**Appendix I. System implementation**

Figure 8 depicts the process flow and the main components of this system which we designed and implemented to generate partial and synchronized captions.



Figure 8 System architecture and process flow

(1) The process starts with the alignment module to generate word-level synchronized captions. The system receives a video and its corresponding transcript as input, both taken from TED talks in this study. First, the audio is ripped from the input video and the embodied speech is transcribed automatically by the ASR system Julius v4.3.1 (Lee & Kawahara, 2009). Synchronization is realized by the word-level alignment feature of ASR. For precise alignment, however, a dedicated acoustic-phonetic model is necessary. For example, standard speech recognizers that are trained with a corpus of read speech do not work well for spontaneous speech (e.g., lectures) even in the alignment task. The acoustic and language models used in this system were trained using 780 TED talks (Naptali & Kawahara, 2012). The resultant ASR system outputs highly accurate transcripts with estimated timestamps for tokenized words. Finally, the original transcript and the ASR output are aligned using the force-alignment procedure for precise synchronization.

(2) The feature extraction module processes the caption text thus generated and converts it into a feature vector. This module calculates the speech rate and frequency of each word in the transcript and detects the specific words.

(3) The rule engine in the decision-making module decides inclusion or exclusion of a word in the final caption. In the first stage, a decision about a word is made based on whether it has high speech rate (i.e., above the learner’s ability), it has low-frequency (i.e., beyond the learner’s vocabulary size), or it is an academic term. In the second stage, the system handles special instances such as abbreviations, proper names, numbers, interjections, transparent compounds, and repeated appearances of words. These general features act on each word, either as excitatory or inhibitory. For instance, abbreviations and proper names are always shown while interjections are discarded. The third stage regards the sequence of the words that should be readable and understandable for the learners. The rules handle words after numbers and words after apostrophe –s in this version.

(4) The formatting and display module generates the final PSC using the user display parameters. If a word is selected to appear in the caption, it will be copied intact to the output caption; otherwise, a character mask (here we use dots) replaces every letter of the word to emulate the speech flow (see demonstrations on <http://sap.ist.i.kyoto-u.ac.jp/psc/>).

**Appendix II. Sample of comprehension questions**

The following is a sample of comprehension questions used in our experiment. Please note that the students saw questions one at a time, and could only see the next question after they had answered the previous one without being able to go back. The students were allowed to take notes if they wanted.

1. Who created the candle problem first? (B)
2. Sam Glucksberg
3. Karl Dunker
4. Daniel Pink
5. What is the candle problem about? (C)
6. Melt the candle and adhere it to the wall without using the matches.
7. Melt the candle and attach it to the wall without using any thumbtacks.
8. Adhere the lit candle to the wall so that it will not drip wax onto the table.
9. What was the reward offered in Glucksberg’s first experiment? (A)
10. If you are the fastest to solve the problem, you receive 20$.
11. If you are in the top 25% of the fastest times, you receive 10$.
12. If you are in the top 10% of the fastest times, you receive 5$.
13. What was the result of Glucksberg’s first experiment? (B)
14. The group ‘with’ reward solved the problem more quickly.
15. The group ‘without’ reward solved the problem more quickly.
16. None of the groups could find the solution in a given time
17. According to the speaker, which one is **true**: (A)
18. Business is focusing on wrong motivators and incentives.
19. Business should offer more rewards to improve thinking.
20. Business should include more punishment for better results.
21. The incentivized group beat the other group in the second experiment, because… (C)
22. they were more creative and intelligent.
23. they were given more time and reward.
24. they were given a clear goal plus a reward.
25. The solution to the candle problem can be found by… (B)
26. sticking to functional fixedness.
27. looking for solution on periphery.
28. concentrating on the incentives .
29. What does the speaker say about white workers in the 21st century? (B)
30. They are still doing rule-based tasks.
31. They are less involved in routine tasks.
32. They are doing more left-brain activities.

The following is a sample of the cloze tests. The students could fill in the blanks while listening to the audio.



Answers: Solution – Fixedness – Platform – Experiment

**Appendix III. Questionnaire Data**

A 5-point Likert-scale questionnaire with the scale of 1 (strongly disagree) to 5 (strongly agree) was used to get learner feedback on PSC. Table 1 presents the results.

Table 1 5-point Likert-Scale Survey Results

|  |  |  |
| --- | --- | --- |
| *Statements* | *Mean* | *SD* |
| **S1:** I think PSC is a good idea. | 3.92 | 0.86 |
| **S2:** I think PSC helps me understand. | 3.84 | 1.00 |
| **S3:** I think PSC helps me follow the audio without being distracted. | 3.64 | 0.89 |
| **S4:** I think PSC is better than FC. | 2.84 | 1.00 |
| **S5:** I think PSC is enough to understand. | 3.18 | 1.16 |
| **S6:** I think PSC helped me use my listening skill more. | 3.20 | 1.06 |
| **S7:** I think PSC is better than FC as I can't read all text  | 2.93 | 1.01 |
| **S8:** I think Synchronized Caption is very helpful. | 3.75 | 0.84 |
| **S9:** I think showing "…" instead of hidden words is a good idea.  | 3.57 | 1.02 |
| **S10:** I could find most of words I did not know in PSC. | 3.40 | 0.93 |
| **S11:** I could find most of the words with high speech rate in PSC. | 3.30 | 0.95 |
| **S12:** I think the captions of videos were easy to read.  | 3.15 | 1.28 |

The first two statements explored the attitudes toward our method, and received mostly positive responses and support (72% and 69% respectively of the participants agreed with those statements with 4 or higher on the Likert scale). Items S4 through S7 elicit views on the effectiveness of PSC compared to FC. Data on these items ranged from 2.84 to 3.20 (average 3.03), indicating that the learners are still not sure if PSC can be substituted for FC.

Statements S3 and S8 focused on the word-level synchronization aspect of PSC. Responses to S3 do not indicate that PSC is distractive (only 2% of the participants found it very distractive). The results of S8 show mainly positive feedback on the synchronization feature of PSC (67% selected 4 or higher). Items S10 and S11 investigated views on the partialization aspect of PSC and the selected words. Approximately 68% of the subjects (40 out of 58) chose neutral (score=3) or agree (score=4) and 16% chose strongly agree (score=5) in response to these items. To check PSC’s readability, findings from items S9 and S12 suggest some improvement should be considered in presenting the captions.

**References**

Lee, A. & Kawahara, T. (2009) Recent development of open-source speech recognition engine Julius, *Proceedings of Asia-Pacific Signal and Information Processing Association*, Sapporo, Japan, 131–137.

Naptali, W. & Kawahara, T. (2012) Automatic speech recognition for ted talks, *Proceedings of the 6th Spoken Document Processing Workshop*, Toyohashi, Japan.