Vol.2014-NL-216 No.15
Vol.2014-SLP-101 No.15
2014/5/23

Partial and Synchronized Caption Generation to Enhance the Listening Comprehension Skills of Second Language Learners

MARYAM SADAT MIRZAEI[1,a)] TATSUYA KAWAHARA[1]

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. Moreover, it performs better than other conditions for a new segment of the same video without any captions.

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 route and makes a phase of sophistication and concern for many language learners. 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 technique of captioning 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 a word-level alignment feature of the 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 the 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/unrecognizable words.

Following this introduction, this paper reviews the previous studies and describes the proposed technique of captioning. Then, the experimental procedures together with the results are demonstrated. 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

[1] Graduate School of Informatics, Kyoto University, Japan
[a)] maryam@ar.media.kyoto-u.ac.jp

"visual text delivered via multimedia that matches the target language auditory signal verbatim" (Leveridge and Yang [19] , p.1). Captions neatly demonstrate the word boundaries without being affected by the accent, pronunciation and audio deficiencies (Vanderplank, [37]) and allow the learners to parse the speech stream into meaningful chunks, the process known as essential for learning (Ellis [10]).

A considerable amount of literature has been published on the different beneficial effects of captions. Some of these studies have investigated the effect of captioning on vocabulary acquisition (Bird and Williams [4]; Duquette and Painchaud [9]; Griffin and Dumestre [14]; Neuman and Koskinen [29]), grammar acquisition (Lommel et al. [20]), reading development (Bean and Wilson [3]), word recognition (Bird and Williams [4]; Markham [21]) and listening comprehension (Baltova [2]; Borras and Lafayette [5]; Danan [7]; Garza [12]; Huang and Eskey [16]; Markham [21]; Markham et al. [22,23]; Price [32]; Perez et al. [24]; Taylor, [35]; Vanderplank, [36, 37]; Winke et al., [39]).

For instance, Garza [12] 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 Obari et al. (1993), Suzuki [34], and Ikeda (1994) reported the positive effect of English caption on Japanese listening comprehension development [31]. The type and manner of captioning may influence the effect of this assistive tool on language learning. Garza (1991, p.246) suggests using various types of open captioning, such as verbatim, paraphrase and keywords.

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. (2006) used ASR to generate transcripts of a webcast lecture to examine the native speakers' comprehension. 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 by Shimogori et al. [33] who suggest that captions with 80% accuracy will increase the understanding of the Japanese intermediate 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 [1] developed a system that exploits the alignment of text with audio at various levels (letters, phones, syllables, words, chunks, 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 [15] 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" (p.95).

In a recent study by Perez et al. [24]), 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, causing 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 [15] 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 Limitation 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. [40] investigated learners' use of captions and reported that learners read the captions on average 68% of the time.

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 Caption: PSC)

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 is consisted

Table 1 Comparison of Different Methods of Captioning

Advantage - Caption Type	Full	Keyword	Proposed Partial	Synchronized	**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		✓	✓		✓

of two components; synchronization and partialization where the two are complementary and have counteracted the demerits of one another.

First, synchronized caption is automatically generated; a word-level synchronization of text with speech is realized by ASR. The word-level alignment, which synchs each word with the speakers 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 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, making 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 needs, the method involves words which are beyond theproficiency 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 (Perez et al., 2014). Nevertheless, this problem is mitigated by synchronization, as 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.

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 because they have been identified as major contributing factors for listening comprehension impair. Besides, these factors could be quantified automatically and were relatively easy to be implemented.

3.1.1 Speech Rate

Previous studies showed that high speech rate can negatively affect L2 listeners' comprehension (Dunkel [8]) and this is even true for native speakers; the number of words accurately heard by native speakers will decrease as the speaking speed increases (Wingfield et al. [38]). Same results were obtained for non-native speakers (Tomita, 1998). For Japanese learners of English, fast rates of speech and inability to perceive the sounds in English could most impair comprehension (Naoko [25]).

Some studies suggested modification of speech rate as a solution; however this is not close to real-world situation. Instead, we can provide the learner with PSC that represent words or phrases uttered faster than the normal rate of speech, or that of tolerable to 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 word 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 he/she missed (Goh [13]).

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 et al. [30]). Word frequency is calculated based on its occurrence in spoken or written corpora. A well-cited paper by Nation [27] 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 (Nation [28]).

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 impair (Goh [13]). 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 the proposed method.

4. System Architecture

Figure 1 depicts the dataflow 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, vocabulary frequency, 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 feature vector for the decision making module.

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

In the formatting and display module, the captions are altered

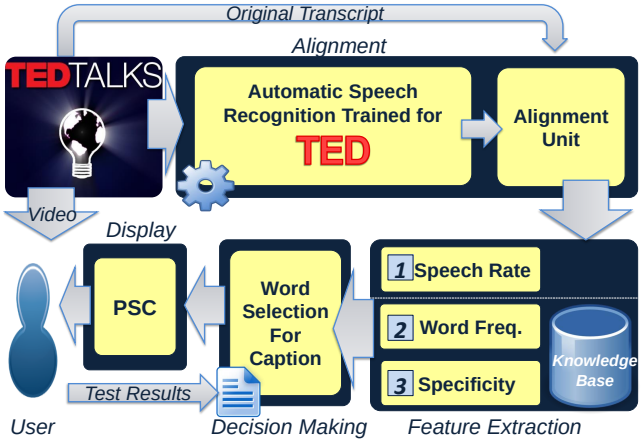


Figure 1 Dataflow and Main Components of the System

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.

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 [18]; Lee et al. [17]). 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 the 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 [26]). 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.

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 may result in misinterpretations in measurement while the latter, SPS, is more suitable to measure short speeches and thus is selected as the unit of measurement 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 the speech rate requires the duration of the word, which is calculated using the time-tags obtained in the alignment phase and excluding the long pauses.

4.2.2 Word Frequency

Word frequency is defined by referring to corpus-based studies. Nation [27] 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. Besides, learner's knowledge of vocabulary is assessed by a vocabulary size test (Nation [28]) and is used as a feature input to the system. 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, to get an exact frequency value.

4.2.3 Word Specificity

In this method specific words are determined using a popular catalogue called Academic Word List (AWL) by Coxhead [6] 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 [11]) for more accuracy.

The system is also capable of dealing with general features such as *abbreviations, proper names, interjections, numbers, transparent compounds and repeated appearance of words*.

4.3 Decision Making Module

Having calculated 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, and readability.

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 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 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 acts on each word. The features are either excitatory or inhibitory. The decision based on general features are made on top of the first stage.

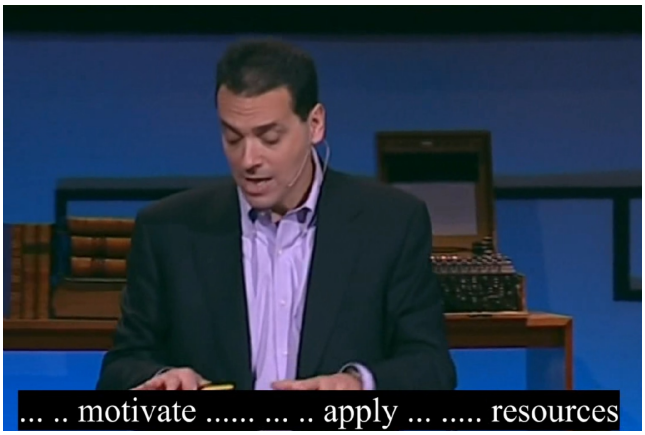


Figure 2 Screenshot of PSC on a TED talk. The caption is made out of the original transcript "how we motivate people how we apply our human resources" based on the extracted features.

For instance, abbreviations and proper names are being marked to be displayed while interjections are marked to be discarded directly. The third stage of decision-making is about the sequence of the words that should be readable and understandable for the students. The rules also handle words after numbers and words after apostrophe s in this version.

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 "dont" will be replaced by "...."). Figure 2 shows the screenshot of the generated PS.

5. Experiment

Given the novelty of partial and synchronized captioning method, the effectiveness of this technique needs to be evaluated. Thus, the study investigates the following questions:

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 in 2007 was used to evaluate the vocabulary reservoir

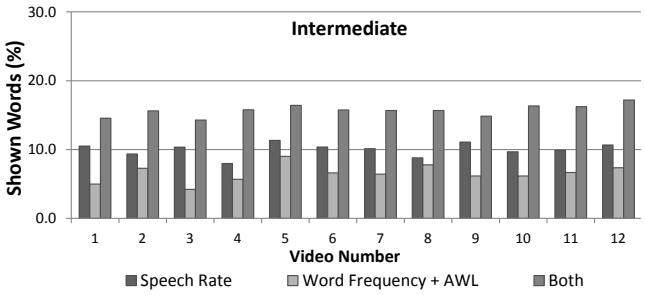


Figure 3 Percentage of words shown in PSC for intermediate learners (speech rate = 5.30 sps, vocabulary size = 4000 word families)

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 a specific treatment (caption) type, the students were asked to take a listening comprehension test on the video in the form of multiple-choice and cloze test on summary.

Questionnaire: In the final session of the experiment, the students were also asked to fill out a 5-point Likert-scale questionnaire about their experience of using our method.

5.3 Procedure

The study was conducted in CALL classes where students were provided with 20 inch-wide screens and headphones. Although the experiment was carried out in two different classes, the same procedure was adopted to maintain similar conditions except that the students in the second class had a trail session to familiarize with the new method. The same videos were being captioned with a different method (PSC ↔ FC) for each class.

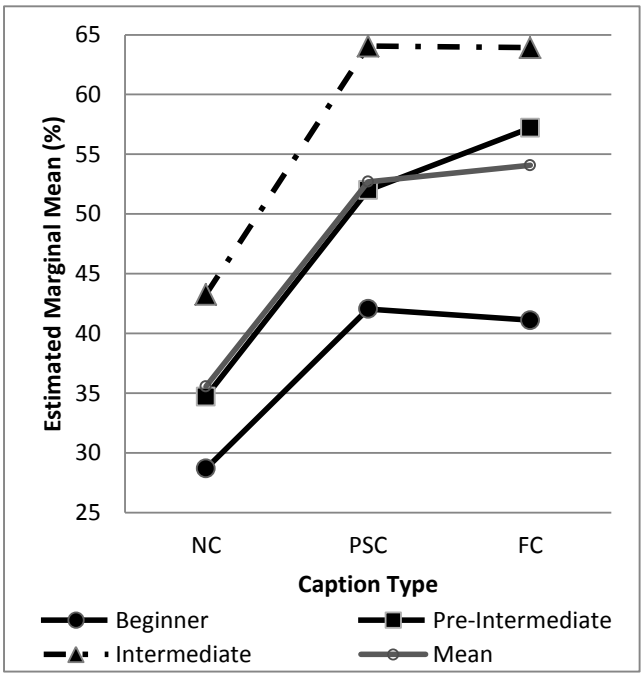
We considered learner's proficiency as a blocking factor, with three levels: "beginner", "pre-intermediate" and "intermediate". For the purpose of dividing our students into these three groups, the aforementioned vocabulary size test together with the students' results on a TOEIC or CASEC test were considered.

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),

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

(a) Descriptive statistics



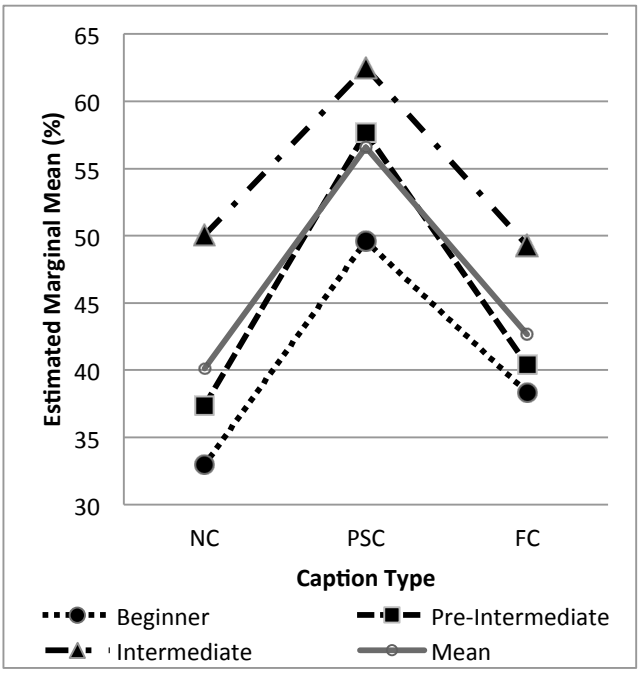
(b) Profile plot

Figure 4 Comprehension performance of the two classes on the first part of the video with different treatments (NC, PSC, FC)

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

(a) Descriptive statistics



(b) Profile Plot

Figure 5 Comprehension performance of the two classes on the second part of the video without caption after watching the first part with a treatment

and took another test. The procedure was 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 like real-world condition.

6. Results, Analysis 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 Analysis of Variance (ANOVA) test was applied to the participants' comprehension scores on different types of captions. Paired-samples t-test was used to see whether equal effects could or could not be assumed on different conditions for each group. The significance level is set to 0.05.

The overall result of both classes are presented in Table 4(a) and Figure 4(b) for part 1 of the experiment (70% of the video with/without caption) and Table 5(a) and Figure 5(b) for part 2 of the experiment (30% of the video without caption).

The findings reveal a significant difference between the NC (M =35.7) condition and the PSC or FC condition (M =52.9; M=54.2). Hence, we have answered our first research question by showing that the students' scores on PCS condition are significantly higher than NC condition (p=0.000).

However, no significant difference was found between the score on PSC and FC condition in part 1 (p = 1.000). This finding provides the answer to our second research question, and suggests that our method can be used as an alternative to full captioning while providing the learner with only what he needs for comprehension (less than 30% of the full-text caption)

To answer our third research question on whether learners with different proficiency levels can benefit from our method, we used a paired-sample t-test to check the differences between the scores gained by beginner (p = 0.681), pre-intermediate (p = 0.479) or intermediate (p = 0.785) students when using PSC as compared to FC. Analysis of the results revealed no significant difference across proficiency groups; all groups gained approximately similar scores on PSC and FC condition. This finding indicates that this method has positive effect on students with different proficiency levels. A possible assumption is that, these results are derived because of the adjustable characteristic of our method that can generate partial captions compatible with the learner's proficiency level and hence useful for any learner.

In the second part of the experiment, PSC showed a significant influence on learner's performance when they later watched a video without caption. In other words, the best performance on the 30% of the video without any caption is associated with the condition in which the learners first watched 70% of that video with PSC as compared to FC: p = 0.024 (beginner), p = 0.002 (pre-intermediate), p = 0.004 (intermediate).

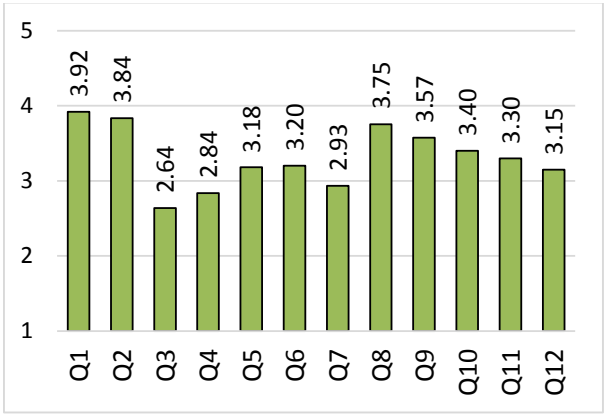
The data on this section provide a positive answer to our fourth research question, which is concerned about the effectiveness of PSC on preparing the learner for real-world situation as compared to NC and FC. Although this is a short-term enhancement partly because of adaptation to the video, this finding is still valuable.

A 5-point likert-scale questionnaire was used to get learner feedback on the proposed method. For each item, learners had the choice from 1 (strongly disagree) to 5 (strongly agree) to show the degree they agreed to the corresponding statement.

As Figure 6 illustrates, in general learner feedback on PSC method was positive (Q1 and Q2). Further, we did not find any strong belief on the idea that PSC is distracting (Q3). Data on items Q4 to Q7 reflect that learners are still not sure whether the proposed method can be substituted commonly used FC method.

No.	Questions
Q1	I think PSC is a good idea.
Q2	I think PSC helps me understand.
Q3	I think PSC is distracting.
Q4	I think PSC is better than FC.
Q5	I think PSC is enough to understand.
Q6	I think PSC helped me use my listening skill more.
Q7	I think PSC is better than FC as I can't read all text.
Q8	I think Synchronized Caption is very helpful.
Q9	I think showing "..." instead of hidden words is a good idea.
Q10	I could find most of words I did not know in PSC.
Q11	I could find most of the words with high speech rate in PSC.
Q12	I think the captions of videos were easy to read.

(a) Survey questions



(b) Average of Likert scores

Figure 6 5-point Likert-scale questionnaire results on PSC (1= strongly disagree, 2 = disagree, 3 = neutral, 4 = agree 5 = strongly agree)

That is probably because this method is new to these participants, they are not still used to it. A closer look at this data reveals that learners like the synchronization characteristics of the proposed methods (Q8). For the purpose of enhancing readability of our method, we attempted to elicit learners' ideas through items Q9 and Q12 that resulted in almost positive feedback from the learners, but also suggest that some improvement should be considered. Finally, by items Q10 and Q11, we investigated the learner feedback on the selection of the words to appear on PSC which seem to have gained almost positive responses.

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 captions that deal 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 the proposed method overtook

that of the no-caption condition while resulted in almost equal comprehension as 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 the learner for listening in real-world situations.

The findings of this study 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

[1] Bailly, G., Barbour, W.-S. et al.: Synchronous reading: learning French orthography by audiovisual training, *Proceedings of Interspeech 2011*, pp. 1153–1156 (2011).

[2] Baltova, I.: Multisensory language teaching in a multidimensional curriculum: The use of authentic bimodal video in core French, *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, Vol. 56, No. 1, pp. 31–48 (1999).

[3] Bean, R. M. and Wilson, R. M.: Using closed captioned television to teach reading to adults, *Literacy Research and Instruction*, Vol. 28, No. 4, pp. 27–37 (1989).

[4] Bird, S. A. and Williams, J. N.: The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling, *Applied Psycholinguistics*, Vol. 23, No. 04, pp. 509–533 (2002).

[5] Borrás, I. and Lafayette, R. C.: Effects of multimedia courseware subtitling on the speaking performance of college students of French, *The Modern Language Journal*, Vol. 78, No. 1, pp. 61–75 (1994).

[6] Coxhead, A.: A new academic word list, *TESOL Quarterly*, Vol. 34, No. 2, pp. 213–238 (2000).

[7] Danan, M.: Captioning and subtitling: Undervalued language learning strategies, *Meta: Journal des traducteurs Meta:/Translators' Journal*, Vol. 49, No. 1, pp. 67–77 (2004).

[8] Dunkel, P. A. and Davis, J. N.: 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*, pp. 55–74 (1994).

[9] Duquette, L. and Painchaud, G.: A comparison of vocabulary acquisition in audio and video contexts, *Canadian modern language review*, Vol. 53, No. 1, pp. 143–172 (1996).

[10] Ellis, N. C.: Constructions, chunking, and connectionism: The emergence of second language structure, *The handbook of second language acquisition*, pp. 63–103 (2003).

[11] Gardner, D. and Davies, M.: A New Academic Vocabulary List, *Applied Linguistics*, p. amt015 (2013).

[12] Garza, T. J.: Evaluating the use of captioned video materials in advanced foreign language learning, *Foreign Language Annals*, Vol. 24, No. 3, pp. 239–258 (1991).

[13] Goh, C.: A cognitive perspective on language learners' listening comprehension problems, *System*, Vol. 28, No. 1, pp. 55–75 (2000).

[14] Griffin, R. and Dumestre, J.: An initial evaluation of the use of captioned television to improve the vocabulary and reading comprehension of navy sailors, *Journal of Educational Technology Systems*, Vol. 21, No. 3, pp. 193–206 (1992).

[15] Guillory, H. G.: The effects of keyword captions to authentic French video on learner comprehension, *CALICO Journal*, Vol. 15, No. 1-3, pp. 89–108 (1998).

[16] Huang, H.-C. and Eskey, D. E.: The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students, *Journal of Educational Technology Systems*, Vol. 28, No. 1, pp. 75–96 (2000).

[17] Lee, A. and Kawahara, T.: Recent development of open-source speech recognition engine julius, *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, pp. 131–137 (2009).

[18] Lee, A., Kawahara, T. and Shikano, K.: Julius—an open source real-time large vocabulary recognition engine (2001).

[19] Leveridge, A. N. and Yang, J. C.: Testing learner reliance on caption supports in second language listening comprehension multimedia environments, *ReCALL*, Vol. 25, No. 02, pp. 199–214 (2013).

[20] Lommel, S., Laenen, A. and d'Ydewalle, G.: Foreign-grammar acquisition while watching subtitled television programmes, *British Journal of Educational Psychology*, Vol. 76, No. 2, pp. 243–258 (2006).

[21] Markham, P.: Captioned Videotapes and Second-Language Listening Word Recognition, *Foreign Language Annals*, Vol. 32, No. 3, pp. 321–328 (1999).

[22] Markham, P. and Peter, L.: The influence of English language and Spanish language captions on foreign language listening/reading comprehension, *Journal of Educational Technology Systems*, Vol. 31, No. 3, pp. 331–341 (2003).

[23] Markham, P., Peter, L. A., McCarthy, T. J. et al.: The effects of native language vs. target language captions on foreign language students' DVD video comprehension, *Foreign Language Annals*, Vol. 34, No. 5, pp. 439–445 (2001).

[24] Montero Perez, M., Peters, E. and Desmet, P.: Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension, *ReCALL*, pp. 1–23 (2014).

[25] Naoko, O.: What factors affect Japanese EFL learners' listening comprehension, *JALT2007 Challenging Assumptions*, Vol. 40, No. 5, pp. 337–344 (2007).

[26] Naptali, W. and Kawahara, T.: Automatic Transcription of TED Talks (2012).

[27] Nation, I. S.: How large a vocabulary is needed for reading and listening?, *Canadian Modern Language Review/La revue canadienne des langues vivantes*, Vol. 63, No. 1, pp. 59–82 (2006).

[28] Nation, I. and Beglar, D.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).

[29] Neuman, S. B. and Koskinen, P.: Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students, *Reading Res. Quarterly*, pp. 95–106 (1992).

[30] Nissan, S., DeVincenzi, F. and Tang, K. L.: *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*, Educational Testing Service Princeton, NJ (1996).

[31] Nitta, H., Okazaki, H. and Klinger, W.: Speech Rates and a Word Recognition Ratio for Listening Comprehension of Movies, *Bulletin of English Movie Education Society*, No. 16, pp. 5–16 (2011).

[32] Price, K.: Closed-captioned TV: An untapped resource, *Matsol Newsletter*, Vol. 12, No. 2, pp. 1–8 (1983).

[33] Shimogori, N., Ikeda, T. and Tsuboi, S.: Automatically generated captions: will they help non-native speakers communicate in English?, *Proceedings of the 3rd international conference on Intercultural collaboration*, ACM, pp. 79–86 (2010).

[34] Suzuki, J.: An Empirical Study on a Remedial Approach to the Development of Listening Fluency: the Effectiveness of Pausing on Students' Listening Comprehension Ability, *Language Laboratory*, No. 28, pp. 31–46 (1991).

[35] Taylor, G.: Perceived processing strategies of students watching captioned video, *Foreign Language Annals*, Vol. 38, No. 3, pp. 422–427 (2005).

[36] Vanderplank, R.: The value of teletext sub-titles in language learning, *ELT journal*, Vol. 42, No. 4, pp. 272–281 (1988).

[37] Vanderplank, R.: A very verbal medium: Language learning through closed captions, *TESOL journal*, Vol. 3, No. 1, pp. 10–14 (1993).

[38] Wingfield, A., Poon, L. W., Lombardi, L. and Lowe, D.: Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time, *J, of Gerontology*, Vol. 40, No. 5, pp. 579–585 (1985).

[39] Winke, P., Gass, S. and Sydorenko, T.: The effects of captioning videos used for foreign language listening activities, *Language Learning & Technology*, Vol. 14, No. 1, pp. 65–86 (2010).