Project for production of closed-caption TV programs for the hearing impaired

Takahiro Wakao
Telecommunications Advancement Organization of Japan
Uehara Shibuya-ku, Tokyo 151-0064, Japan
wakao@shibuya.tao.or.jp

Terumasa Ehara
NHK Science and Technical Research Lab / TAO

Eiji Sawamura
TAO

Ichiro Maruyama
TAO

Katsuhiko Shirai
Waseda University, Department of Information and Computer Science / TAO

Abstract

We describe an on-going project whose primary aim is to establish the technology of producing closed captions for TV news programs efficiently using natural language processing and speech recognition techniques for the benefit of the hearing impaired in Japan. The project is supported by the Telecommunications Advancement Organisation of Japan with the help of the ministry of Posts and Telecommunications.

We propose natural language and speech processing techniques should be used for efficient closed caption production of TV programs. They enable us to summarise TV news texts into captions automatically, and synchronise TV news texts with speech and video automatically. Then the captions are superimposed on the screen.

We propose a combination of shallow methods for the summarisation. For all the sentences in the original text, an importance measure is computed based on key words in the text to determine which sentences are important. If some parts of the sentences are judged unimportant, they are shortened or deleted. We also propose keyword pair model for the synchronisation between text and speech.

Introduction

The closed captions for TV programs are not provided widely in Japan. Only 10 percent of the TV programs are shown with captions, in contrast to 70 % in the United States and more than 30 % in Britain. Reasons why the availability is low are firstly the characters used in the Japanese language are complex and many. Secondly, at the moment, the closed captions are produced manually and it is a time-consuming and costly task. Thus we think the natural language and speech processing technology will be useful for the efficient production of TV programs with closed captions.

The Telecommunications Advancement Organisation of Japan with the support of the ministry of Posts and Telecommunications has initiated a project in which an electronically available text of TV news programs is summarised and syncrhorinised with the speech and video automatically, then superimposed on the original programs.

It is a five-year project which started in 1996, and its annual budget is about 200 million yen.

In the following chapters we describe main research issues in detail and the project schedule, and the results of our preliminary research on the main research topics are presented.

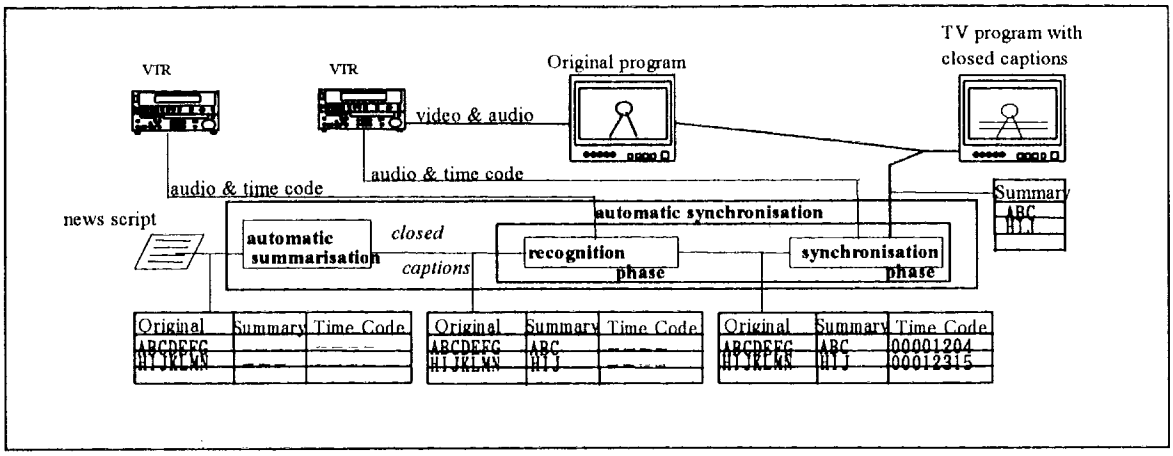

Figure 1 System Outline

1 Research Issues

Main research issues in the project are as follows:

- automatic text summarisation
- automatic synchronisation of text and speech
- building an efficient closed caption production system

The outline of the system is shown in Figure 1. Although all types of TV programs are to be handled in the project system, the first priority is given to TV news programs since most of the hearing impaired people say they want to watch closed-captioned TV news programs. The research issues are explained briefly next.

1.1 Text Summarisation

For most of the TV news programs today, the scripts (written text) are available electronically before they are read out by newscasters. Japanese news texts are read at the speed of between 350 and 400 characters per minute, and if all the characters in the texts are shown on the TV screen, there are too many of them to be understood well (Komine *et al.* 1996).

Therefore we need to summarise the news texts to some extent, and then show them on the screen. The aim of the research on automatic text summarisation is to summarise the text fully or partially automatically to a proper size to obtain closed captions. The current aim is 70% summarisation in the number of characters.

1.2 Synchronisation of Text and Speech

We need to synchronise the text with the sound, or speech of the program. This is done by hand at present and we would like to employ speech recognition technology to assist the synchronisation.

First, synchronising points between the original text and the speech are determined automatically (recognition phase in Figure1). Then the captions are synchronised with the speech and video (synchronisation phase in Figure1).

1.3 Efficient Closed Caption Production System

We will build a system by integrating the summarisation and synchronisation techniques with techniques for superimposing characters on to the screen. We have also conducted research on how to present the captions on the screen for the handicapped people.

2 Project Schedule

The project has two stages: the first 3 years and the rest 2 years. We research on the above issues and build a prototype system in the first stage. The prototype system will be used to produce closed captions, and the capability and functions of the system will be evaluated. We will focus on improvement and evaluation of the system in the second stage.

3 Preliminary Research Results

We describe results of our research on automatic summarisation and automatic synchronisation of text and speech. Then, a study on how to present captions on TV screen to the hearing impaired people is briefly mentioned.

3.1 Automatic Text Summarisation

We have a combination of shallow processing methods for automatic text summarisation. The first is to compute key words in a text and importance measures for each sentence, and then select importanct sentences for the text. The second is to shoten or delete unimportant parts in a sentence using Japanese language-specific rules.

3.1.1 Sentence Extraction

Ehara found that compared with newspaper text, TV news texts have longer sentences and each text has a smaller number of sentences (Ehara *et al* 1997). If we summarise TV news text by selecting sentences from the orignal text, it would be 'rough' summarisation. On the other hand, if we devide long sentences into smaller units, thus increase the number of sentences in the text, we may have finer and better summarisation (Kim & Ehara 1994). Therefore what is done in the system is that if a sentence in a given text is too long, it will be partitioned into smaller units with minimun changes made to the original sentence.

To compute importance measures for each sentence, we need to find first key words of the text. We tested high-frequency key word method (Luhn 1957, Edumundson 1969) and a TF-IDF-based (Text frequency, Inverse Document Frequency) method. We evaluated the two methods using ten thousand TV news texts, and found that high-frequency key word method showed slightly better results than the method based on TF-IDF scores (Wakao *et al* 1997).

3.1.2 Rules for shortening text

Another way of reducing the number of characters in a Japanese text, thus summarising the text, is to shorten or delete parts of the sentences. For example, if a sentence ends with a *sahen* verb followed by its inflection, or helping verbs or particles to express proper politeness, it does not change the meaning much even if we keep only the verb stem (or *sahen* noun) and delete the rest of it. This is one of the ways found in the captions to shorten or delete unimportant parts of the sentences.

We analysed texts and captions in a TV news program which is broadcast fully captioned for the hearing impaired in Japan. We complied 16 rules. The rules are devided into 5 groups. We describe them one by one below.

1) **Shotening and deletion of sentence ends**

We find some of phrases which come at the end of the sentence can be shortened or deleted. If a *sahen* verb is used as the main verb, we can change it to its *sahen* noun.
For example:

- ... *keikakushiteimasu* (計画しています)
  → ... *keikaku* (計画)

(note: keikakusuru = plan, sahen verb)

If the sentence ends in a reporting style, we may delete the verb part.

- ... *bekida to nobemashita*
  (べきだと述べました)
  → *bekida* (べきだ)

(bekida = should, nobemashita = have said)

2) **Keeping parts of sentence**

Important noun phrases are kept in captions, and the rest of the sentence is deleted.

- *taihosaretano ha Matumoto shachou*
  (逮捕されたのは松本社長)
  → *taiho Matumoto shachou*
  (逮捕 松本社長)

(taiho = arrest, shachou = a company president, Matumoto = name of a person)

3) **Replacing with shorter phrase**

Some nouns are replaced with a simpler and shoter phrase.

- *souridaijin* (総理大臣)→ *shushou* (首相)

(souridaijin, shushou both mean a prime minister)

4) **Conneticting phrases omitted**

Connecting phrases at the beginning of the sentence may be omitted.

- *shikashi* (しかし = however),
  ippou (一方 = on the other hand)

5) **Time expressions deleted**

Comparative time expressions such as today (kyou 今日), yesterday (kinou, 昨日) can be deleted. However, the absolute time expressions such as May, 1998 (１９９８年５月) stay unchanged in summarisation.

When we apply these rules to selected important sentences, we can reduce the size of text further 10 to 20 percent.

3.2 Automatic Synchronisation of Text and Speech

We next synchronise the text and speech. First, the written TV news text is changed into a stream of phonetic transcriptions. Second, we try to detect the time points of the text and their corresponding speech sections. We have developed 'keyword pair model' for the synchronisation which is shown in Figure 2.

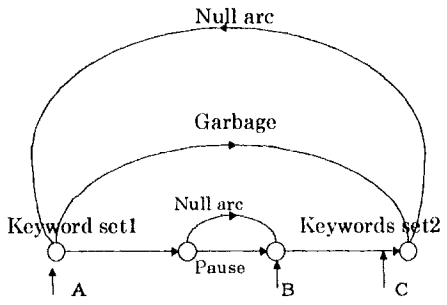

Figure 2 Keyword Pair Model

The model consists of two sets of words (keywords1 and keywords2) before and after the synchronisation point (point B). Each set contains one or two key words which are represented by a sequence of phonetic HMMs (Hidden Markov Models). Each HMM is a three-loop, eight-mixture-distribution HMM. We use 39 phonetic HMMs to represent all Japanese phonemes.

When the speech is put in the model, non-synchronising input data travel through the garbage arc while synchronising data go through the two keyword sets, which makes the likelihood at point B increase. Therefore if we observe the likelihood at point B and it becomes bigger than a certain threshold, we decide it is the synchronisation point for the input data.

Thirty-four (34) keywords pairs were taken from the data which was not used in the training and selected for the evaluation of the model. We used the speech of four people for the evaluation.

The evaluation results are shown in Table 1. They are the accuracy (detection rate) and false alarm rate for the case that each keyword set has two key words. The threshold is computed as logarithm of the likelihood which is between zero and one, thus it becomes less than zero.

Threshold	Detection rate (%)	False Alarm Rate (FA/KW/Hour)
-10	34.56	0
-20	44.12	0
-30	54.41	0
-40	60.29	0
-50	64.71	0.06
-60	69.12	0.06
-70	69.85	0.06
-80	71.32	0.12
-90	78.68	0.18
-100	82.35	0.18
-150	91.18	0.54
-200	94.85	1.21
-250	95.59	1.81
-300	99.26	2.41

Table 1 Synchronisation Detection

As the threshold decreases, the detection rate increases, however, the false alarm rate increases little (Maruyama 1998).

3.3 Speech Database

We have been gathering TV and radio news speech. In 1996 we collected speech data by simulating news programs, *i.e.* TV news texts were read and recorded sentence by sentence in a studio. It has seven and a half houses of recordings of twenty people (both male and female). In 1997 we continued to record TV news speech by simulation, and recorded speech data from actual radio and TV programs. It has now ten hours of actual radio recording and ten hours of actual TV programs. We will continue to record speech data and increase the size of the database in 1998.

3.4 Caption Presentation

We have conducted a study, though on small scale, on how to present captions on TV screen

to the hearing impaired people. We superimposed captions by hand on several kinds of TV programs. They were evaluated by the hadicapped people (hard of hearing persons) in terms of the following points :

- characters : size, font, colour
- number of lines
- timing
- location
- methods of scrolling
- inside or outside of the picture (see two examples below).

Figure 3 Captions in the picture

Figure 4 Captions outside of the picture

Most of the subjects preferred 2-line, outside of the picture captions without scrolling (Tanahashi, 1998). This was still a preliminary study, and we plan to conduct similar evaluation by the hearing impaired people on large scale.

Conclusion

We have described a national project, its research issues and schedule, as well as preliminary research results. The project aim is to establish language and speech processing technology so that TV news program text is summarised and changed into captions, and synchronised with the speech, and superimposed to the original program for the benefits of the hearing impaired. We will continue to conduct research and build a prototype TV caption production system, and try to put it to a practical use in the near future.

Acknowledgements

We would like to thank Nippon Television Network Corporation for letting us use the pictures (Figure 3, 4) of their news program for the purpose of our research.

References

Edmundson, H.P. (1969) *New Methods in Automatic Extracting* Journal of the ACM, 16(2), pp 264-285.

Ehara, T., Wakao, T., Sawamura, E., Maruyama I., Abe Y., Shirai K. (1997) *Application of natural language processing and speech processing technology to production of closed-caption TV programs for the hearing impaired* NLPRS 1997

Kim Y.B., Ehara, T. (1994) *A method of partitioning of long Japanese sentences with subject resolution in J/E machine translation,* Proc. of 1994 International Conference on Computer Processing of Oriental Languages, pp.467-473.

Komine, K., Hoshino, H., Isono, H., Uchida, T., Iwahana, Y. (1996) *Cognitive Experiments of News Captioning for Hearing Impaired Persons* Technical Report of IECE (The Institution of Electronics, Information and Communication Engineers), HCS96-23, in Japanese, pp 7-12.

Luhn, H.P. (1957) *A statistical approach to the mechanized encoding and searching of literary information* IBM Journal of Research and Development, 1(4), pp 309-317.

Maruyama, I., Abe, Y., Ehara, T., Shirai, K. (1998) *A Study on Keyword spotting using Keyword pair models for Synchronization of Text and Speech,* Acoustical Society of Japan, Spring meeting, 2-Q-13, in Japanese.

Tanahashi D. (1998) *Study on Caption Presentation for TV news programs for the hearing impaired.* Waseda University, Department of Information and Computer Science (master's thesis) in Japanese.

Wakao, T., Ehara, E., Sawamura, E., Abe, Y., Shirai, K. (1997) *Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired.* ACL 97 workshop, Natural Language Processing for Communication Aids, pp 55-58.