Partial and synchronized captioning: A new tool for second language listening development

Maryam Sadat Mirzaei[1], Yuya Akita[2], and Tatsuya Kawahara[3]

Abstract. This study investigates a novel method of captioning, partial and synchronized, as a listening tool for second language (L2) learners. In this method, the term partial and synchronized caption (PSC) pertains to the presence of a selected set of words in a caption where words are synced to their corresponding speech signal, using a state-of-the-art automatic speech recognition (ASR) technology. The system automatically selects words/phrases which are likely to hinder the learner's listening comprehension and discards the rest. To evaluate the system, the performance of 58 Kyoto University students was assessed by a listening comprehension test on TED talks, under three conditions: no caption, full caption and PSC. Analysis of results revealed that while reducing the textual density of captions to less than 30%, PSC realizes comprehension performance as well as the full caption condition. Moreover, it gains higher scores compared to other conditions for a new segment of the same video without any captions. The findings suggest that PSC can be incorporated into CALL systems as an alternative method to enhance L2 listening comprehension.

Keywords: listening comprehension, partial and synchronized caption, word frequency, speech rate, ASR, CALL.

1. Introduction

In recent years authentic audio/visual materials have become more accessible, increasingly used by L2 learners. While these resources provide rich content and

1. Kyoto University; maryam@ar.media.kyoto-u.ac.jp.

2. Kyoto University; yuya@media.kyoto-u.ac.jp.

3. Kyoto University; kawahara@i.kyoto-u.ac.jp.

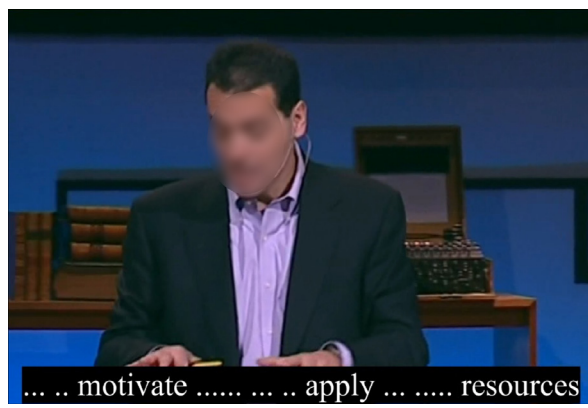
How to cite this article: Mirzaei, M. S., Akita, Y., & Kawahara, T. (2014). Partial and synchronized captioning: A new tool for second language listening development. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*; *Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 230-236). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000223

reflect real-world language, they often entail complex listening comprehension skills (Rogers & Medley, 1988). To facilitate the comprehension of these materials, adding captions is considered as an effective solution. Along with its effectiveness, captioning has received critical attention for bringing too much textual assistance and impeding the development of listening strategies (Pujolà, 2002; Vandergrift, 2004).

The type of captioning may influence the effect of this assistive tool on language learning. Although the conventional full captioning method is still the mainstream of contemporary education, other methods such as keyword/paraphrase captioning have drawn some attention (Garza, 1991). Moreover, the advances of the ASR technology have enabled the generation of synchronized captions. Unlike the typical captions where chunks of words appear on the screen, in synchronized captions the emergence of words on the screen is concurrent to the speaker's utterance. This method fosters word recognition, but promotes word-by-word decoding known as a hindering strategy.

With the purpose of providing adequate textual assistance to L2 listeners while encouraging them to use their listening skills, this study introduces a new method of captioning, partial and synchronized captioning (Figure 1).

Figure 1. Screenshot of PSC on a TED talk made from the original transcript "how we motivate people how we apply our human resources"

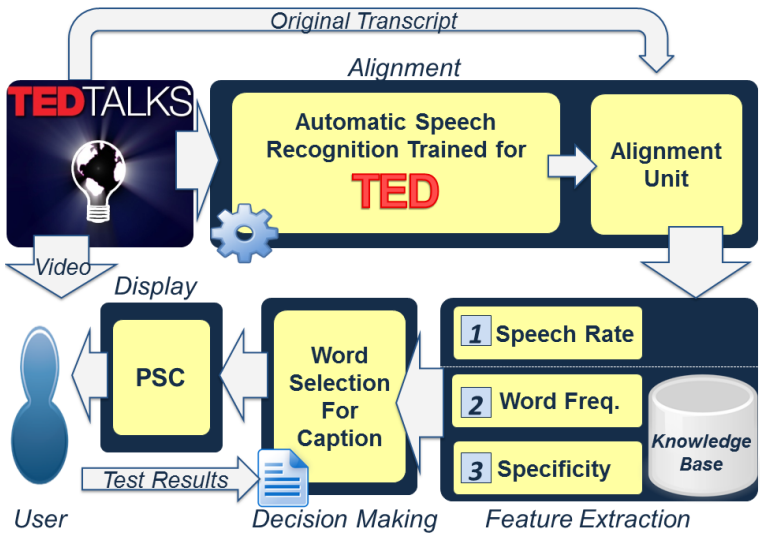


In this method, the original transcript is automatically reduced to a partial caption, which includes only a selected set of words/phrases (partialization). Particular to this method, each word in the caption is synchronized to its respective speech signal (synchronization).

2. Partial and synchronized caption

PSC focuses on assisting the listeners to cope with aural input difficulties without constantly referring to the verbatim caption. To this end, a system was developed based on two main modules: synchronization and partialization (Figure 2). These two are complementary and their integration counteracts the demerits of one another. For instance, synchronized captioning neatly presents the word boundaries, but it is criticized for promoting dependence on caption and encouraging word-by-word decoding. Partial captioning builds on the synchronized caption to avoid over-reliance on reading, however, the irregular and salient appearance of the words in this caption is only handled by the alignment feature of synchronized captioning (Table 1).

Figure 2. Data flow and main modules of the system



2.1. Synchronized caption

Selecting TED talks as the medium of this study, an ASR system was trained by the TED corpus and employed to make alignment. This provides accurate alignment in word level, which in turn enables text-to-speech mapping and fosters word recognition. Synchronized captioning, although in favor of many language learners, has several disadvantages that are alleviated in partial captioning.

2.2. Partial caption

In the partial captioning process, the system selects a subset of words that are likely to be incomprehensible for L2 listeners. This type of captioning attempts to actively mediate the comprehension by bringing a sort of scaffold. As a result, learners' current level of competence should be taken into account when preparing captions. A prudent choice to define credible criteria for selecting the target words is to consider major obstacles of listening comprehension as follows.

2.2.1. Speech rate

High speech rate can negatively affect listeners' comprehension of both native and non-native speakers (Griffiths, 1992; Wingfield, 2000). The proposed method precisely calculates the speech rate of words in syllables per second and represents words/phrases uttered faster than the normal rate of speech or the tolerable rate for the learner.

2.2.2. Word frequency

When listening to an audio, unfamiliar words often confine listeners' attention and impede comprehension. To address this issue, the proposed method selects the difficult words and presents them on screen while masking the rest. The frequency of words in written/spoken corpora is a reliable measure to assess word difficulty. The study measured the frequency of each word using the word family lists based on British National Corpus (Nation & Webb, 2011) and Corpus of Contemporary American English (Davies, 2008-). We also handled instances such as academic words (Coxhead, 2000), proper names and interjections.

Table 1. Comparison of caption methods

Caption Type / Advantage	Full Caption	Keyword Caption	Proposed Partial Caption	Synchronized Caption	PSC
Aid word boundary detection	✓			✓	✓
Speech-to-text mapping				✓	✓
Avoid over-reliance on reading		✓	✓		✓
Avoid being distractive	✓			✓	✓
Automatic	✓		✓	✓	✓
Adjustable to learners' knowledge			✓		✓
Adjustable to the content		✓	✓		✓

3. Evaluation

Given the novelty of the method, the following questions have been investigated to evaluate the system:

- Do captioned videos result in better comprehension compared to non-captioned ones?
- Can PSC substitute the conventional full-text captioning?
- Does PSC help the learner comprehend the video later without any captions?

The participants of this study were 58 Japanese students enrolled in CALL courses at Kyoto University, ranging from 19 to 22 years old. Videos of American speakers were selected from the TED website (www.TED.com) and trimmed to approximately 5-minute meaningful segments. The Vocabulary Size Test (Nation & Beglar, 2007) was used to evaluate the participants' vocabulary reservoir.

The students were grouped into three proficiency levels based on their TOEIC or CASEC scores: beginners, pre-intermediates and intermediates. In each group, the learners' vocabulary size and their tolerable rate of speech were evaluated to generate PSC. Thus, for each group a particular PSC was generated with a different percentage of words to be shown, ranged from 20% to 30% of the original transcript. The students watched the videos under three conditions: no caption (NC), full caption (FC) and PSC. The experiment had two parts: first, the students watched 70% of the video under one of the above conditions and took a listening test; next, the subjects watched the rest of the same video (30%) without caption (presuming a real-world situation) and took another test.

4. Results and discussion

The result of one-way ANOVA test on the first part (70%) revealed a significant difference between NC (M=35.7, SD=14.7) condition and PSC (M=52.9, SD=19.4) or FC condition (M=54.2, SD=17.3) at p<.05. This answers the first research question by showing that the students' scores on PSC condition are significantly higher than NC condition. However, no significant difference was found between the score on PSC and FC condition in this part [$F(1, 57)$=25, p=.62]. The findings provide the answer to the second research question and suggest that PSC leads to the same level of comprehension as FC while providing less than 30% of the transcript.

In the second part of the experiment (30% without caption), the best scores were gained when the learners first watched videos with PSC [$F(2,118)=20.5$, $p<.05$] compared to other conditions. The results provide a positive answer to the third research question and suggest the effectiveness of PSC on preparing the learners for real-world situations. Although this is a short-term enhancement partly because of adaptation to the video, this finding is still valuable.

A Likert-scale questionnaire reflected positive learner feedback on PSC. However, learners were skeptical about substituting FC by PSC.

5. Conclusion

The study introduced a smart type of captions that allows the use of limited textual clues and promotes listening to the audio in order to comprehend the material. The findings highlighted the positive effect of this method in enhancing listening comprehension by presenting less than 30% of the text. Given the nature of listening skills, however, a long-term experiment is required to evaluate the overall listening improvement of the L2 learners.

Acknowledgements. We would like to thank Mark Peterson for his inspiration and brilliant comments and Kourosh Meshgi for his invaluable assistance.

References

Coxhead, A. (2000). A new academic word list. *TESOL quarterly, 34*(2), 213-238. doi:10.2307/3587951

Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Retrieved from http://corpus.byu.edu/coca/

Garza, T. J. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals, 24*(3), 239-258. doi:10.1111/j.1944-9720.1991.tb00469.x

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly, 26*(2), 385-390. doi:10.2307/3587015

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Pujolà, J. T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL, 14*(2), 235-262. doi:10.1017/S0958344002000423

Rogers, C. V., & Medley, F. W. (1988). Language with a purpose: Using authentic materials in the foreign language classroom. *Foreign Language Annals, 21*(5), 467-478. doi:10.1111/j.1944-9720.1988.tb01098.x

Vandergrift, L. (2004). 1. Listening to learn or learning to listen? *Annual Review of Applied Linguistics, 24*, 3-25. doi:10.1017/S0267190504000017

Wingfield, A. (2000). Speech perception and the comprehension of spoken language in adult aging. In D. Park & N. Schwarz (Eds), *Cognitive Aging: A Primer* (pp.175–195). Philadelphia, PA: Psychology Press.