Listening difficulty detection to foster second language listening with a partial and synchronized caption system

Maryam Sadat Mirzaei[1], Kourosh Meshgi[2], and Tatsuya Kawahara[3]

Abstract. This study proposes a method to detect problematic speech segments automatically for second language (L2) listeners, considering both lexical and acoustic aspects. It introduces a tool, Partial and Synchronized Caption (PSC), which provides assistance for language learners and fosters L2 listening skills. PSC presents purposively selected words along with a video in the form of synchronized text-to-speech captions. It uses corpus-based information and conducts speech-data analysis to detect difficulties in speech, thus achieving effective word selection. To this end, PSC uses an Automatic Speech Recognition (ASR) system as a model for L2 listeners to elucidate speech difficulties for these learners. In this method, misrecognized words by the ASR system are analyzed to find useful patterns that could signal problematic speech segments for L2 listeners. The identified patterns were evaluated by experiments to ensure that they cause difficulties for L2 listeners in the same way that they impede ASR performance. Experimental findings confirm that adding these instances to PSC significantly improves learners' recognition of the respective segments.

Keywords: L2 listening, partial and synchronized caption, automatic speech recognition, error analysis.

1. Introduction

L2 listening is a transient process which entails sophisticated skills to recognize speech and achieve comprehension (Rost, 2013). While L2 listeners need to process

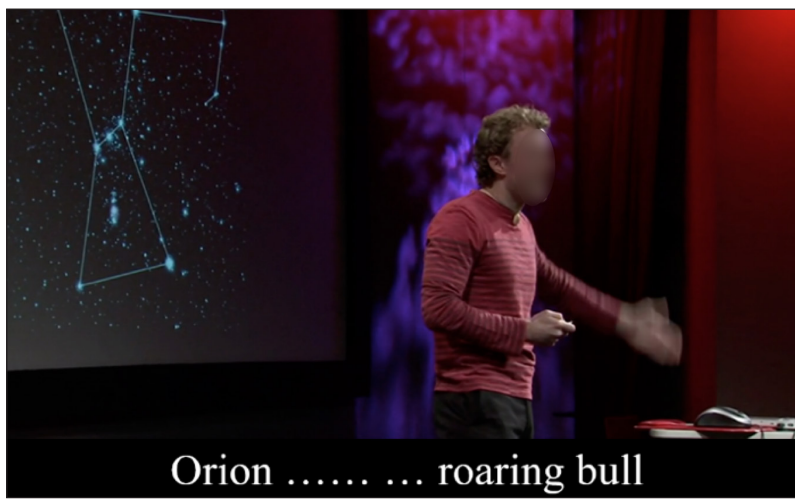
1. Kyoto University, Kyoto, Japan; maryam@sap.ist.i.kyoto-u.ac.jp
2. Kyoto University, Kyoto, Japan; meshgi-k@sys.i.kyoto-u.ac.jp
3. Kyoto University, Kyoto, Japan; kawahara@i.kyoto-u.ac.jp

How to cite this article: Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2017). Listening difficulty detection to foster second language listening with a partial and synchronized caption system. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 211-216). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.715

input attentively and utilize skills adeptly to gain adequate comprehension, some factors associated with the speech input itself can impede their listening (Bloomfield et al., 2010). The lexical units used in speech and the ambiguities involved in articulation, such as uncertain word boundaries, can lead to difficulties for many learners (Field, 2008). Despite many studies focusing on L2 listening difficulties, analyzing the nature of speech and identifying problematic speech segments for individual L2 listeners has not been systematically investigated. This highlights the need for a tool that can predict speech-related difficulties for language learners and provide a learner-specific scaffold to assist in learner comprehension. This paper describes an attempt to achieve this goal by using ASR errors as a source of information to indicate difficult speech segments for L2 listeners.

ASR systems process speech signals to generate a transcript of an audio track. This process often involves some errors, which are often (but not always) the product of intrinsic speech difficulties (Meyer, Brand, & Kollmeier, 2011). Through a comparison between ASR errors and L2 speech recognition mistakes, we found specific patterns of ASR errors such as homophones, minimal pairs, negative cases, and breached boundaries. These kinds of errors cause similar problems in comprehension for both for ASR and L2 learners (Mirzaei, Meshgi, & Kawahara, 2016). The study described here focuses on the automatic detection of such ASR error patterns and how they can be embedded in PSC to provide better assistance to language learners.

Figure 1. Screenshot of a TED talk with PSC (Original sentence: "Orion facing the roaring bull")



PSC anticipates learners' listening difficulties when using authentic materials by detecting difficult-to-recognize words/phrases and presenting them in the caption to provide assistance to the learner (Figure 1). To this end, it draws on features that impede L2 listening comprehension, based on L2 studies of areas of difficulty. The baseline PSC system is based on a heuristic approach, mainly using lexical features that influence L2 listening, such as word frequency and specificity (Webb, 2010). The only speech-related factor used in the baseline system is the speech rate; it overlooks other patterns. We regard ASR errors as a source to provide other useful cases for PSC. Moreover, as the system strives to serve individual learners, it adjusts word selection to the proficiency levels of the learners. However, for PSC to be more effective, its parameters need to be better tuned.

2. Method

Using TED talks as the material for PSC, we employed an ASR system to generate a transcript and make word-level synchronization. Next, we compared it with the human-annotated transcript, provided by TED, to detect the mismatches (ASR errors). These errors were further analyzed to extract useful instances for speech difficulties. Meanwhile, lexical features were assessed to unveil useful words/phrases for L2 listeners. Based on the outcome of this process, coupled with the assessment of the learners' proficiency levels, the system can automatically decide on the words/phrases that impede listening. To prioritize the most useful instances, we enhanced the baseline system selection by using the ASR error clues.

2.1. Automatic detection of useful ASR errors

We mark a word as (1) a homophone (e.g. "rain" */R EY N/* and "reign" */R EY N/*) if the Levenshtein distance between the phonetic sequences of the ASR erroneous phrase and its corresponding transcript is zero, and (2) a minimal pair (e.g. "pin" */P IH N/* and "bin" */B IH N/*) if the distance is one. Negative cases are detected by considering the negative particle "not", plus prefixes and suffixes that form negation (e.g. "legal" and "illegal"). To detect breached boundaries, we checked for the following patterns based on L2 studies (Cutler, 2005; Field, 2008):

- Strong-syllable strategy: learners tend to insert word boundaries when they encounter a strong syllable and set it as the beginning of the word (e.g. "disguise" heard as "the skies"). In addition, learners tend to delete the boundary before a weak syllable and thus merge the words (e.g. "ten-to-two" heard as "twenty to").

- Assimilation rule: learners have difficulty in setting the right word boundaries due to the common phonological process which alters a word ending sound in expectation of the following sound (e.g. “right you are” as “rye chew are”).

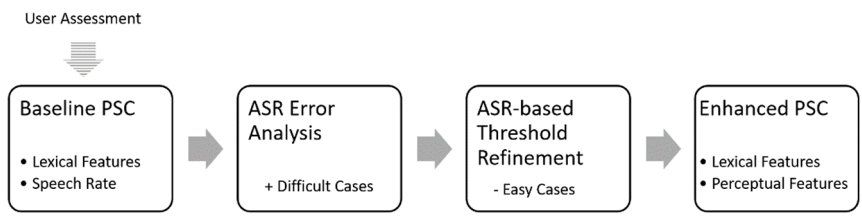
- Frequency rule: learners have a general tendency to insert word boundaries to perceive more frequent words than the actual word (e.g. “achieve her” heard as “a cheaper”).

- Resyllabification: resyllabification happens when the final consonant of a word attaches to the following syllable (e.g. “made out” heard as “may doubt”).

2.2. Enhancing PSC using ASR clues

The baseline PSC system is enhanced not only by embedding the detected cases, but also by improving the word selection through refining the system thresholds (Figure 2). If the learner prefers to receive more words, we may simply add ASR erroneous cases to the baseline system to provide more choices that relate to speech difficulty. However, if the learner prefers minimal, but targeted assistance, the system trims the number of cases by refining the thresholds based on ASR erroneous cases. Therefore, if ASR has difficulty in recognizing a word, the system prioritizes that word to those deemed difficult by the baseline.

Figure 2. Enhanced PSC using ASR clues



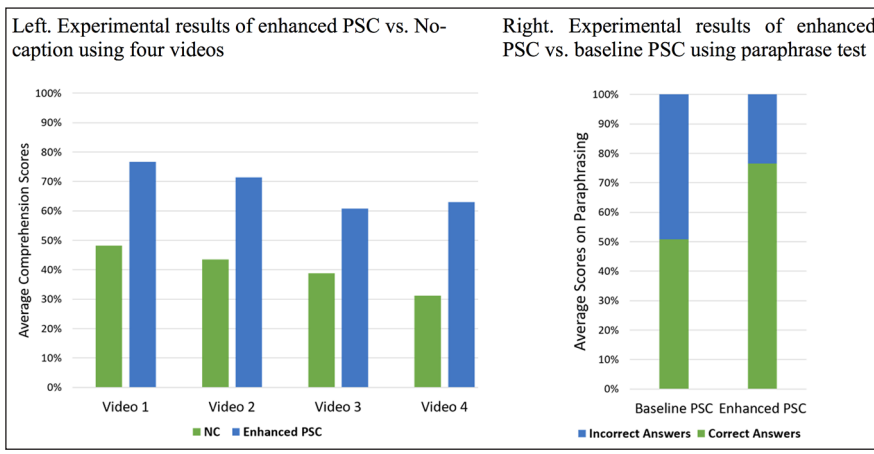
3. Experimental evaluations

Enhanced PSC was compared with the no-caption as well as baseline PSC in experiments with L2 learners of English. In the first experiment (enhanced PSC vs. no-caption), 18 learners of English with TOEIC scores above 750 listened to TED

talks of about three minutes long, either with enhanced PSC or without captions. After each video, the participants were asked to answer the comprehension questions and the listening cloze tests that followed (videos were rotated among participants). Figure 3 Left, shows the participants' scores on no-caption versus enhanced PSC condition. As the figure demonstrates, learners' scores were significantly higher when they received enhanced PSC as opposed to no captions (p<.001).

In the second experiment, 38 intermediate-level learners were divided into two groups and watched a series of short video segments either with baseline PSC or with enhanced PSC, followed by several paraphrasing questions. The numbers of shown words to both groups were controlled to be the same, while the choices of words in the baseline and enhanced versions were different. The results are shown in Figure 3 Right, which suggests using the enhanced version led to a significant increase in participants' paraphrasing scores.

Figure 3. Experimental results of enhanced PSC



4. Conclusions

In this study, we focused on the detection of speech-related difficulties for L2 learners and proposed the use of ASR errors as indicators of such difficulties. Compared with other studies, which consider ASR errors as a drawback of these systems, in this study we tried to make use of these errors and considered them as indicators of speech difficulties. To this end, useful cases of ASR errors for L2 listeners are automatically detected and embedded into the PSC system to provide

better assistance. Moreover, we prioritized PSC choices based on the ASR clues. Experimental results showed increased recognition with the use of enhanced PSC over no captions or the baseline. However, more experiments are needed to reflect PSC's learner-specific adjustments for varying levels of language proficiency. PSC aims to promote learner autonomy in developing L2 listening skills by allowing learners to adjust the number of shown words through the course of study and to customize the system's features based on their preference. This method can be used easily for an ample number of listening materials to provide learners with an adjustable amount of scaffolding and allow teachers to recognize learners' listening difficulties.

References

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*. University of Maryland. https://doi.org/10.21236/ADA550176

Cutler, A. (2005). The lexical statistics of word recognition problems caused by L2 phonetic confusion. *Interspeech Lisboa 2005: 9th European Conference on Speech Communication and Technology, September 4-8*, 413-416. Causal Productions.

Field, J. (2008). Bricks or mortar: which parts of the input does a second language listener rely on? *TESOL quarterly*, *42*(3), 411-432. https://doi.org/10.1002/j.1545-7249.2008.tb00139.x

Meyer, B. T., Brand, T., & Kollmeier, B. (2011). Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *The Journal of the Acoustical Society of America*, *129*(1), 388-403. https://doi.org/10.1121/1.3514525

Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2016). Leveraging automatic speech recognition errors to detect challenging speech segments in TED talks. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 313-318). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.581

Rost, M. (2013). *Teaching and researching: listening* (2nd ed.). Routledge.

Webb, S. (2010). Using glossaries to increase the lexical coverage of television programs. *Reading in a Foreign Language*, *22*(1), 201-221.

Research-publishing.net

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

CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017
Edited by Kate Borthwick, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net

Cover design based on © Josef Brett's, Multimedia Developer, Digital Learning, http://www.eurocall2017.uk/, reproduced with kind permissions from the copyright holder.

Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "frog" on cover by © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-04-7 (Ebook, PDF, colour)
ISBN13: 978-2-490057-05-4 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-03-0 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

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

Legal deposit: Bibliothèque Nationale de France - Dépôt légal: décembre 2017.