Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Evaluation of Speech Balloon Captions for Auditory Information Support in Small Meetings

Ippei Hisaki , Hiroaki Nanjo , and Takehiko Yoshimi

Graduate School of Science and Technology, Ryukoku University, Seta Otsu 520-2194, Japan
{hisaki, nanjo, yoshimi}@nlp.i.ryukoku.ac.jp

PACS: 43.72.Kb

ABSTRACT

This paper addresses information support for hearing-impaired people. Automatic speech recognition, which converts speech to text, is promising support for hearing-impaired people, and studies such include automatic captioning for TV programs or the automatic transcription of oral presentations, lectures, and meetings. These studies mainly focused on how to recognize speech accurately without paying attention how to display the caption texts. The display of caption texts has not been a significant problem because a single speaker usually talks in TV news, oral presentations, or lectures. But, how to display caption texts easily so that who is talking can be understood is important in meetings in which more than one person participates. In TV programs or movies, caption text is just displayed on the bottom side of the screen. The display method, which we call "TV-type caption" in this paper, is inadequate for meetings because it is hard to understand who is talking. Accordingly, we propose a caption display system that shows caption texts with speech balloons near speaker faces based on automatic face detection and speech recognition. In this paper, we evaluate speech balloon captions and compare them with TV-type captions through a questionnaire for appearance, readability of caption text, and comprehension. We confirmed that speech balloon captions are adequate for appearance and comprehension when several speakers exist. TV-type captions are suitable for appearance and readability of caption text when a single speaker talks.

INTRODUCTION

This paper addresses information support for handicapped people in small discussions in which several speakers take part. For supporting auditory handicapped persons, the manual transcription of speech or translation to sign language are typically adopted in schools and universities. However, for such manual support, people with special skills are required. Moreover, even such skilled people have difficultly keeping up with corresponding speech on manual translation to texts or sign language. Manual captioning is too expensive, and thus, automatic speech recognition (ASR) is promising for automatic captioning.

Several automatic captioning studies have been investigated, including captioning for TV programs[1], oral presentations, lectures, and meetings[2]. These conventional studies have mainly focused on how to perform ASR accurately. However, how to display caption texts has not been considered. For lectures or meetings, captions are usually shown on a special screen. Auditory handicapped persons have difficulty grasping at once caption texts and speaker faces, which include moving lips and emotions. From this viewpoint, it is preferable to show both speaker images and caption texts in a single display. Actually, in TV programs or movies, caption texts are shown at the bottom, and such information support is acceptable. In this paper, this display method is called "TV-type caption." However, caption texts are just shown on the bottom of movies, and how to display them has not been well considered. Usually, a single speaker talks on TV news, oral presentations, and lectures, and thus, how to display caption texts has not been a significant problem. On the contrary, for meetings in which several speakers take part, the display of caption texts is crucial to identify who is talking. Showing caption texts on the bottom of movies should be avoided since we cannot easily comprehend the speaker of each utterance. Therefore, more studies for displaying caption texts are needed.

Based on this background, we propose a captioning system that displays captions with speech balloons near speaker faces based on automatic face detection and speech recognition. In this paper, we evaluate this display method, which we call "speech balloon caption," in small meetings. Fig. 1 shows images of TV-type and speech balloon captions.

EVALUATION OF SPEECH BALLOON CAPTIONS

In this section, we compare speech balloon captions with TV-type captions through a survey for appearance, readability of caption text, and comprehension.

Appearances and readability of caption text

First, we evaluate speech balloon captions for appearance and the readability of caption text and investigate the influence of the number of speakers on the caption appearance. Then we investigate the influence of a chairperson's presence.

Experiment

We conducted our experiment as follows. A movie with TV-type caption and a movie with speech balloon caption are displayed to subjects at the same time. Afterwards, subjects answered questions about the movies.

We prepared three movies: 1) prime minister's address, 2) meeting with a chairperson, and 3) meeting without a chairperson. Examples of the movies are shown in Fig. 2. There was one

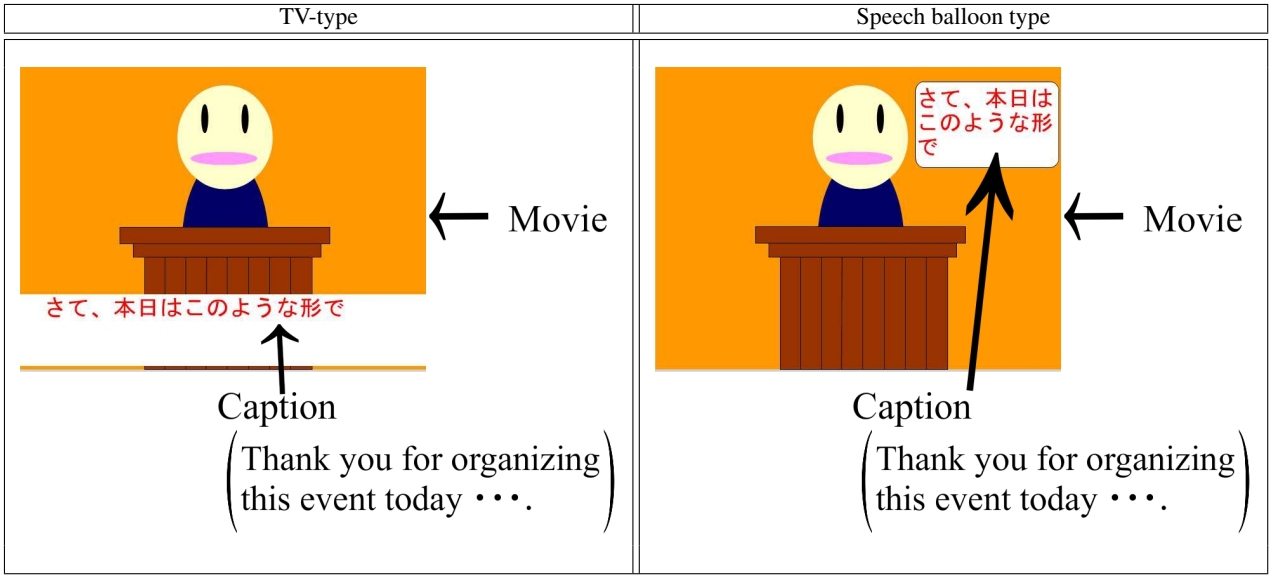

Figure 1: Definitions of TV-type and speech balloon captions

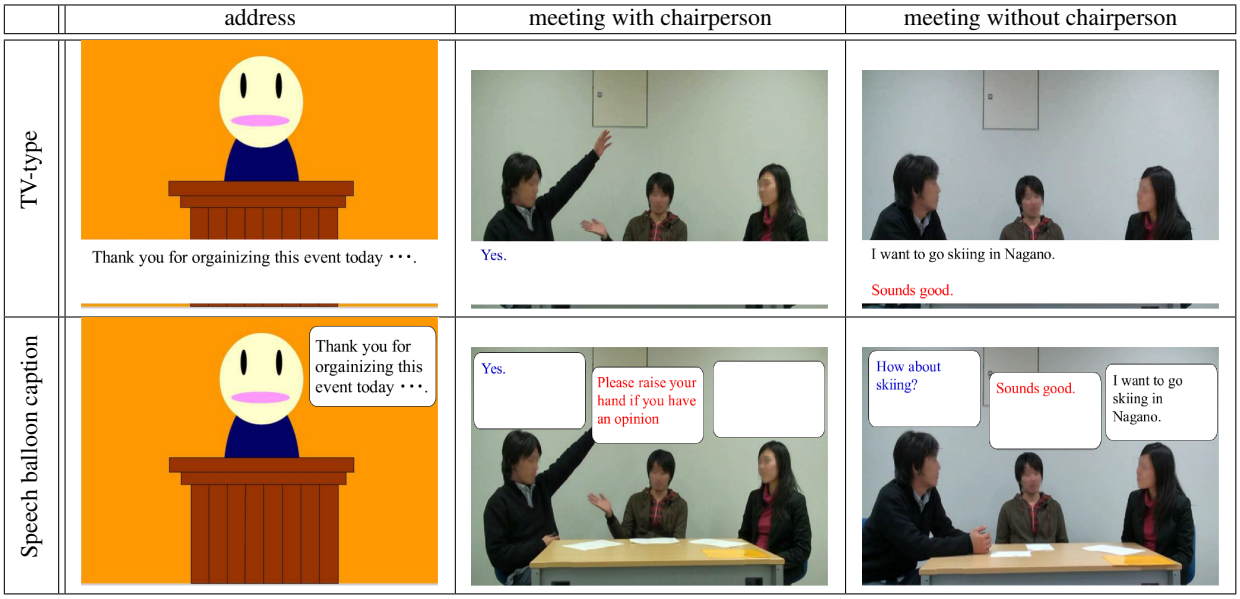

Figure 2: Examples of movies used for questionnaire

Table 1: Questionnaire results of appearance and readability of caption text

	Question topic
Question 1	appropriateness of caption position
Question 2	readability
Question 3	comprehension

Table 2: Answers for each question

	Choices
Answer 1	TV-type caption
Answer 2	speech balloon caption
Answer 3	same
Answer 4	no opinion

speakers in the movie of the address and three in the movies of meetings. For the questionnaire, we addressed three topics: 1) appropriateness of caption position, 2) readability of caption text, and 3) comprehension of caption text. For each question, we prepared four answers: 1) TV-type caption, 2) speech balloon caption, 3) same, and 4) no opinion. The questionnaire is listed in Table 1. Answers are listed in Table 2. In this test, 110 university students participated as subjects.

Questionnaire results

Tables 3, 4, and 5 show the questionnaire results of the caption text for appearance, readability, and comprehension, respectively. 95% confidence intervals are indicated. No influence of the chairperson's presence was confirmed in all the questions.

First, we explain the questionnaire results about the appropriateness of the caption position (Table 3). For the address movie, the approval rating of the TV-type caption was 95%, and the approval rating of the speech balloon caption was 4%. For movies of meetings with chairperson, the approval ratings of the TV-type and speech balloon captions were 25% and 62%, respectively. For the movie of a meeting without a chairperson, the approval rating of the TV-type caption was 30%. The approval rating of the speech balloon caption was 57%. In the appearances of the captions (position), TV-type captions are appropri-

Table 3: Appropriateness of caption position

		TV-type	Speech balloon caption	Same	No opinion
address	Number of people	105	4	0	1
	Approval rating (%)	95	4	0	1
	95% confidence interval of approval rating	92∼99	0∼7	0∼0	0∼3
meeting with chairperson	Number of people	27	68	10	5
	Approval rating (%)	25	62	9	4
	95% confidence interval of approval rating	17∼33	53∼71	4∼14	1∼8
meeting without chairperson	Number of people	33	63	8	6
	Approval rating (%)	30	57	7	6
	95% confidence interval of approval rating	21∼39	48∼67	10∼12	8∼10

Table 4: Readability of caption text

		TV-type	Speech balloon caption	Same	No opinion
address	Number of people	96	11	1	2
	Approval rating (%)	87	10	1	2
	95% confidence interval of approval rating	81∼94	4∼16	0∼3	0∼4
meeting with chairperson	Number of people	63	42	4	1
	Approval rating (%)	57	38	4	1
	95% confidence interval of approval rating	48∼67	29∼47	0∼7	0∼3
meeting without chairperson	Number of people	67	37	3	3
	Approval rating (%)	60	34	3	3
	95% confidence interval of approval rating	52∼70	25∼42	0∼6	0∼6

Table 5: Comprehension of contexts

		TV-type	Speech balloon caption	Same	No opinion
address	Number of people	86	11	10	3
	Approval rating (%)	78	10	9	3
	95% confidence interval of approval rating	70∼86	4∼16	4∼14	0∼6
meeting with chairperson	Number of people	31	64	12	3
	Approval rating (%)	28	58	11	3
	95% confidence interval of approval rating	20∼37	49∼67	5∼17	0∼6
meeting without chairperson	Number of people	39	60	6	5
	Approval rating (%)	35	55	5	5
	95% confidence interval of approval rating	27∼44	45∼64	1∼10	1∼8

ate when the number of speakers in a movie is one, and speech balloon caption is appropriate when the number of speakers in the movie exceeds one.

Next, we explain the questionnaire results about the readability of the caption text (Table 4). For the address movie, the approval rating of the TV-type captions was 87%, and the one of the speech balloon captions was 10%. For the movie of a meeting with a chairperson, the approval ratings of the TV-type and speech balloon captions were 57% and 38%, respectively. For the movie of a meeting without a chairperson, the same tendency was observed. The approval ratings of TV-type and speech balloon captions were 60% and 34%, respectively. For the readability of the caption text, TV-type caption is approved regardless of the number of speakers. One possible reason might be that subjects have many chances to see such TV-type captions. Another explanation might be the problem of displaying character strings in speech balloon captions. Speech balloon captions should display character strings in a narrower space than TV-type captions, and Japanese words that consist of two or more characters should be divided at irrelevant positions for line feeds. This might be a Japanese specific problem since words are not segmented in Japanese text.

Finally, we explain the questionnaire results of the comprehension of caption text (Table 5). For the address movie, the approval rating of the TV-type captions was 78%, and the approval rating of the speech balloon captions was 10%. For the movie of a meeting with a chairperson, the approval ratings of the TV-type and speech balloon captions were 28% and 58%, respectively. For the movie of a meeting without a chairperson, the approval rating of the TV-type caption was 35%. The approval rating of the speech balloon captions was 55%. In the comprehension of the caption texts, TV-type captions are preferable with only one speaker, and speech balloon captions are preferable with more than one speaker.

We investigated speech balloon captions for appearance and readability. When there are several speakers in movies, speech balloon captions are helpful for caption positions. However, they do not improve the readability of caption texts.

Comprehension of caption text

In this section, we evaluate speech balloon captions for comprehension.

Experiment

Our experiment was conducted as follows. Subjects watched a silent movie with caption text and answered questions about its contents. We prepared two identical movies that only differed in how they displayed their caption texts. We prepared

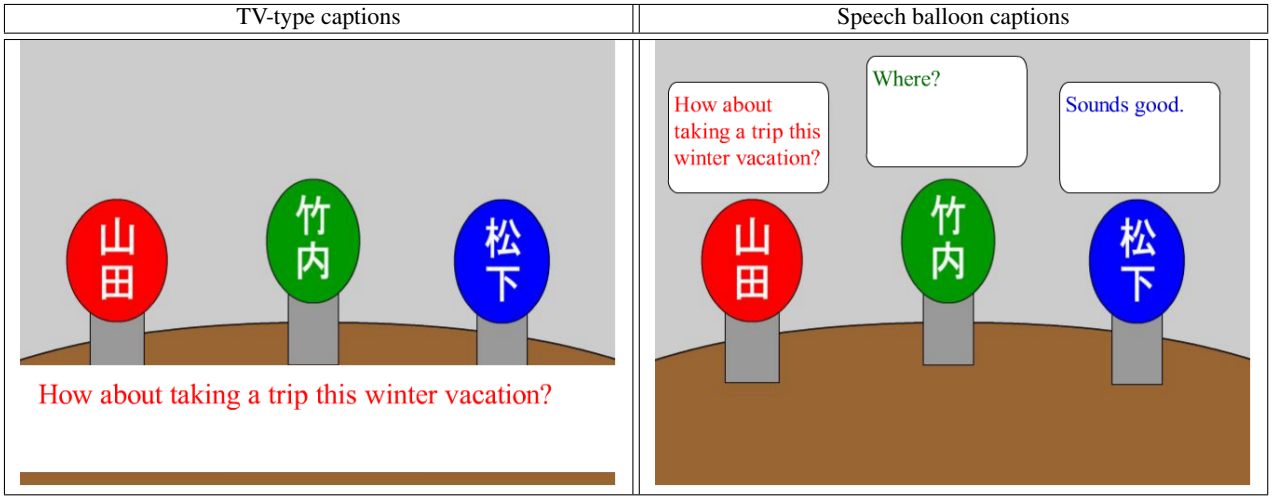

Figure 3: Examples of movies used for comprehension

Table 6: Average question scores about movies

Questions	Average score after watching TV-type captions number of correct answers / correct rate	Average score after watching speech balloon captions number of correct answers / correct rate
About speakers	1.8 / 36%	2.85 / 57%
About events	2.5 / 83%	2.45 / 82%

Table 7: Four text patterns to remove order effect

Pattern	Order
1	speech balloon caption → questionnaire 1 → TV-type caption → questionnaire 2
2	speech balloon caption → questionnaire 2 → TV-type caption → questionnaire 1
3	TV-type caption → questionnaire 1 → speech balloon caption → questionnaire 2
4	TV-type caption → questionnaire 2 → speech balloon caption → questionnaire 1

two question sets about the movie contents. Each caption text has a specific color that corresponds to a particular speaker. Each question set consists of eight questions: five about the person and three about the event. Subjects watched one movie and answered one question set. Afterward, they watched another movie and answered another question set. Examples of questions are “Who recommended the hot spring?”, “Who recommended skiing?”, and “Where is the meeting place?” Examples of the movies are shown in Fig. 3.

Each subject watched two movies. Their discussing about a trip. Duration of the movies is about three minutes. Since they are the same, the number of correct answers for the second question set should be higher than for the first question set. So we prepared four test patterns to remove any order effect. They are listed in Table 7. The subjects were 40 students from 18 to 25 years old. Ten subjects were assigned to each test pattern.

Questionnaire results

Table 6 shows the average number of correct answers after watching each caption type movie.

For the questions about persons, the average correct answers after watching the TV-type captions was 1.8 and 2.85 after watching the speech balloon captions. We performed a paired t test and confirmed that the difference is statistically significant ($P < 0.01$). For the questions about events, the average correct answers after watching the TV-type captions was 2.5 and 2.45 after watching the speech balloon captions. According to a paired t test, the difference is not statistically significant. The effect of the speech balloon captions on comprehension resembles the TV-type captions for event type questions. The speech balloon captions outperformed the TV-type captions for questions about persons.

Figure 4: Image of designed system

DESIGN OF INFORMATION SUPPORT SYSTEM

In previous sections, we confirmed that speech balloon captions are appropriate for appearance and comprehension when several speakers exist. So we designed an information support system that used them. We made an information support system with speech balloon captions on a notebook computer. Fig. 4 shows the system screenshot.

Overview

We designed a speech balloon captioning system that is based on ASR and automatic face detection (Fig. 5). The system consists of ASR and image processing modules. In the ASR

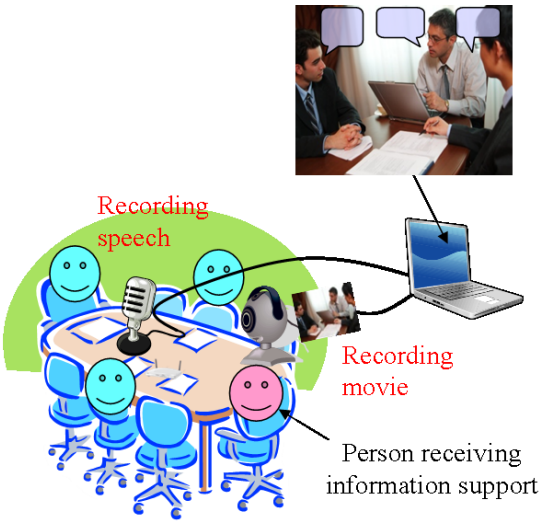

Figure 5: Image of system

module, utterances are recorded with microphones, and then speaker identification and speech recognition are performed. For each speaker, caption text is embedded to the speech balloon image file corresponding to the speaker. In the image processing module, face detection and face identification are performed. Finally, based on the detected face and the speaker's speech recognition result, speech balloon caption image files are embedded to each movie frame. Fig. 6 shows this system configuration. The following sections describe these modules in detail.

Speech recognition module

The ASR module consists of a conventional HMM acoustic model, an N-gram language model, and a decoder Julius[3]. For each component, the most appropriate one should be selected based on a theme that corresponds to the meeting and its participants; the model should be adapted to the topics and the users. The ASR module generates speech balloon image files, as described above. In this system, the maximum number of characters in each line is set to six, and the maximum number of lines in each speech balloon caption is set to three. As for Japanese, length of words is shorter than six unlike european languages. The length is not enough but not so tight. When the recognized text length is longer than 18 (= 6 x 3), a new image file is generated and soon displayed. Our system assumes brief discussions, and all participants are required to wear head-set microphones. Therefore, speaker identification is not difficult: it is just checking a microphone channel ID.

Image processing module

Image processing is designed with OpenCV[4][5]. Movies are captured with a simple web camera, and then face detection is performed. Since our system assumes that the seats of the meeting participants are fixed, speaker identification is performed based on the x-coordinates of the detected faces. The position of the speech balloon caption is set to the upper left of the face considering the face size.

CONCLUSION

We evaluated speech balloon captions with information support for appearance and comprehension. When the number of speakers is one, TV-type captions are appropriate for appearance. When the number of speakers exceeds one, speech balloon captions are appropriate for appearance and comprehension.

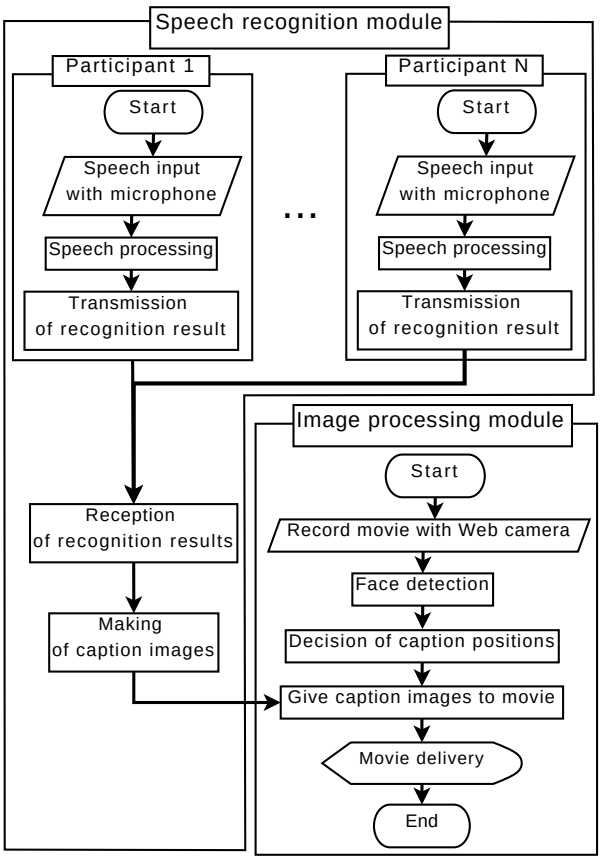

Figure 6: System configuration

As future work, we will make information support system which uses speech balloon caption, with a head mounted display. It is thought that the user takes part in a meeting easily. And we will evaluate the system in supporting hearing-impaired people.

REFERENCES

[1] S. Homma, A. Kobayashi, T. Oku, S. Sato, T. Imai, and T. Takagi, "New real-time closed-captioning system for japanese broadcast news program." in *In Proceedings of the 11th internationnal Conference on Computers Helping People with Special Needs (ICCHP 2008)*, 2008, pp. 651–654.

[2] M. Wald, "Captioning multiple speakers using speech recognition to assist disabled people." in *In Proceedings of the 11th internationnal Conference on Computers Helping People with Special Needs (ICCHP 2008)*, 2008, pp. 617–623.

[3] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine." in *In Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1691–1694.

[4] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Press, October 2008.

[5] http://opencv.willowgarage.com/wiki/.