Facilitating L2 listening through automatic detection of speech and lexical difficulties

Maryam Sadat Mirzaei[1] and Kourosh Meshgi[2]

Abstract. This paper focuses on Partial and Synchronized Caption (PSC) as a tool to train L2 listening and introduces new features to facilitate speech-related difficulties. PSC is an intelligent caption that extensively processes the audio and transcript to detect and present difficult words or phrases for L2 learners. With the new features, learners can benefit from repetition and slowdowns of particular audio segments that are automatically labeled difficult. When encountering high speech rates, the system slows down the audio to the standard rate of speech. For disfluencies in speech (e.g. breached boundaries), the system generates the caption and repeats that video segment. In our experiments, intermediate L2 learners of English watched videos with different captions and functionalities, provided feedback on new PSC features, and took a series of tests. Smart repetition and slowdown components received positive learner feedback and led to significant improvement in L2 listening recognition.

Keywords: captioned videos, listening, scaffold, repetition, slowdown, language learning.

1. Introduction

Training L2 listening skills has long been a challenge for L2 learners and teachers, however few tools are developed to foster this process. While listening, learners need to go through complicated procedures such as input decoding at syllable-level, lexical search to find the right words, parsing to apply syntactic rules, and finally, meaning understanding (Field, 2019). Failures in understanding authentic materials can often be attributed not to the general comprehension process but

1. RIKEN AIP, Tokyo, Japan; maryam.mirzaei@riken.jp; https://orcid.org/0000-0002-0715-1624
2. RIKEN AIP, Tokyo, Japan; kourosh.meshgi@riken.jp; https://orcid.org/0000-0001-7734-6104

How to cite this article: Mirzaei, M. S., & Meshgi, K. (2021). Facilitating L2 listening through automatic detection of speech and lexical difficulties. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation: short papers from EUROCALL 2021* (pp. 214-219). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.54.1335

to the recognition of individual words and their segmentation. Some difficulties arise from the listener's unfamiliarity with a word, while others lie in the learner's inability to segment the incoming speech stream into separate words (Salverda, Dahan, & McQueen, 2003). A serious difficulty in L2 listening is caused by the increased number of activated lexical competitors and the challenge to find the right match between the activated candidates and what is just heard. It is more difficult to recognize words when the number of candidates that partially match the input is larger (Broersma, 2012). Given the transient nature of speech, many learners fail to quickly resolve this situation, which distorts subsequent understanding.

Captions can facilitate the listening process by making a phonological visualization of the aural cues (Bird & Williams, 2002). However, matching the selected candidate against the intended input and modifying the hypothesis on-the-fly based on the evidence in the caption can be confusing and may lead to cognitive overload. Thus, learners may require more time to resolve the situation when there is a mismatch between the activated hypothesis and the evidence in the caption.

Apart from these, speakers often change their speaking rate to get the listener's attention or express excitement, anger, etc. Comprehension is impeded when L2 listeners listen to audios with fast speech rates (Renandya & Farrell, 2011). Thus, increased speech rate exuberates the situation as the learners need to accelerate the recognition process and often fail to decode the speech signal. They may know the words in isolation, but fast speech rate often incurs connected speech and imperceptible boundaries, which impedes segmentation. In such cases, even captions may not help the listeners as the words disappear before they can read or recognize them. Thus, learners need another type of scaffold. Speed controllers allow learners to adjust the speech rate to a level they can tolerate. However, listening self-regulation may not assist all learners, especially lower-proficiency listeners, as they need to know when to pause (Roussel, 2011).

We leverage PSC which does word-level synchronization, and omits easy words from the caption to encourage more listening than reading. PSC automatically detects lexical difficulties and problematic speech segments, such as difficult word boundary detection (Mirzaei, Meshgi, & Kawahara, 2018). We introduced a repetition function so that the identified hard segments are automatically replayed after a short pause (Figure 1). Another functionality is a smart and smooth slowdown. PSC calculates speech rate for individual words and finds excessively fast speech segments. It slows down these segments to maintain the average rate of the video but preserves naturalness by following standard speaking rate. Our study

aims to provide a useful scaffold for L2 listeners by targeting specific areas that impede L2 listening.

Figure 1. PSC caption with smart slowdown and repetition functionalities

2. Proposed method

In PSC, we detect speech difficulties by using Automatic Speech Recognition (ASR) systems. Similar to L2 listeners, these systems produce recognition errors when encountering problematic speech segments and disfluencies. The generated errors of the ASR system can indicate cases of breached boundaries, minimal pairs, homophones, and acoustically similar words. Thus, it helps to identify where in the input learners may face perceptual difficulties. Using this, our system detects potentially challenging segments, shows them in the caption, and sets for repeating those segments automatically after a short pause (with adjustable number of repetitions and the default being one). Resolving such ambiguous segments requires a more complicated process. Therefore, through repetition, our system provides more time for processing the input and matching/readjusting the candidate words and the boundaries.

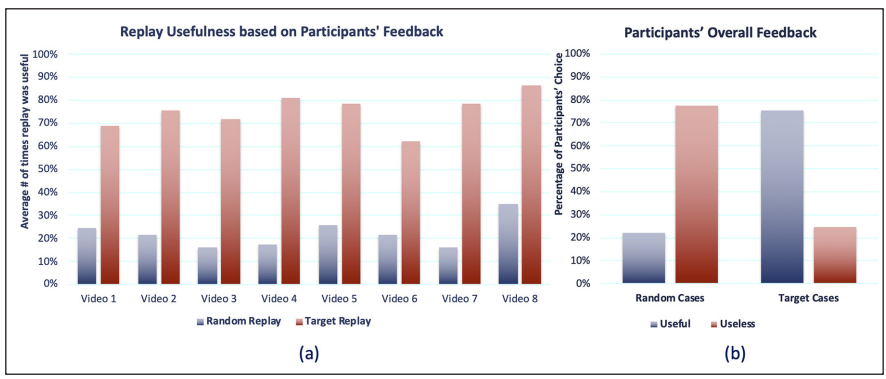
PSC uses a force-alignment procedure that matches the ASR-generated transcripts with human-annotated transcripts to find both the ASR errors and the time tag for each word. We used a TED corpus that provides human annotations to conduct the forced-alignment procedure. PSC calculates the speech rate of each word in syllables per second, then monitors the speech rate of the adjacent words to identify

the segments with significantly faster rates of speech compared to the speaker's average speaking rate. This usually happens due to changes in the emotion of the speaker. The system smoothly lowers the speed of the video as far as it normally blends with the rest of the video without being too slow or unnatural. The reason is that we avoid over-simplification, as learners should be able to handle normal rates of speech. The system looks for instances of high speech rate and speech disfluencies to determine which facilitative strategy to use while allowing the learners to change the settings or turn it off if needed.

3. Experiments and results

We conducted experiments with 37 intermediate learners of English (university students) and asked them to watch a series of videos (1~2 minutes) with several target and random replays that they were not aware of. After each replay, learners chose whether the replay was useful or not. Figure 2a shows their feedback for each video. Most participants found the target repetition more useful than random ones, indicating that our system could accurately detect the problematic segments, and the repetition was beneficial to the listeners (Figure 2b). However, we also noticed that some participants prefer not to have replays at all, whereas some welcome any extra scaffold and always choose to have repetition even for easy segments (random). Thus, it is important to allow learners to customize the system based on their preferences.

Figure 2. Learner feedback on random and target replays

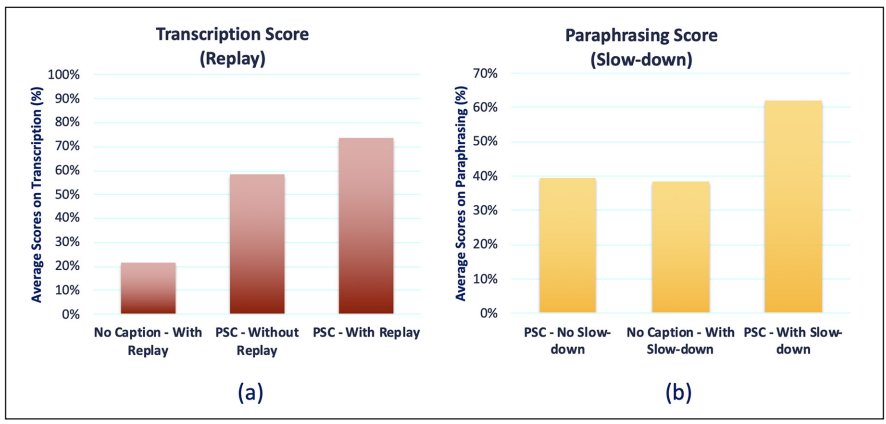


Next, the participants were divided into three groups: no-caption with repetition, PSC without repetition, and PSC with repetition. At the end of the videos, learners transcribed a target replay segment. Results in Figure 3a show that learners

struggled in transcribing the target segments that involved perceptual difficulties. Their transcription included incorrect segmentation and boundary detection. Adding the repetition feature to PSC significantly improved the performance and better assisted the learners. A possible explanation is that learners may be unaware of alternative segmentation or stick to their guess and cannot abandon it (Field, 2019). However, repetition provides another chance to reidentify the boundaries, decode more accurately, and consider other hypotheses.

Finally, we maintained the same grouping but used paraphrasing tests for the slowdown: no-caption with slowdown, PSC without slowdown, and PSC with slowdown. Figure 3b shows that without the slowdown functionality, PSC is not very helpful when the speech rate is considerably fast since the learners do not get a chance to follow the caption. However, we need to consider that speech rate may not be the only factor, and lexical difficulties can also cause problems.

Figure 3. Learners' performance when having replays (a) and slowdowns (b) as assistive features



4. Conclusions

We used PSC as a tool to provide a timely scaffold for the L2 listeners. We detected L2 listeners' challenges in the input and enabled three forms of assistance: presenting difficult words/phrases in the caption, repeating potential breached boundary cases, and slowing down the parts with excessively fast speech rates. Learner feedback indicated that the system could successfully detect the hard-to-recognize segments and highlighted the usefulness of our assistive strategy. Moreover, with the replay function in PSC, learners could better transcribe the video and set the correct

boundaries. Findings also revealed that the automatic detection of speed variation and slowing down the relative segment could assist learners to better comprehend the input. Since PSC's features are adjustable, learners can gradually decrease the amount of shown words or disable the scaffolding features to cope with real-life situations.

References

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*(4), 509-533. https://doi.org/10.1017/S0142716402004022

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and cognitive processes, 27*(7-8), 1205-1224. https://doi.org/10.1080/01690965.2012.660170

Field, J. (2019). Second language listening: current ideas, current issues. In J. Schwieter & A. Benati (Eds), *The Cambridge handbook of language learning*. Cambridge University Press. https://doi.org/10.1017/9781108333603.013

Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2018). Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening. *Computer Speech & Language*, *49*, 17-36. https://doi.org/10.1016/j.csl.2017.11.001

Renandya, W. A., & Farrell, T. S. (2011). 'Teacher, the tape is too fast!' Extensive listening in ELT. *ELT journal, 65*(1), 52-59. https://doi.org/10.1093/elt/ccq015

Roussel, S. (2011). A computer assisted method to track listening strategies in second language learning. *ReCALL, 23*(2), 98-116. https://doi.org/10.1017/S0958344011000036

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*(1), 51-89. https://doi.org/10.1016/S0010-0277(03)00139-2

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

CALL and professionalisation: short papers from EUROCALL 2021
Edited by Naouel Zoghlami, Cédric Brudermann, Cedric Sarré, Muriel Grosbois, Linda Bradley, and Sylvie Thouësny

Publication date: 2021/12/13

Typeset by Research-publishing.net

Cover Theme by © 2021 DIRCOM CNAM; Graphiste : Thomas Veniant
Cover Photo by © 2021 Léo Andres, Sorbonne Université
Cover Photo by © 2021 Sandrine Villain, Le Cnam
Cover Layout by © 2021 Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-97-9 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2021.