Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 49 (2018) 17–36

www.elsevier.com/locate/csl

Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening☆

Maryam Sadat Mirzaei*, Kourosh Meshgi, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo Ward, Kyoto 606–8501 Japan

Received 14 November 2016; received in revised form 27 August 2017; accepted 4 November 2017
Available online 10 November 2017

Abstract

This paper addresses the viability of using Automatic Speech Recognition (ASR) errors as the predictor of difficulties in speech segments, thereby exploiting them to improve Partial and Synchronized Caption (PSC), which we have proposed to train second language (L2) listening skill by encouraging listening over reading. The system uses ASR technology to make word-level text-to-speech synchronization and generates a partial caption. The baseline system determines difficult words based on three features: speech rate, word frequency and specificity. While it encompasses most of the difficult words, it does not cover a wide range of features that hinder L2 listening. Therefore, we propose the use of ASR systems as a model of L2 listeners and hypothesize that ASR errors can predict challenging speech segments for these learners. Among different cases of ASR errors, annotation results suggest the usefulness of four categories of homophones, minimal pairs, negatives, and breached boundaries for L2 listeners. A preliminary experiment with L2 learners focusing on these four categories of the ASR errors revealed that these cases highlight the problematic speech regions for L2 listeners. Based on the findings, the PSC system is enhanced to incorporate these kinds of useful ASR errors. An experiment with L2 learners demonstrated that the enhanced version of PSC is not only preferable, but also more helpful to facilitate the L2 listening process.

Keywords: Computer-assisted language learning; Second language listening skill; Automatic speech recognition; Partial and synchronized caption

1. Introduction

The advancement of Information and Communication Technology (ICT) has formed new avenues of research and promoted further opportunities in different domains. The application of these technologies in language learning and teaching is known as computer-assisted language learning – CALL (Levy, 1997), which is quickly changing the teaching materials and the learning environment. CALL systems provide the materials that meet the requirements of different language learners and foster exposure to the contextualized and authentic resources including multimedia presentations, web-based distribution of print-media, radio, and TV programs (Amaral and Meurers, 2011).

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author.
E-mail address: maryam@sap.ist.i.kyoto-u.ac.jp (M.S. Mirzaei).

http://dx.doi.org/10.1016/j.csl.2017.11.001
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

While the effectiveness of using these authentic materials is undeniable, the fact that these resources are often highly challenging for L2 learners is equally evident (Gilmore, 2007). To overcome the difficulties of authentic materials, which may cause a stage of frustration and demotivation, captioning can be used (Danan, 2004). Captioning provides the textual clues and phonological visualization of what is being heard and hence allows the use of reading while listening to comprehend the audio. Nevertheless, many learners prioritize reading the caption text over listening to the audio (Osada, 2004). These strategies assist learners in comprehending the audio but apparently do not promote the use of listening skill if not hinder it (Pujolà, 2002; Vandergrift, 2004).

In order to overcome the shortcomings of conventional captioning, we have proposed a novel captioning system called PSC (Mirzaei et al., 2014; Mirzaei and Kawahara, 2015), which automatically detects difficult words and presents them on the screen to scaffold the L2 listeners, while hiding easy words to encourage more listening than reading. Fig. 1 shows a screenshot of the system. PSC synchronizes the text to speech in word-level using ASR technology. As a baseline, the detection of difficult words is realized based on three defined features: speech rate, word frequency, and word specificity. The level of difficulty and the amount of shown words in PSC is tailored to the requirement of different learners at different levels.

Studies on L2 listening difficulties indicate that learners may encounter a miscellaneous collection of factors that impede their listening (Bloomfield et al., 2010). Among those, the above-mentioned features are of special importance for the main causes of listening difficulties (Griffiths, 1992; Révész and Brunfaut, 2013). However, not all listening challenges could be explained by these features. As a result, PSC's selected words sometimes include several easy to recognize words and occasionally exclude difficult words or phrases, which highlights the importance of exploring other features. One main source of difficulties for many L2 listeners is the wrong boundary detection (Field, 2008). For many language learners, finding the right boundaries between the words in connected speech is often difficult, thus many L2 learners end up being confused with breached boundaries. Such difficulties severely hinder listening, but are not easy to detect without analyzing the nature of the speech and hence are missing in baseline PSC's selected words.

To decipher listening challenges, in this paper, we propose the use of ASR errors as a source to predict difficulties for L2 listening. ASR systems process the speech signal to generate a transcript of the audio file. This process, however, often involves some errors, which can be the product of some intrinsic speech difficulties. In this view, the performance of ASR systems is similar to L2 listeners when it comes to the transcription task. In other words, ASR errors in transcribing speech may derive from the same sources that lead to L2 misrecognition. Therefore, these errors can provide useful clues for the enhancement of PSC.

In this paper, we focus on finding useful patterns or features in the ASR errors to detect problematic speech segments for L2 listeners. The discovered patterns are tested in actual language learning environment to ensure that they cause difficulties for L2 listeners as they impede ASR performance. Then, useful errors are incorporated to the baseline PSC to provide better assistance. Finally, through an experiment, the enhanced version of PSC is compared with the baseline PSC by assessing L2 listeners' preferences and performance on using each version.

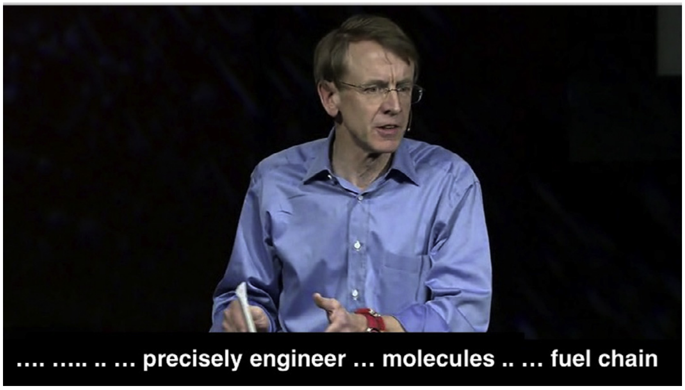
.... precisely engineer ... molecules .. ... fuel chain

Fig. 1. Screenshot of the PSC System: The caption text is presented incrementally in synch with the speech. The original transcript was: "That means we can precisely engineer the molecules in the fuel chain." TED talk by John Doerr: Salvation (and profit) in greentech.