Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Partial and Synchronized Caption Generation to Develop Second Language Listening Skill

Maryam Sadat MIRZAEI [a*], Yuya AKITA [a], Tatsuya KAWAHARA [a]
[a] *Graduate School of Informatics, Kyoto University, Japan*
*maryam@ar.media.kyoto-u.ac.jp

Abstract: Captioning is widely used by second language learners as an assistive tool for listening. However, the use of captions often leads to word-by-word decoding and over-reliance on reading skill rather than improving listening skill. With the purpose of encouraging the learners to listen to the audio instead of merely reading the text, the study introduces a novel technique of captioning, partial and synchronized, as an alternative listening tool for language learners. Using TED talks as a medium for training listening skill, the system employs the ASR technology to synchronize the text to the speech. Then, the system uses the learner's proficiency level to generate partial captions based on three features that impair comprehension: speech rate, word frequency and specificity. To evaluate the system, the performance of Kyoto University students in two CALL classes was assessed by a listening comprehension test on TED talks under three conditions: no caption, full caption and the partial-and-synchronized caption. Results revealed that while reducing the textual density of captions to less than 30%, the proposed method realizes comprehension performance as well as full caption condition. Besides, it performs better than other conditions on new segments of the video without captions.

Keywords: Computer-Assisted Language Learning, Automatic Speech Recognition, Listening Comprehension, Word Frequency, Speech Rate

1. Introduction

The process of learning a foreign language involves mastering different skills such as listening, speaking, reading and writing. Of these, acquiring listening often entails a complex cognitive process and demands the use of different strategies which in turn make a phase of frustration for many language learners (Leveridge and Yang, 2013). In order to improve listening, one must be exposed to authentic and comprehensible input. Authentic input, however, makes listening more challenging especially when the phonological systems of the first and the second language are distant (e.g. Japanese vs. English).

Listeners can overcome this problem by benefiting from assistive tools such as "captioning" that textualizes the verbatim speech and makes it more recognizable through neatly dividing the word boundaries. Nevertheless, when it comes to using captions, both language learners and teachers face a dilemma. In fact, when reading captions is part of watching a video, learners often rely on their reading skill to compensate for their listening skill deficiencies, whereas in a real-world communication, learners should solely use their listening skill as no assistive tools are available.

To address these issues, this study proposes a new method of captioning, "partial and synchronized" as an alternative tool for enhancing second language (L2) learners' listening comprehension skills. The term "synchronized" captioning is to present caption text word by word aligned in precise timing with the speech signal of the respective words, which effectively shows the correspondence between words and the audio channel. This method is realized by the automatic word-level alignment feature of automatic speech recognition (ASR) technology, which precisely maps each word to its corresponding speech signal. In the "partial" captioning method we select a subset of words from the transcript and present them in the caption while hiding the rest of the words. Although seems similar to keyword captioning, in this method "important" words are not the selection criteria. Instead, words that impair comprehension or the ones beyond the learner's current level of competence form the basis of this selection. Moreover, the selection of keywords is content-specific

and does not consider the proficiency level of the learners, whereas the features of the proposed method are tuned to the learner's knowledge to meet the requirements of each individual.

Unlike conventional captions, in Partial and Synchronized Captioning, comprehension cannot be gained by solely reading the captions, but by listening to the audio and reading only for difficult or unrecognizable words. Thus the method is effective for reducing learners' dependence on captions.

Following this introduction, this paper reviews previous studies and describes the proposed technique of captioning. Then, the experimental procedure together with the results is demonstrated and a discussion over the findings is addressed. This paper ends with conclusion and future directions.

2. Literature Review

2.1 Captioning and L2 Listening Comprehension

To overcome the listening problems, assistive materials, such as captions, are used to help L2 listeners. Captioning is defined as "visual text delivered via multimedia that matches the target language auditory signal verbatim" (Leveridge and Yang, 2013, p.1). Captions neatly demonstrate the word boundaries without being affected by accent, pronunciation and audio deficiencies (Vanderplank, 1993) and allow the learners to parse the speech stream into meaningful chunks, an essential process for learning (Ellis, 2003). A considerable amount of literature has been published on various beneficial effects of captions. Some of these studies have investigated the effect of captioning on vocabulary acquisition (Bird and Williams, 2002; Griffin and Dumestre, 1992), reading development (Bean and Wilson, 1989), word recognition (Bird and Williams, 2002; Markham, 1999) and listening comprehension (Danan, 2004; Garza, 1991; Markham, 1999; Montero Perez et al., 2014; Vanderplank, 1993; Winke et al., 2010).

For instance, Garza (1991) conducted an experiment with 70 high-intermediate learners of English and 40 three to four year learners of Russian, and assessed their comprehension of videos with/without captions. His results indicated significant improvement on the captioning condition in both groups. Studies in Japan such as Suzuki (1996) reported the positive effect of English caption on Japanese listening comprehension development.

The type and manner of captioning may influence the effect of this assistive tool on language learning. Garza (1991) suggests using various types of open captioning, such as verbatim, paraphrase and keywords as means of training listening skill.

2.2 Aligned and Synchronized Captioning

Correspondence between caption and speech may also affect the learning process. Advancement of speech technology has enabled precise text-to-speech alignment. Munteanu et al. (2007) used ASR to generate transcripts of webcast lectures for examining native speakers' comprehension on the videos. They found out that ASR generated transcripts are useful when word error rate (WER) is lower than 20%. This finding was generalized to L2 learners in a study by Shimogori et al. (2010) who suggest that captions with 80% accuracy improve the understanding of Japanese learners of English.

Accordingly, "karaoke-style" display, where the text is highlighted in colors as the audio moves by, has gained some instructional value. Bailly and Barbour (2011) developed a system that exploits the alignment of text with audio at various levels (letters, phones, syllables, etc.). This system uses a data driven phonetizer trained on an aligned lexicon of 200,000 French entries to display a time-aligned text with speech at phoneme level. The results showed that the multimodality of synchronous reading systems is beneficial for overcoming the problem of word decoding in a text/audio-only environment.

It should be noted that this method may lead to over-reliance on the caption and needs to be refined. This can be accomplished through highlighting only particular words or sentences in the caption, as in keyword captioning.

2.3 Keyword Captioning

Guillory (1998, p.95) examined the use of keyword captioning for learners of French. The results demonstrated that students who received keyword captions performed as well as those who received full captions. Guillory discussed that "learners no longer need to be subjected to a volume of text to read; they can in fact comprehend authentic video with considerably less pedagogical support".

In a recent study by Montero Perez et al. (2014), the perceived effectiveness of keyword captioning is criticized. The study investigated the effect of full text captions and keyword captions versus no captioned condition. The results demonstrated that full captioning group outperformed the other two groups on the global comprehension questions while both the keyword captioning and the no-captioning group had equal performance on this test. Analysis of the responses received from the keyword-captioning group revealed that this type of captioning is distractive. According to the researchers, a plausible explanation may be the salient and irregular appearance of the keywords on the screen, which causes distraction. However, not every learner can benefit from presenting the keywords in captions since the selection of keywords is content-specific and may not provide each learner with his/her required amount of support. In line with this assumption, Guillory (1998) noted that the keyword captions used for her study contained a tiny portion of the total script, which may not have provided enough information for the beginners.

2.4 Limitations on Captioning

In spite of the beneficial aspects of captioning, there are some criticisms on the use of this assistive tool. It is skeptical whether learners provided with captions are training their listening or their reading skills. Kikuchi (1995) examining subtitles in Japanese and captions in English reported that students who watched the movie with Japanese subtitles merely read the text without listening to the movie. Using an eye tracker, Winke et al. (2013) investigated learners' use of captions and reported that learners read the captions on average 68% of the time it is on the screen.

On one hand, the learner needs to be able to deal with real-world situation where there is no access to any supportive tool, and on the other hand we cannot expect a non-native listener to follow the authentic input without any support. Hence, the listening instruction should focus first and foremost on assisting the language learners to cope with aural input difficulties while maintaining a tendency to develop compensatory strategies for listening in real-time. Thus, further research should be conducted to investigate an effective method for assisting learners to gain adequate comprehension, without becoming too much dependent on captions.

3. Proposed Method: Partial and Synchronized Captioning

We propose a new type of captioning called Partial and Synchronized Captioning (hereinafter, PSC). In this method the text is synchronized to the speech in word-level and only a subset of words are shown in the caption while the rest are masked to keep the learner listening to the speech. Thus, this method consists of two components; synchronization and partialization where the two are complementary and counteract the demerits of one another.

First, synchronized caption is automatically generated; word-level synchronization of text with speech is realized by ASR. The word-level alignment, which synchs each word with the speaker's utterance, presents the phonological visualization of the words and thus leads to improvement in aural word recognition skills through mapping between the speech stream and the verbatim text.

Moreover, this method neatly presents word boundaries, which often cannot be easily recognized in authentic speech input. Synchronized captions, although in favor of many language learners, may bring too much assistance for the learner and makes them more and more dependent on the caption (Vandergrift, 2004; Garza, 1991). In order to solve the disadvantages of this method, we propose partial captioning which builds on synchronized captions to provide the students with reduced transcription of the videos in order to better train them for real-world situations.

This method can act as an intermediary stage before the learner is totally independent of captions. In this method, the filtering process of words to be presented takes into account not only the hindering factors of comprehension, but also the assessed knowledge of the learner. Hence, adjusted

to a particular learner's need, the method selects words which are beyond the proficiency level of the learner. However, if using partial captions alone, as in keyword captioning, the students are often distracted by the sudden and irregular appearance of a word on the screen (Montero Perez et al., 2014). Nevertheless, this problem is mitigated by synchronization in PSC.

To conclude, this new tool, PSC, is anticipated to make the learner less dependent on caption and more prepared to handle listening in real-world situations. Table 1 summarizes the advantages of PSC compared to other captioning methods and Figure 1 shows the screenshot of a generated PSC.

Table 1: Comparison of different caption methods.

Advantage \ Caption Type	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		✓	✓		✓

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

3.1 Feature Selection

In order to decide which words to show in the caption and which ones to hide, the following features were picked as the selection criteria. These features were chosen for being identified as major contributing factors in listening comprehension impair. Besides, these factors can be quantified automatically and are easy to be implemented.

3.1.1 Speech Rate

Previous studies showed that high speech rate can negatively affect L2 listeners' comprehension (Dunkel, 1994) and this is even true for native speakers (Wingfield et al., 1985). For Japanese learners of English, particularly, fast rates of speech and inability to perceive the sounds in English are the major factors to impair comprehension (Osuka, 2007). Some studies suggested modification of speech rate as a solution, however, this is not close to real-world situation. Instead, we provide the learner with PSC that presents words/phrases uttered faster than normal rate of speech, or that of tolerable for the learner.

3.1.2 Word Frequency

When the vocabulary chosen by the speaker exceeds the vocabulary size of the listener, comprehension will be impeded. In such cases the unknown words confine the learner's attention, and as the speech proceeds the learner cannot pursue the subsequent parts. In other words, the listener invests a lot of time trying to understand what s/he missed (Goh, 2000).

The frequency of word usage in a language is a measure to assess word difficulty. For instance, learners are less likely to be familiar with low-frequency words (Nissan, 1996). Word frequency is calculated based on its occurrence in spoken or written corpora. A well-cited paper by Nation (2006) categorizes English vocabulary into High-frequency (the most frequent 2000-3000 word families), Mid-frequency (anything between 3000-9000 word families), and Low-frequency (beyond the 9000 frequency band). The term word family here refers to a base word and all its derived and inflected forms that can be recognized by a learner without having to learn each form separately.

To assist L2 listeners, PSC presents words or phrases, which are less frequent and hence make comprehension difficult.

3.1.3 Word Specificity

The occurrence of specific words in a video would make comprehension difficult since limited knowledge of academic words is often seen as a reason for L2 listening comprehension deficiency (Goh, 2000). Thus, when considering word frequency, it is important to consider word specificity as well. Using academic talks as the material for this study, this feature is also taken into account in PSC.

4. System Architecture

Figure 2 depicts the data flow and main components of the system. The procedure of generating a PSC starts with an alignment phase where the ASR system outputs the transcript with estimated word timing, which is aligned and adjusted with the given transcript of the caption. Next, word frequency, word specificity and speech rate are used to serve as the selection criteria for making PSC. The feature extraction module further processes the transcript and converts it into a feature vector for the decision making module.

A rule engine in the decision making module decides whether a word should be shown or not. This decision not only depends on the features, but also relies on the user input (i.e. quiz results).

In the formatting and display module, the captions are altered as the desired output of the system. Being synchronized with the utterance of the word, the corresponding dictation of the word (or character mask) should appear on the screen. Eventually this module plays back the media with the generated caption, and offers a pre-made comprehension test afterwards.

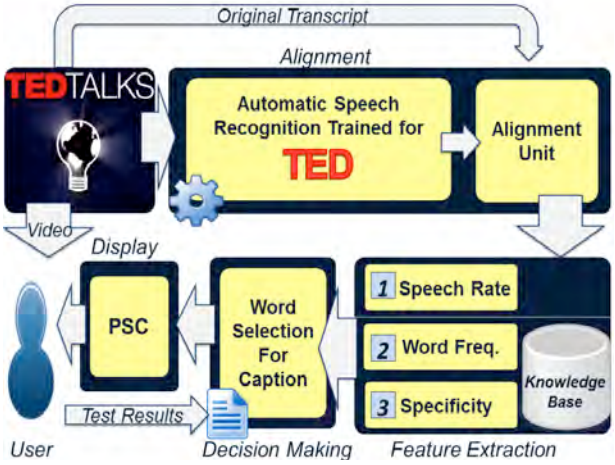


Figure 2. Data flow and the main components of the system.

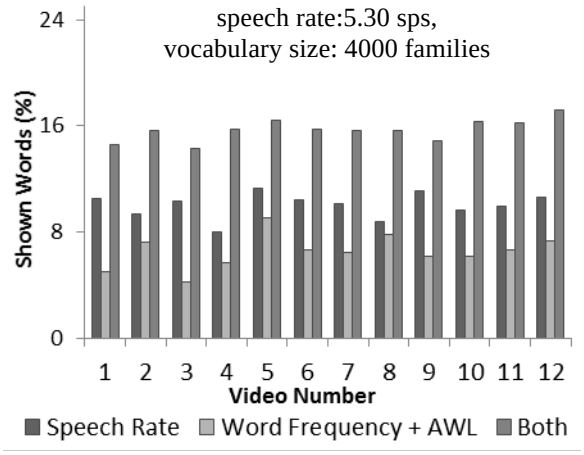


Figure 3. Percentage of words shown in PSC for intermediate learners

4.1 Alignment Module

The input data is composed of a video and its transcript text. To obtain the time tag of the tokenized words automatically, the audio should be ripped from the video to be passed to our ASR system, Julius v4.3.1 (Lee and Kawahara, 2009). Since Julius itself is a language-independent decoding program, it is possible to make a recognizer of a language, given an appropriate language model and acoustic model for the target language. The performance of ASR largely depends on these models. In this study TED talks were selected as the material. Thus, for precise alignment to take place, it is necessary to train the ASR models using a matched corpus, in this case TED talks. This model training was done in our laboratory, based on the lightly-supervised training approach using 780 TED talks (Naptali and Kawahara, 2012). The transcript and ASR output then got aligned using the force-alignment procedure.

4.2 Feature Extraction Module

This module extracts the main features and calculates them. The following elaborates on these features.

4.2.1 Speech Rate

The speech rate is often measured in Words per Minute (WPM) or Syllables per Second (SPS). The former may be affected by pauses and change of speech rate within a minute which causes inaccurate measurement while the latter is more suitable to measure short speeches and thus is used in this study.

The first step to calculate this feature is to estimate the speech rate where we need to count the number of syllables in each word, and then calculate the duration of its utterance. Calculation of the syllables is based on the structural syllabification of the corresponding text, which was realized using Natural Language Toolkit (NLTK). The full calculation of speech rate requires the duration of a word, which is calculated by the time tags obtained in the alignment phase after excluding the long pauses.

4.2.2 Word Frequency

Word frequency is defined by referring to corpus-based studies. Nation (2006) has designed 25 word family lists each including 1000 word families, plus four additional lists: (i) an ever-growing list of proper names; (ii) a list of marginal words including swear words and exclamations; (iii) a list of transparent compounds; and (iv) a list of abbreviations. The first two lists are carefully hand-selected while the rest are based on the following two famous corpora.

- The British National Corpus (BNC) which involves 100 million word collections of samples of written and spoken language from British English.
- Corpus of Contemporary American English (COCA), gathered by Mark Davies (from 1990 to 2012), includes 450+ million words. The corpus is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

This study is based on aforementioned word family lists and COCA. Every word is lemmatized first, and the result is looked up for the word family, created offline from the COCA and BNC corpus. The family of the lemmitized word serves as the difficulty index. The word is also cross-checked with the spoken genre section of COCA.

4.2.3 Word Specificity

In this method specific words are determined using a popular catalogue called Academic Word List (AWL) by Coxhead (2000) which includes 570 headwords and about 3000 of academic words altogether. Besides, these words are cross-referenced with COCA's academic words (Gardner and Davies, 2013) for more accuracy. The system is also capable of handling other features such as abbreviations, proper names, numbers, transparent compounds, and repeated appearance of words.

4.3 Decision Making Module

Based on the features, the system decides whether a word should be included in the final partial caption or not. This decision not only relies on the value of the features, but also considers general features.

In the first stage, the main features - word frequency, speech rate, and specificity - are accounted. If only one of them require a word to be shown, the word is marked to appear in caption. To decide on the word frequency feature, a vocabulary size test (Nation and Beglar, 2007) is employed to assess the vocabulary size of the learner and to determine the appropriate frequency threshold for him/her. Similarly, a decision about whether a word should be a candidate for being shown in partial caption is taken by comparing the calculated speech rate of the word to that of preferable for the learner. Thus, if the utterance of the word (measured by speech rate feature) is faster

than the tolerable threshold of the learner, the word will be shown in caption as a textual clue. This threshold can be adjusted by the user.

In the second stage, the general features act on each word. The features are either excitatory or inhibitory. The decision on general features is made on top of the first stage. For instance, abbreviations and proper names are being marked to be displayed while interjections are marked to be discarded.

The third stage of decision-making is about the sequence of the words that should be readable for the learners. The rules also handle words after numbers and words after "apostrophe s".

4.4 Formatting and Display Module

This module generates the final partial and synchronized caption using the user display parameters. If the word is decided to be shown, it will be copied intact in the partial caption; otherwise a character mask (here we use "dots") replaces every letter of the word. This will emulate the speech flow, by showing each and every word in the given speech in synch with their utterance. (e.g. "express" will be replaced by "....." and "don't" will be replaced by "....").

5. Experiment

Given the novelty of partial and synchronized captioning method, experiments were needed to evaluate the effectiveness of this technique. Thus, the study investigates the following questions:

- Do captioned videos result in better comprehension of video compared to non-captioned ones?
- Can the proposed captioning method substitute the conventional full-text captioning?
- Do proficiency differences affect the benefits obtained from the proposed captioning method?
- Does the proposed method help the learner comprehend the video better without any captions?

5.1 Participants

The participants were 28 and 30 Japanese students at Kyoto University ranging from 19 to 22 years old. These students were undergraduates of different majors who enrolled at a CALL course. The experiments were carried out over this course, in two different classes, for three consecutive sessions.

5.2 Material

Videos: The video materials of this research were selected from TED website which provides us with authentic videos plus almost accurate captions without the copyright issue (www.TED.com). The selection criteria were bound to "popularity" and "recentness" of the videos. The selection was carefully done to include only videos of native American speakers, to avoid the influence of other accents. All videos were trimmed to 5-mintue meaningful segments.

Pre-study Vocabulary Size Test: A vocabulary size test created by Nation (2007) was used to evaluate the vocabulary reservoir of each student. The results of this test were used both as a placement criteria of dividing students into groups of proficiency and as a value to determine the frequency threshold for our caption generator. This test consists of 140 multiple-choice questions, with 10 items from each 1000 word family level. Since the caption generator uses the same word family lists as its references, the result of the test is appropriate to be set as our threshold.

Partial and Synchronized Caption Statistics: Taking into account the result of the vocabulary size test and the tolerable rate of speech, the system generates appropriate captions for learners with different proficiency levels. The percentage of words to be shown in the final caption does not exceed 30% for any of the videos as illustrated in Figure 3. This figure presents how the generated captions show fairly equal amount of words per video for a particular intermediate learner.

Comprehension Tests: After watching each video with assigned caption, the students were asked to take a listening comprehension test in the form of multiple choice and cloze test on summary.

5.3 Procedure

The study was conducted in CALL classes where students were provided with a 20 inch-wide screen and a headphone. Although the experiment was held in two different classes, the same procedure was adopted for both. Same videos were captioned with a different method (PSC↔FC) for each class.

We considered learner's proficiency as a blocking factor, with three levels: "beginner", "pre-intermediate" and "intermediate". For the purpose of dividing the students into these three groups, the assessed vocabulary size together with the students' TOEIC/CASEC scores were considered.

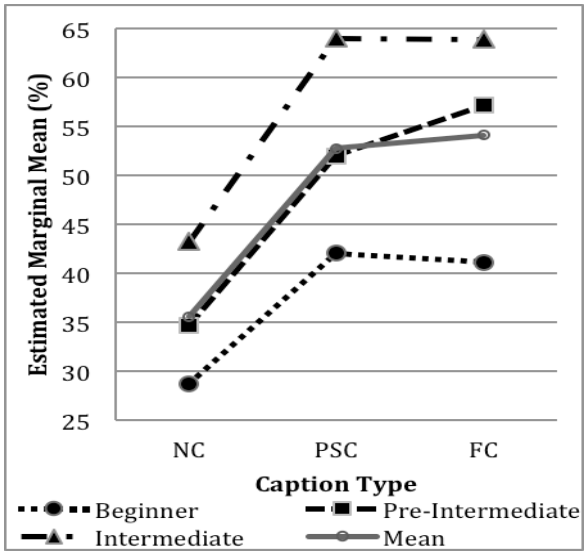
Each video, regardless of the caption type assigned to that was divided into two segments; 70% from the beginning and the rest of 30%. The students watched the first part of the video (70%) under one of these three conditions: no-caption (NC), full-caption (FC) and partial and synchronized caption (PSC). This was followed by a listening comprehension test. Next, the subjects were asked to watch the rest of the same video (30%) "without any caption" (regardless of the type of caption in the previous phase), and took another test. The procedure remained the same for all videos, while the type of caption was changed. To be more specific, the second part of each video is dedicated to evaluate students' performance on a non-captioned video as in real-world condition.

6. Results and Discussion

The scoring system was easily constructed because of the objective format of multiple-choice and cloze-on summary items. One point was awarded for each correct answer to multiple-choice questions while partial credit (0.25) was given to each item in cloze test. The total score was finally calculated in percentage for all participants in each group. One-way ANOVA test was used to analyze the result of the tests and to investigate whether any statistically significant difference is found between different conditions under which the learners watched the videos.

Table 2: Comprehension performance of both classes on the first part of video with (NC, PSC, FC)

Proficiency Level		Mean	SD	N
NC	Beg.	28.7	13.6	19
	Pre. Int.	34.7	11.8	19
	Int.	43.3	15.1	20
	Total	35.7	14.7	58
PSC	Beg.	42.0	16.7	19
	Pre. Int.	52.0	17.5	19
	Int.	64.0	18.0	20
	Total	52.9	19.4	58
FC	Beg.	41.1	12.3	19
	Pre. Int.	57.2	14.8	19
	Int.	63.9	16.4	20
	Total	54.2	17.3	58



*** NC:** No Caption **PSC:** Partial and Synchronized Caption **FC:** Full Caption

As shown in Table 2, analysis of the first part of the experiment (watching 70% of the videos) revealed a significant difference between NC (M =35.7, SD =14.7) condition and PSC (M =52.9, SD =19.4) and also FC condition (M =54.2, SD = 17.3) at the $p < .05$. The results provide a positive answer to our first research questions, which concerns the effectiveness of PSC as compared to NC. However, no significant difference was found between the scores gained under PSC and FC conditions in this part of the experiment [F (1, 57) = 25, p =.62]. The findings suggest that PSC leads to the same level of comprehension as FC while providing less than 30% of the transcript.

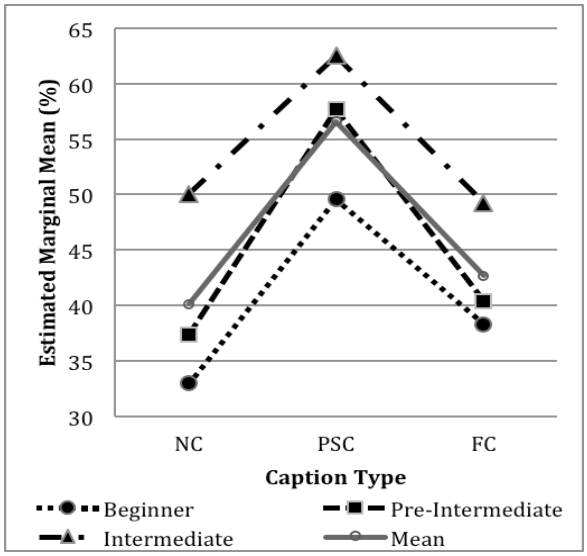
Consequently, the assumption of our second research question is plausible and hence PSC can be used as an alternative to full captioning method for training L2 listening. Furthermore, the results reveal that students with different proficiency levels gained almost equal scores under PSC and FC conditions and could benefit from our method. A possible explanation for deriving such results may lie in the adaptability of PSC that considers the proficiency level of the learners for generating appropriate amount of caption for them and provides adequate assistance for any learner.

Table 3 presents the results of comprehension tests on the second part of the experiment where students watched the rest of videos without any captions immediately after having watched the first parts under different conditions (NC, FC, PSC).

In the second part of the experiment (30% without caption), the best performance is associated with the condition in which the learners first watched the video with PSC [F (2,118) = 20.5, $p < .05$] as compared to FC and NC. The findings highlight the effectiveness of PSC on preparing the learner for real-world situation where captioning is not available. While this result indicate a short-term enhancement partly because of adaptation to the video, this finding is still of value.

Table 3: Comprehension performance of both classes on the second part of video without caption

Proficiency Level		Mean	SD	N
NC	Beg.	33.0	16.0	19
	Pre. Int.	37.4	16.6	19
	Int.	50.0	15.6	20
	Total	40.1	17.4	58
PSC	Beg.	49.6	15.8	19
	Pre. Int.	57.7	17.2	19
	Int.	62.5	17.4	20
	Total	56.6	17.3	58
FC	Beg.	38.3	13.5	19
	Pre. Int.	40.4	11.9	19
	Int.	49.3	12.7	20
	Total	42.7	13.4	58



*** NC:** No Caption **PSC:** Partial and Synchronized Caption **FC:** Full Caption

7. Conclusion and Future Work

The study introduced a novel technique of captioning, partial and synchronized, which is based upon speech rate, word frequency and specificity, to generate a smart type of caption that deals with limitation of previous methods. This method is based on the premise that the presence of infrequent or specific words and fast delivery of speech by the speaker hinder learner's listening comprehension. Additionally, by synchronization, the system emulates the speech flow which facilitates text-to-speech mapping and avoids the salient appearance of the words on the screen. Besides, to generate a suitable caption for a particular learner, the system assesses the tolerable rate of speech and vocabulary size of the leaner and prepares the captions in accordance to his/her level of competence.

Evaluated in two CALL classes, the results of the experiment showed that students' scores using PSC overtook that of the no-caption condition while resulted in almost equal comprehension as the full-caption condition. Furthermore, learner's scores on a new segment of the video without caption was significantly higher than other conditions when they watched the video with PSC first. The finding highlights the positive effect of PSC in preparing learners for listening in real-world situations.

The results also indicate that our method can assist learners to obtain adequate comprehension of the video by presenting less than 30% of the transcript to them. Such a method is assumed to be

effective particularly for Japanese students who heavily rely on caption text in order to comprehend the content of the video. The findings further suggest that this form of captioning can be effectively incorporated into CALL systems as an alternative method to enhance L2 listening comprehension.

Long-term study requires both time and dedicated resources such as CALL classes that in this stage of the study was infeasible. Instead, we considered the immediate effect of the proposed method presuming a real-world situation by checking the learner's comprehension of a new segment of the video without any caption after being exposed to our proposed method. Although the findings has shown comprehension improvement on a short-time adaptation experiment, given the nature of listening skill, overall improvement could not be realized unless the participants undertake long-term experiments, hence such an experiment is suggested.

References

Bailly, G., & Barbour, W. S. (2011). Synchronous reading: learning French orthography by audiovisual training. *Proceedings of Interspeech 2011*, 1153-1156.

Bean, R. M., & Wilson, R. M. (1989). Using closed captioned television to teach reading to adults. *Literacy Research and Instruction*, *28*(4), 27-37.

Bird, S. A., & Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics*, *23*(04), 509-533.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal des traducteursMeta:/Translators' Journal*, *49*(1), 67-77.

Dunkel, P. A., & Davis, J. N. (1994). The effects of rhetorical signaling cues on the recall of English lecture information by speakers of English as a native or second language. *Academic listening: Research perspectives*, 55-74.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. The handbook of second language acquisition, 63-103.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, amt015.

Garza, T. J. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals*, *24*(3), 239-258.

Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(1), 55-75.

Griffin, R., & Dumestre, J. (1992). An initial evaluation of the use of captioned television to improve the vocabulary and reading comprehension of navy sailors. *Journal of Ed. Tech. Systems*, *21*(3), 193-206.

Guillory, H. G. (1998). The effects of keyword captions to authentic French video on learner comprehension. *Calico Journal*, *15*(1-3), 89-108.

Lee, A., & Kawahara, T. (2009). Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC* 2009, 131-137.

Leveridge, A. N., & Yang, J. C. (2013). Testing learner reliance on caption supports in second language listening comprehension multimedia environments. *ReCALL*, *25*(02), 199-214.

Markham, P. (1999). Captioned Videotapes and Second-Language Listening Word Recognition. *Foreign Language Annals*, *32*(3), 321-328.

Montero Perez, M., Peters, E., & Desmet, P. (2014). Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL*, *26*(01), 21-43.

Munteanu, C., Penn, G., & Baecker, R. (2007). Web-based language modelling for automatic lecture transcription. In *Interspeech*, 2353-2356.

Naptali, W. and Kawahara, T. (2012): Automatic Transcription of TED Talk, *Proceedings IWSLT 2012*.

Nation, I. S. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review/La revue canadienne des langues vivantes*, 63(1), 59-82.

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

Nissan, S. (1996). An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension. TOEFL Research Reports, 51.

Nitta, H., Okazaki, H., & Klinger, W. (2011). Speech Rates and a Word Recognition Ratio for Listening Comprehension of Movies. *Bulletin of English Movie Education Society*, (16), 5-16.

Osuka, N. (2007). What factors affect Japanese EFL learners' listening comprehension, *JALT2007 Challenging Assumptions*, *40*(5), 337-344.

Shimogori, N., Ikeda, T., & Tsuboi, S. (2010). Automatically generated captions: will they help non-native speakers communicate in english?. In *Proc. of the 3rd Int'l conf. on Intercultural collaboration*, 79-86.

Suzuki, H. (1996). The effects of closed-captioned videos on the listening and viewing processes of EFL learners. *Studies in Language*, 19, 19-34.

Vandergrift, L. (2004). Listening to Learn or Learning to Listen? *Annual Rev. of Applied Linguistics*, *24*(1), 3-25.

Vanderplank, R. (1993). A very verbal medium: Language learning through closed captions. *TESOL*, *3*(1), 10-14.

Wingfield, A., Poon, L. W., Lombardi, L., & Lowe, D. (1985). Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *Journal of Gerontology*, *40*(5), 579-585.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65-86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors Influencing the Use of Captions by Foreign Language Learners: An Eye-Tracking Study. *The Modern Language Journal*, *97*(1), 254-275.