceedings of the Spoken Document Processing Workshop 2, pp. 113-120 (2008) (in Japanese)
16) T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato: "Speech Recognition with a Seamlessly Updated Language Model for Real-Time Closed-Captioning," Proc. Interspeech, pp. 262-265 (2010)
17) S. Homma, S. Sato, T. Oku, A. Kobayashi, T. Imai and T. Takagi: "Improvements in Automatic Speech Recognition of Spontaneous Speech in News-related Talk Shows," Proceedings of Spring Meeting of Acoustic Society of Japan, 3-Q-17, pp. 243-244 (2009) (in Japanese)
18) T. Oku, S. Sato, A. Kobayashi, S. Homma, and T. Imai: "Low-Latency Speaker Diarization Based on Bayesian Information Criterion with Multiple Phoneme Classes," Proc. IEEE ICASSP, pp. 4189-4192 (2012)
19) A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai and T. Takagi: "Discriminative Rescoring Based on Minimization of Word Errors for Transcribing Broadcast News," Proc. Interspeech, pp. 1574-1577 (2008)
20) A. Kobayashi, T. Oku, S. Homma, S. Sato and T. Imai: "A Broadcast News Transcription System for Content Application," IEICE Transactions on Information and Systems, Vol. J93-D, No. 10, pp. 2085-2095 (2010) (in Japanese)