Leveraging automatic speech recognition errors to detect challenging speech segments in TED talks

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

Abstract. This study investigates the use of Automatic Speech Recognition (ASR) systems to epitomize second language (L2) listeners' problems in perception of TED talks. ASR-generated transcripts of videos often involve recognition errors, which may indicate difficult segments for L2 listeners. This paper aims to discover the root-causes of the ASR errors and compare them with L2 listeners' transcription mistakes. Our analysis on the ASR errors revealed several categories, such as minimal pairs, homophones, negative cases, and boundary misrecognition, which are assumed to denote the challenging nature of the respective speech segments for L2 listeners. To confirm the usefulness of these categories, we asked L2 learners to watch and transcribe a short segment of TED videos, including the above-mentioned categories of errors. Results revealed that learners' transcription mistakes substantially increase when they transcribe segments of the audio in which ASR made errors. This finding confirmed the potential of using ASR errors as a predictor of L2 learners' difficulties in listening to a particular audio. Furthermore, this study provided us with valuable data to enrich the Partial and Synchronized Caption (PSC) system we proposed earlier to facilitate and promote L2 listening skills.

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

1. Introduction

Among the four skills that play crucial roles in L2 learning, listening is of paramount importance for being a prerequisite of speaking and for providing the language input (Rost, 2005). Mastery in listening needs exposure to authentic

[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. (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

audio/visual input, which in turn makes the listening process more difficult. The main reason lies in the fact that listening is a complicated process, from perceiving the speech to comprehending the content, and when using authentic materials, L2 listeners are more likely to encounter difficulties and make recognition mistakes (Gilmore, 2007).

Despite a body of work (Bloomfield et al., 2010; Révész & Brunfaut, 2013) on L2 listening difficulties, there is no reliable resource to detect those segments of an audio/visual material that cause recognition difficulties for L2 listeners. Motivated by this demand, we decided to find a means for discovering these difficulties in order to scaffold the language learners. Undoubtedly, L2 learners, in accordance with their language proficiency level and individual skills, may encounter different types of problems in recognizing speech. However, certain factors are more likely to result in misrecognition for the majority of L2 learners.

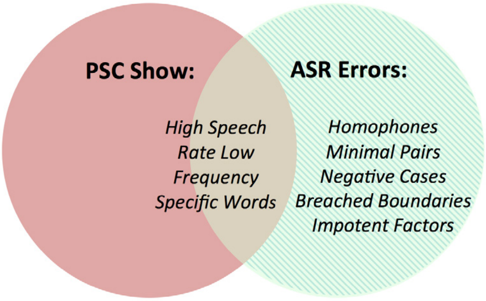
Detecting difficult words and phrases in audio/visual material is initially developed in our work (Mirzaei, Meshgi, Akita, & Kawahara, forthcoming), the PSC system. PSC identifies difficult words in TED talks and presents them in the caption while hiding easy ones to encourage listening over reading. The PSC system detects difficult words based on high speech rate, low word-frequency, and specificity (specific or academic terms). Nevertheless, these factors alone may not encompass all problematic speech segments in different videos.

To consider other factors, this study aims to establish the viability of using ASR as a predictor of L2 listeners' problems. ASR systems can generate transcriptions of spoken language, but this transcript often involves some errors. Despite common criticisms that regard these errors as major drawbacks of ASR systems, this paper investigates their usefulness to discover challenging speech segments of TED talks and to seek means for enhancing the PSC system.

2. Method

The Julius ASR system was used to transcribe 72 TED talks; the ASR errors were derived from the system and given to a linguist expert to perform a root-cause analysis. These errors were then compared with PSC's selected words to show in the caption. Since PSC also chooses difficult words for L2 listeners based on speech rate, frequency and specificity factors, some overlaps are anticipated, while some mismatches can also be detected (Figure 1). These mismatches are automatically extracted and further analysed by the expert to discover the challenging speech segments that are not yet handled by the PSC system (dashed area in Figure 1).

Figure 1. Diagram of the PSC shown words vs. ASR error categories: dashed area presents the scope of analysis



Expert analysis on these ASR errors revealed several clusters, four of which seemed to be the prudent sources of listening difficulties for L2 learners: homophones, minimal pairs, negative cases and breached boundaries.

Homophones (e.g. plain/plane) and minimal pairs (e.g. face/faith) hinder recognition processes by activating several word candidates (Weber & Cutler, 2004) and imposing high-level semantic analysis to select the right choice. Moreover, the wrong choice of words may thoroughly change the meaning of the sentence and deteriorate the comprehension. The same argument can be made for negative cases (e.g. can/can't). Finally, breached boundary (i.e. wrong lexical segmentation) is perhaps the most important category of errors that are very likely to arise when L2 learners watch a video; identifying word or phrase boundary is a challenging part of L2 listening (Field, 2009). A number of factors can be responsible for wrong lexical segmentation, including, assimilation, stress patterns, frequency rule and resyllabification (Cutler, 1990; Field, 2003). Regardless of their causes, breached boundary occurrences are hard to be predicted. ASR errors, however, provide helpful clues to detect such kinds of potentially misrecognized boundaries.

3. Experiment

A transcription experiment is conducted to verify whether these four categories of ASR errors are actually problematic for L2 listeners. The data from this experiment allows comparing ASR errors to those of L2 learners' recognition mistakes.

The participants of this study were 11 Japanese and 10 Chinese students who were undergraduates and graduates majoring in different fields. All participants were intermediates with TOEIC scores (or equivalents) of above 650. There were 8 female and 13 male students.

The material of this experiment included twenty talks selected out of 72. The selected videos were delivered by native American English speakers in order to exclude the effect of other accents (e.g. British English). Two segments were chosen from each video:

- one including four ASR error categories which PSC hid ('difficult' cases that PSC failed to predict);

- one including ASR correct cases which PSC hid ('easy' cases both for ASR and PSC).

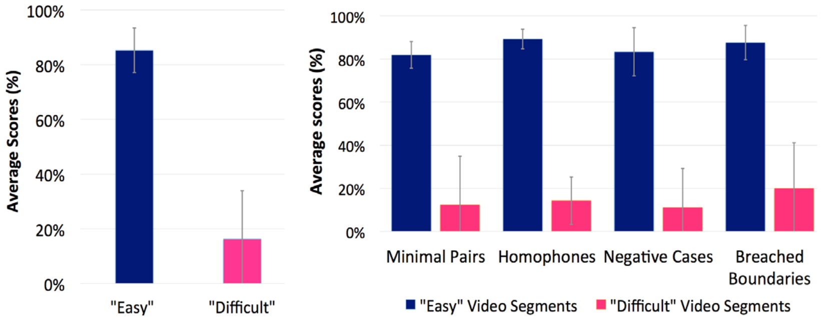
The former involved minimal pairs, homophones, negative forms, and breached boundaries and were tested to identify the difficulty-level of these categories. The latter were chosen to compare the performance of L2 listeners on transcribing easy versus potentially difficult speech segments.

Each segment of the video lasted 25~35 seconds and stopped at an irregular interval, which ended with 4~6 words, including the target word (easy or difficult case). Irregular pauses were made manually to maintain real life listening and avoid word-by-word decoding. The participants were not aware of when the videos were going to stop or which word would be the target word. They could watch each video only once and were supposed to type the last few words they heard immediately after the pause. Limited time was set to avoid overthinking and analyzing, thereby allowing the participants to input what they recognized. The test was made using iSpring Quiz Maker and was launched online. Spelling errors were ignored, unless affecting the meaning.

4. Results

The results of this experiment include the scores of the participants on transcribing the easy segments of the videos versus the difficult parts that included minimal pairs, homophones, negative forms, and breached boundaries. As Figure 2 (Right) illustrates, the mean scores of the students on the easy segments (*M*=0.85, *SD*=0.08) is much higher compared to the difficult segments (*M*=0.16, *SD*=0.18).

Figure 2. (left) Overall average scores on ‘easy’ vs. ‘difficult’ segments. (right) Detailed result on four subcategories of ASR errors vs. respective easy segments



The results of a t-test indicate that this difference is statistically significant (*t*(19)=18.131; p<.05). Figure 2 (left) also suggests that participants’ average scores on each of the categories of errors, i.e. minimal pairs (*M*=.12; *SD*=.22), homophones (*M*=.14, *SD*=.11), negative cases (*M*=.11, *SD*=.18) and breached boundaries (*M*=.20, *SD*=.21) were statistically lower than their average scores on respective easy segments of the videos. Interestingly, similar mistakes were found in participants’ transcriptions and ASR generated transcript (e.g. “make a new ear” was transcribed as “make a new year” by both ASR and the participants).

5. Conclusion

The study investigated the use of ASR errors in detecting challenging speech segments of TED talks and improving the word selection criteria in PSC. Following a thorough root-cause analysis, it was found that several categories in the ASR errors suggest the difficulties for L2 listeners. These categories included homophones, minimal pairs, negative forms, and breached boundaries. An experiment with L2 listeners confirmed the feasibility of using these ASR errors to predict L2 speech recognition difficulties. This finding can provide means for future advances of the PSC system by exploiting ASR clues to optimize the choice of words in the caption. In this view, the enhanced version should be compared with the current one to investigate any improvement. Furthermore, discovering challenging speech segments allows the learners to identify their listening problems and provide the teachers with useful information to better scaffold the learners.

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*. Maryland University.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: psycholinguistic and perceptional perspectives* (pp.105-121). Cambridge: MIT Press.

Field, J. (2003). Promoting perception: lexical segmentation in L2 listening. *ELT journal, 57*(4), 325-334. https://doi.org/10.1093/elt/57.4.325

Field, J. (2009). *Listening in the language classroom*. Cambridge University Press. https://doi.org/10.1017/cbo9780511575945

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40*(2), 97-118. https://doi.org/10.1017/S0261444807004144

Mirzaei, M. S., Meshgi, K., Akita, Y., & Kawahara, T. (forthcoming) Partial and synchronized captioning: a new tool to assist learners in developing second language listening skill. *ReCALL*.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*(1), 31-65. https://doi.org/10.1017/S0272263112000678

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-527). Lawrence Erlbaum Associates, Inc., Publishers.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50*(1), 1-25. https://doi.org/10.1016/S0749-596X(03)00105-0

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

CALL communities and culture – short papers from EUROCALL 2016
Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it is not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover design by © Easy Conferences, info@easyconferences.eu, www.easyconferences.eu
Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "bridge" on cover by © Andriy Markov/Shutterstock
Photo "frog" on cover by © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-43-8 (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'.

ISBN13: 978-1-908416-44-5 (Ebook, PDF, colour)
ISBN13: 978-1-908416-45-2 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
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 2016.