Errors in automatic speech recognition versus difficulties in second language listening

Maryam Sadat Mirzaei¹, Kourosh Meshgi², Yuya Akita², and Tatsuya Kawahara²

Abstract. Automatic Speech Recognition (ASR) technology has become a part of contemporary Computer-Assisted Language Learning (CALL) systems. ASR systems however are being criticized for their erroneous performance especially when utilized as a mean to develop skills in a Second Language (L2) where errors are not tolerated. Nevertheless, these errors can provide useful information and propose further implications. In this study we investigate the relationships between the underlying features causing ASR errors and those that make L2 listening difficult. This research is inspired by the comparable nature of the difficulties both ASR and L2 listeners encounter in recognizing speech. The aim of this study is to enhance Partial and Synchronized Caption (PSC) systems, which we previously developed for fostering L2 listening skill. PSC presents only a selective set of words (those leading to listening difficulties) in order to encourage listening to the audio and read for problematic words only. To enhance PSC’s word selection, we strive to detect individual difficult sentences/words in terms of recognition by referring to ASR errors. Our system compares these errors with PSC choices to find the overlaps and seek further enhancement. The results revealed a close relationship between ASR errors and factors leading to L2 listening difficulties. The findings indicated that ASR errors can contribute to word selection in PSC.

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

1. Kyoto University, Japan; Maryam@ar.media.kyoto-u.ac.jp
2. Kyoto University, Japan; Meshgi-k@sys.i.kyoto-u.ac.jp; Akita@ar.media.kyoto-u.ac.jp; Kawahara@i.kyoto-u.ac.jp

How to cite this article: Mirzaei, M. S., Meshgi, K., Akita, Y., & Kawahara, T. (2015). Errors in automatic speech recognition versus difficulties in second language listening. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy (pp. 410-415). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000367
1. Introduction

There has been increasing interest in the use of ASR technology in the field of second language acquisition. Some possible applications include the pronunciation evaluation in order to improve oral skills and caption generation in order to facilitate listening comprehension (Shimogori, Ikeda, & Tsuboi, 2010; Thomson, 2013).

In spite of their significant advancement, ASR systems are still endeavoring to achieve better accuracy. The limitations of these systems have raised a number of criticisms when they are utilized as a means for L2 development. This is partly because of the particular challenges involved. For instance, the difference between L2 learner’s speech (non-native speech) and standard speech used to train ASR systems makes recognition cumbersome and leads to poor ASR performance especially for spontaneous speech (Thomson, 2013). In addition, as regards the ASR-generated caption, the recognition errors in the output often make the captions undesirable for the end-users. In such cases, even captions including less than 5% Word Error Rate (WER) – a significant performance for ASR systems – can cause distraction (Vasilescu, Adda-Decker, & Lamel, 2012).

On the other hand, instead of being constantly seen as the major drawback of ASR systems, some research implications could be considered for ASR errors. This exploratory study conducts a root-cause analysis to investigate the potential of using these errors as an inspiring source to determine difficult speech segment.

The goal of this research is to enhance the PSC3 system, which we previously developed (Mirzaei, Akita, & Kawahara, 2014). In PSC, with the aim of improving L2 learners’ listening skill, we created a smart caption that presents a principled selection of words instead of all words in order to encourage listening to the audio by restricting learners to read only for the selected words (Figure 1).

Word selection in PSC is done by referring to factors that lead to comprehension impairs i.e. by focusing on the words, which are difficult to recognize. In this view, we considered several factors such as the speech rate of the speaker, the difficulty level of the words based on their frequency of occurrence in well-known corpora and also the presence of specific words (e.g. academic terms) in speech.

---

3. Watch videos on http://www.ar.media.kyoto-u.ac.jp/psc/
Figure 1. Screenshot of PSC on a TED talk made from the original transcript “We are evolving to be a more collaborative and hearty species”

In order to enhance word selection in PSC, we need to incorporate more features to enrich the selection criteria. However, the correlation of these features is complicated. As an alternative, in this study we investigate the ASR errors to detect difficult words.

ASR-related studies often compare such systems with native speakers of the target language or with those having no knowledge of that language – Human Speech Recognition (Vasilescu et al., 2012). However, comparative analysis of ASR errors and L2 learner speech recognition would be a more prudent choice, as both parties deal with almost similar difficulties, using available resources and background knowledge. As listed in Table 1, for example, literature emphasizes the important role of high speech rate in deteriorating the performance of ASR systems (Goldwater, Jurafsky, & Manning, 2010) and impairing L2 comprehension (Rost, 2013).

Table 1. ASR errors and L2 listening difficulties

<table>
<thead>
<tr>
<th>ASR Errors</th>
<th>L2 Listening Difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation, co-articulation, speaking style, accent, age, physiology and emotions lead to ASR difficulties (Vasilescu et al., 2012).</td>
<td>Pronunciation, stress, intonation patterns and accent affect L2 listening comprehension (Rost, 2013).</td>
</tr>
<tr>
<td>Infrequent words are more likely to be misrecognized (Shinozaki &amp; Furui, 2001).</td>
<td>Infrequent words in speech correlate to complexity (Webb, 2010).</td>
</tr>
<tr>
<td>Fast speech / very slow speech increases error rates (Goldwater et al., 2010).</td>
<td>Whether too fast or too slow, speech rate can impair listening (Rost, 2013).</td>
</tr>
<tr>
<td>Word length is considered as a useful predictor of high WER (Shinozaki &amp; Furui, 2001).</td>
<td>The length of a word can affect its recognition (Field, 2008).</td>
</tr>
<tr>
<td>Open class words have lower error rate compared to closed class (Goldwater et al. 2010).</td>
<td>Learners transcribe the content words significantly better than the function words (Field, 2008).</td>
</tr>
</tbody>
</table>
Errors in automatic speech recognition versus difficulties in second language listening

2. Method

In order to conduct our analysis, we used 64 videos from the TED website (https://www.ted.com/talks). TED provides videos together with human-annotated transcript. This transcript is utilized as a reference to detect ASR errors. Next, the ASR transcripts for these talks are generated by our ASR system based on Julius 4.3 and TED models. The output is then compared with the human-annotated transcript, and the ASR errors together with their confidence measures are stored for the next step. Figure 2 depicts the schematic of this framework.

Figure 2. System schematic

Next, we classify the errors into insertion, deletion or substitution categories and perform a root cause analysis in order to extract profound features such as “speech rate”, “word frequency”, “word length”, etc. As Figure 2 presents, this framework is built on PSC system so that eventually the results can be easily compared with PSC’s output.

3. Results and discussion

In total there were 169,402 word tokens, of which 13,755 words were erroneously recognized. WER averaged 8%. These errors are categorized into:

- Substitution errors: ASR output is different from human-annotated transcript (7.0%).
- Deletion errors: ASR omits a word that exists in human-annotated transcript (0.4%).
- Insertion errors: ASR outputs a word that does not exist in human-annotated transcript (0.7%).
We then further analysed these errors to identify the features that led to their occurrences. Findings suggest that for both ASR and L2 learners, the most contributing features to difficulties are following similar trends. Figure 3 illustrates how each feature affects ASR WER. As the figure presents WER increases for too fast/slow speech rates, infrequent words and also words with shorter length.

Figure 3. ASR error trends for different features; vertical axis depicts WER (%)
4. Conclusions

This study made a comparison between ASR errors, L2 listening difficulties and PSC word selection to diagnose the words/phrases, which are hard to recognize. ASR can serve as a simplified model of a language learner. In this view ASR errors can provide useful information and introduce pedagogical implications. The findings of this research suggest that ASR errors can indicate difficult/problematic speech segments and hence can be incorporated into PSC system to better meet L2 listeners’ requirements.

References


