Automatic Speech Recognition Errors as a Predictor of L2 Listening Difficulties

Maryam Sadat Mirzaei, Kourosh Meshgi, Tatsuya Kawahara
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
maryam@sap.ist.i.kyoto-u.ac.jp

Abstract

This paper investigates the use of automatic speech recognition (ASR) errors as indicators of the second language (L2) learners' listening difficulties and in doing so strives to overcome the shortcomings of Partial and Synchronized Caption (PSC) system. PSC is a system that generates a partial caption including difficult words detected based on high speech rate, low frequency, and specificity. To improve the choice of words in this system, and explore a better method to detect speech challenges, ASR errors were investigated as a model of the L2 listener, hypothesizing that some of these errors are similar to those of language learners' when transcribing the videos. To investigate this hypothesis, ASR errors in transcription of several TED talks were analyzed and compared with PSC's selected words. Both the overlapping and mismatching cases were analyzed to investigate possible improvement for the PSC system. Those ASR errors that were not detected by PSC as cases of learners' difficulties were further analyzed and classified into four categories: homophones, minimal pairs, breached boundaries and negatives. These errors were embedded into the baseline PSC to make the enhanced version and were evaluated in an experiment with L2 learners. The results indicated that the enhanced version, which encompasses the ASR errors addresses most of the L2 learners' difficulties and better assists them in comprehending challenging video segments as compared with the baseline.

1 Introduction

Automatic speech recognition technology has formed the integral part of many language learning tools and CALL systems particularly for evaluating, training and improving L2 pronunciation and speaking skill (Neri et al., 2003; Witt, 2012; Thomson and Derwing, 2014). However, this technology has rarely been used for developing listening skill. When it comes to listening skill, instructors are often disadvantaged by the lack of readily available information about the challenges and difficulties of the audio/visual input. Therefore, they suggest the use of assistive tools, such as caption, to help the learners overcome their listening difficulties (Danan, 2004; Winke et al., 2010).

It is criticized, however, that captioning can allow the learner to comprehend the speech by reading the text even without listening (Pujolà, 2002; Vandergrift, 2011). In this view, captioning cannot serve the purpose of promoting the use of listening skill for language learners. To assist L2 learners in training listening skill, an alternative captioning tool called partial and synchronized caption (PSC) was developed (Mirzaei et al., 2016b). The system attempts to realize effective listening by presenting difficult words in the caption and hiding easy ones. PSC employs ASR technology to synchronize each word to the corresponding speech segment in order to allow text-to-speech mapping. It then selects difficult words based on the speech rate, frequency, and specificity of the words (Figure 1). These three factors are accounted for the major causes of L2 listening difficulties according to many studies (Griffiths, 1992; Révész and Brunfaut, 2013). However, there are a number of other factors such as hesitation, distortion, breached boundaries, etc. that play equal roles in making L2 listening challenging for language learners (Field, 2003; Bloomfield et al., 2010). To investigate such factors, this paper attempts to use the ASR

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 192–201, Osaka, Japan, December 11-17 2016.

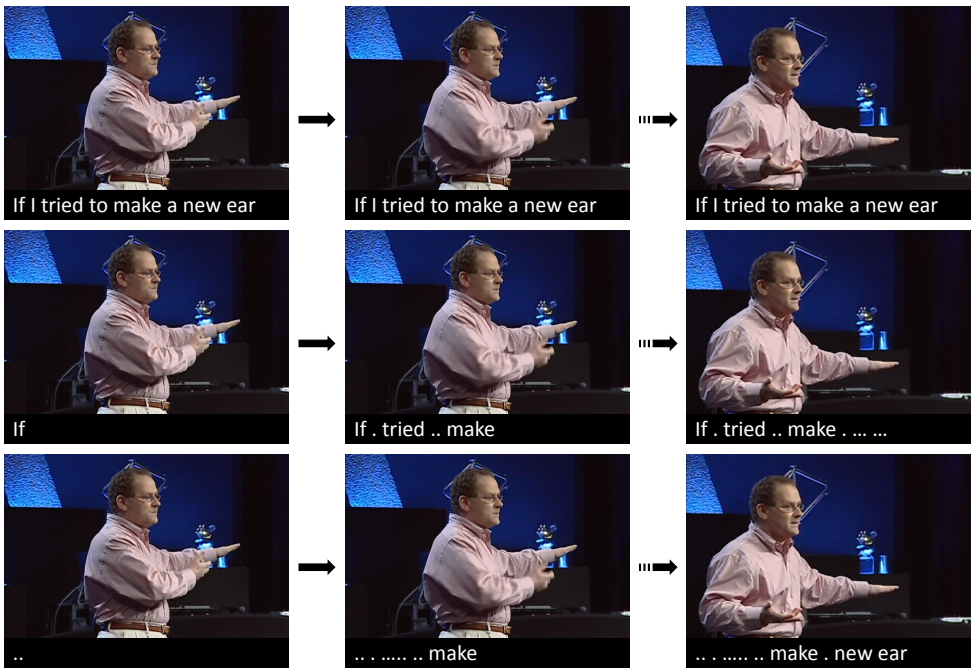


Figure 1: Caption types: (top) Full caption, (mid) Baseline partial and synchronized caption(PSC), (bottom) Enhanced PSC. TED talk by Alan Russell: The potential of regenerative medicine

technology in order to detect the problematic speech segments and to improve the choice of words in the PSC system.

Spontaneous speech such as TED talks presents numerous challenges to both ASR systems and L2 listeners. These challenges often lead to the erroneous performance of the ASR systems (Radha and Vimala, 2012; Goldwater et al., 2010). Such errors potentially involve useful information on the difficulties of speech. Accordingly, the main idea of the paper is to consider the ASR system as a potential model for L2 learners, which will encounter similar difficulties in the perception of the speech and transcription of the audio, thereby could be an indicative of areas of learner difficulties in listening.

Previous studies have compared the ASR systems performance with native and non-native speakers (but not L2 learners) of the target language. These studies are known as ASR-HSR (human speech recognition) research (Meyer et al., 2006; Scharenborg, 2007; Vasilescu et al., 2011). The majority of these studies conclude that HSR outperforms ASR especially in spontaneous speech and suggest different methods to improve the quality of the ASR output and reduce the word error rate (Moore and Cutler, 2001; Scharenborg, 2007). In this study, however, rather than considering ASR errors as major drawbacks of these systems, we are focusing on them as a rich source that elucidates the difficulties of speech. In this view, ASR errors and PSC selected words share some cases as they both refer to similar sources of difficulties. Nevertheless, the ASR errors include a wider range of factors and can introduce undiscovered features, hence can be complementary for the PSC system. In this regard, in an earlier study, we compared the ASR errors with L2 learners' mistakes on transcribing the audio in a contrastive analysis of ASR and L2SR (second language speech recognition). To this end, an in-depth error analysis was performed between the ASR systems and L2 learners (Mirzaei et al., 2015). The results of this study confirmed that some parts of the input were difficult to decode by both ASR systems and language learners and these often led to the erroneous performance of both. Thus, it was concluded that ASR errors could provide insights on detecting challenging speech segments.

To further investigate this concept, in this study, ASR errors and PSC's selected words are compared to specify the degree of overlap and differences. Through this analysis, those ASR errors, which were not included in the PSC's selected words, were detected. Next, the underlying factors that associated with the emergence of these groups of errors were investigated. Drawing upon literature, it was found that some of these root-cause factors lead to listening difficulties for L2 listeners (Field, 1998; Field, 2003). This

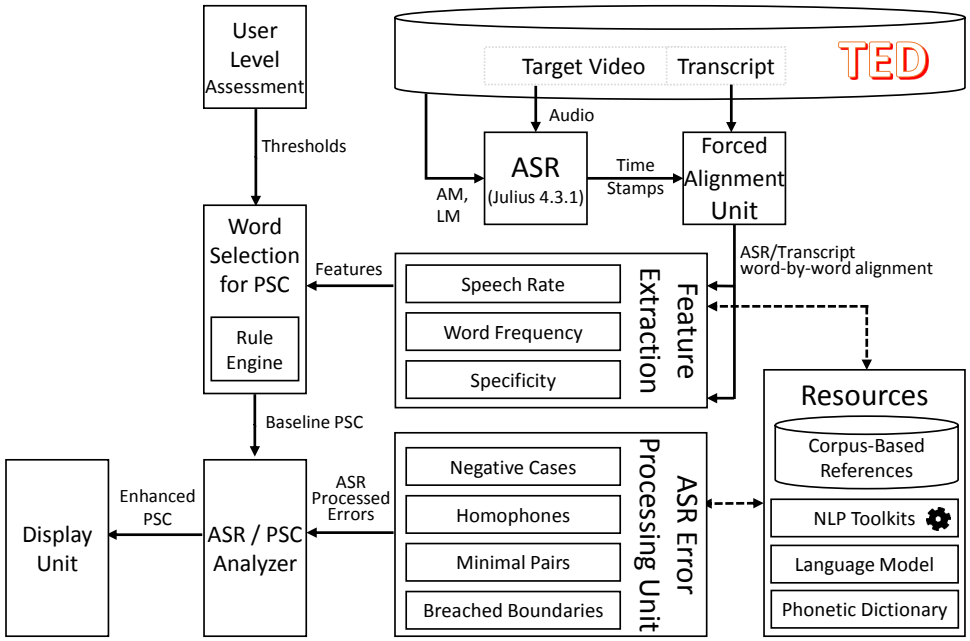


Figure 2: Schematic of the PSC system architecture. Julius ASR system was trained using 780 hours of TED talks and used to transcribe the given video. Via a word-level forced-alignment procedure, the original transcript is synchronized with the speaker's utterance. Several features are extracted for each word in the Feature Extraction module. Considering the user's proficiency level, the Rule Engine marks difficult words and generates the Baseline PSC. The ASR errors, on the other hand, signals other types of listening difficulties, which are extracted using ASR Error Processing Unit. These instances are included to the Baseline PSC to generate the Enhanced PSC.

indicates that these errors can provide useful hints for detecting difficult words or phrases. Accordingly, the aim of the present paper is to detect those parts of the speech that cause listening difficulties for L2 learners by attending to the ASR errors and incorporating them into PSC to provide better assistance for the L2 learners. Moreover, ASR correct cases that were contrarily detected to be difficult by PSC's features, were utilized as a signal for relatively easier segments of speech and used to remove too easy words from the PSC (Figure 2). The enhanced PSC will then build on the baseline system, employing the ASR clues. This enhanced version aims to outperform the baseline PSC in providing essential clues for recognition of the listening tasks to the L2 learners.

2 Method

52 TED talks (approximately 15 hours) were used in this study. `Julius 4.3.1` ASR system (Lee and Kawahara, 2009) was employed to generate the transcripts for these talks. This ASR system was trained using 780 hours of TED talks based on the lightly-supervised training approach (Naptali and Kawahara, 2012). Human-annotated transcripts for these talks were readily available from the TED website and used to evaluate the ASR output. Using forced alignment method, the two transcripts were aligned in word-level to enable error detection. Next, the errors were classified into deletion, insertion and substitution categories as shown in Table 1. ASR error rate was 19.8% with the majority of errors belonging to substitution categories.

2.1 Comparison of ASR Output and PSC Selection

PSC was generated for these videos controlling for high speech rate, low frequency, and specific or academic words. The selected words to be shown in the PSC based on the above categories were compared with the ASR errors to find the degree of overlap. We assumed that ASR errors and PSC's selected words should share some cases, as both refer to the difficult words. Table 2 indicates that 3.7 percent of the

Categories	Frequency	(%)
Total Words	145,663	
ASR Correct	116,807	(80.2%)
ASR Errors	28,856	(19.8%)
ASR Error Substitution	24,269	(16.7%)
ASR Error Insertion	2,928	(2.0%)
ASR Error Deletion	1,659	(1.1%)

Table 1: ASR error analysis

	ASR Correct (80.2%)	ASR Errors (19.8%)
PSC shown words (17.7%)	**(a)** 14.0%	**(b)** 3.7%
PSC hidden words (82.3%)	**(c)** 66.2%	**(d)** 16.1%

Table 2: ASR performance versus baseline PSC's choice of words. In **(a)** PSC categorized the words as difficult cases based on its three features, but ASR managed to correctly recognize them. Cell **(b)** indicates difficult sections according to both PSC and ASR, whereas cell **(c)** accounts for easy regions of the speech. Cell **(d)** counts the challenging words for ASR that PSC missed to select. The proposed enhancement on PSC aims to utilize ASR errors for hiding too easy words and showing the difficult words that the baseline PSC missed, i.e., to move words from **(a)** to **(c)** and from **(d)** to **(b)**.

cases are common between ASR errors and PSC shown words, however, 16.1 percent of the ASR errors could not be explained by PSC's features.

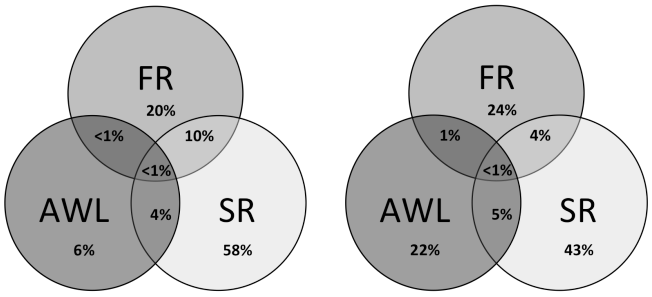
To better understand the results, the ASR errors were further analyzed taking PSC's features into account, i.e. frequency, speech rate and specificity features. Similar to PSC, the speech rate of the ASR errors were calculated in syllables per second, the frequency was estimated based on the corpus of contemporary American English - COCA (Gardner and Davies, 2013) and the specific words were detected by referring to the Academic Word List (Coxhead, 2000) and academic corpus of COCA. Figure 3 illustrates the comparison between ASR errors with PSC selection.

Figure 3(a) depicts the distribution of the mutual cases between the ASR errors and PSC's selected words based on the speech rate, frequency, and specificity. As the figure suggests speech rate is the primary factor that selects the words for PSC and is also the major factor that leads to the emergence of the ASR errors (58%). The frequency factor shows 20 percent of overlap between the PSC shown words and the ASR errors. Finally, specific words are by default set to be always shown in the PSC system, yet only a small number of these words cause ASR errors (6%).

2.2 Analysis of ASR Correct Cases

While our assumption is that ASR errors can indicate problematic speech segments for L2 listeners. ASR correct cases can specify easy items, which may not be necessarily needed to appear in PSC. Looking back to Table 2, our analysis reveals that 14.0% of the ASR correct cases are categorized as difficult words by PSC and are shown in the caption. To explore these cases, we made similar analysis considering the speech rate, frequency and specificity factors. Figure 3(b) demonstrates how these three features are responsible for words in this category, i.e. PSC shown words and ASR correct output. As the figure shows, speech rate (43%) is still the main reason to show these words in PSC, while many of the words brought by the speech rate factor could be correctly transcribed by our ASR system; indicating that these words were not too difficult. Examples include the words such as *one, every, open, look,* etc., which have high frequency and can be simply excluded from PSC without causing a barrier for L2 listeners. Thus, the findings notify the importance of refining our speech rate threshold on ASR correct cases to prevent the inclusion of difficult cases in PSC.

On the other hand, our investigation revealed that few instances of the words shown in PSC based on the frequency feature seem to be unnecessary. For instance, words such as *dystopia, piggybacking,*



(a) ASR errors versus PSC shown words

(b) ASR correct versus PSC shown words

Figure 3: Feature analysis of PSC shown words regarding the correctness of ASR output

pandemic, larceny, abyss, could be correctly transcribed by the ASR, but are infrequent to many L2 listeners and hence likely to be unknown.

Based on our findings, while specific words are always shown in PSC, many of them are not infrequent. For example, words such as *positive, science* and *research* are categorized as academic terms. However, these words are very frequent and the majority of L2 listeners should have no problem with them. Likewise, highly frequent Proper nouns (e.g. *China* and *Obama*) could be simply omitted from or repeated less in the PSC. Meanwhile, our ASR system could also correctly transcribe these words. These findings suggest that frequency of specific words and proper nouns should be considered when deciding on their inclusion in PSC.

2.3 Analysis of ASR Erroneous Cases

The strategies to refine the choice of words in PSC based on excluding easy cases are taken for granted in this study. However, a thorough investigation is needed to ensure that segments including ASR errors are actually difficult for L2 learners. Our earlier experiment with L2 listeners, who were asked to transcribe ASR erroneous cases, revealed that learners have substantial difficulties in transcribing segments that include ASR errors (Mirzaei et al., 2016a).

Based on these results, we augmented the baseline PSC to automatically detect these groups of errors (ASR errors not shown in PSC). We have limited the scope of the automation in this stage to the nine categories of **homophones, minimal pairs, negatives, breached boundaries, verb inflections, determiners, prefixes/suffixes, possessives and plural cases**, since the root-causes of other ASR errors were difficult to discover, hence these categories were discarded in this study. Our earlier study (Mirzaei et al., 2016a) showed that from among all categories, four of them are beneficial for L2 learners: homophones, minimal pairs, breached boundaries and negative cases.

Verb inflections and prefix/suffix derivations of the words are detected using word lemmatizers and COCA word stem list, while language-specific grammar rules are used to detect possessives and plurals. Detecting homophones and minimal pairs rely on word-to-phone mappings empowered by CMU Pronouncing Dictionary. The dictionary allows mappings from words to their pronunciations in the ARPAbet phoneme set, which in turn enables us to detect homophones and minimal pairs. Homophones are defined as two words with different writings, but identical ARPAbet transcripts (e.g., *blue* /B L UW/ and *blew* /B L UW/). In the case of ASR substitution errors, we select the closest pronunciation of a word in the transcript to its ASR-hypothesized utterance. Some special cases are handled, for instance, American and British spelling differences were considered.

Two words are minimal pairs if their phonetics has a Levenshtein distance of one, and they have distinct meaning. This distance enables the detection of different types of minimal pairs: initial consonant (e.g., *rot* /R AA T/ and *lot* /L AA T/), vowels (e.g., *pen* /P EH N/ and *pan* /P AE N/), and final consonant (e.g., *hat* /HH AE T/ and *had* /HH AE D/). Miniml pairs are found to be difficult to distinguish for the L2 learners (e.g., *lay* /L EY/ and *clay* /C L EY/) (Weber and Cutler, 2004). This

category also includes the third person in the present tense and past tense for regular verbs, which we preferred to exclude from the list.

Wrong boundary detection can be attributed to numerous factors, many of which are rare or hard to regulate. In this sense, some of the most studied phenomena are considered that may lead to wrong boundary detection in L2 listening. One of these rules is the frequency rule (Field, 2008; Cutler, 1990), which is built upon the idea that listeners tend to associate what the speaker says to high-frequency words when the speaker actually uses less frequent or unknown word (e.g., *dusty senseless drilling → thus he sent his drill in*). This is in line with what happens in ASR system when facing out-of-vocabulary words (Chen et al., 2013). To implement this rule, the average of ASR-hypothesized error phrase on the false boundary is calculated and compared to the average of the original phrase. Function words are excluded from calculations for having an excessive high frequency, as argued in (Cutler, 1990).

Another pattern that elicits breached boundary is a special arrangement of strong and weak syllables within one word or consecutive words (Cutler, 1990). Strong syllables typically signal for the beginning of the word, which explains why learners tend to insert a boundary when they encounter a strong syllable (e.g., *it was illegal → it was a legal*). On the other hand, learners often remove the boundaries and combine the words when they encounter a week syllable (e.g., *paint with a brush → paint without rush)*. Attending to these rules, we tried to detect such boundary cases in our algorithm.

Resyllabification (Field, 2003) is another rule which often refers to the attachment of final consonant to the following syllable (e.g., *last hour → glass tower*).

Assimilation (Cruttenden, 2014) is another common phenomenon, in which one sound becomes similar or more like a nearby sound (e.g., *did you go? → di due go*). Assimilation patterns are restricted in English. Therefore, we followed the standard patterns in (Cruttenden, 2014) to detect such cases.

Accordingly, we calculate the Levenshtein distance of the phonetic representations of the transcript and ASR-hypothesized phrase. If the distance doesn't exceed a pre-defined threshold (e.g., four differences), we proceed to examine the frequency of the words. If the flagged ASR error involved words with higher frequency (obtained by COCA corpus), then we may have a breached boundary. Along with the frequency check, we draw upon the stress pattern, syllabification, and assimilation rules to precisely detect the misrecognized boundary cases.

Finally, there are numerous situations where acoustic artifacts and speaker disfluencies or high speech rate prevent the listener from hearing the words accurately. Negative forms, in this regard, are the most likely ones to be misrecognized, while they are important to distinguish for understanding the meaning of the sentence. Given the difference between *can and can't*, for example, is a subtle one, these cases are frequently misrecognized by many L2 listeners. Other instances such as *legal and illegal* are equally important as their misrecognition can thoroughly change the meaning. To assist learners in this regard, we detected all the negative forms, which were among the ASR errors to include in the PSC system.

3 PSC Enhancement

To enhance the PSC system we apply the clues derived from the investigation of ASR errors and PSC selected word. The analysis provided us with useful insights on removing easy cases from the baseline PSC in order to provide room for inclusion of more difficult words in the enhanced version. To this end, the frequency of specific words (academic terms) was taken into account in the enhanced version. Referring to COCA academic word corpus, we could retrieve the frequency of these words and by determining a threshold based on an expert suggestion, we aimed to exclude trivial specific words in the enhanced version. The same strategy was used for highly frequent proper nouns *(e.g. America, Paris)*. In the enhanced PSC, the frequency of these cases is based on the frequency of their occurrences in TED corpus. Another enhancement on the exclusion of the easy cases regards the speech rate threshold, which was refined by defining a secondary threshold. In the case of ASR correct output, a stricter threshold was applied to prevent the speech rate factor from bringing in too many easy words. Through these improvements, many of the easy cases were excluded from the enhanced version. In the next step, difficult cases, which were detected based on ASR erroneous output were embedded into the PSC. These cases, which include homophones, minimal pair, breached boundaries and negatives were automatically detected and

incorporated into the PSC system to foster L2 listening for the learners. Table 3 presents the distribution of shown and hidden words in the enhanced PSC based on ASR correct and erroneous cases. As the table shows, using ASR clues, we could successfully move some of the easy cases (3.6%) from **(a)** to **(b)**, i.e exclude trivial words from PSC shown to PSC hide. More importantly, our enhanced version encompasses more of the ASR errors (cases were moved from **(d) to (b)** to provide an effective scaffold for L2 listeners (7.1%). It should be noted that a comparable number of shown words in both the baseline and enhanced version were maintained (the enhanced version includes an even lower amount of words) in order to enable a fair comparison.

	ASR Correct (80.2%)	ASR Errors (19.8%)
PSC shown words (17.5%)	**(a)** 11.4%	**(b)** 7.1%
PSC hidden words (82.5%)	**(c)** 69.8%	**(d)** 12.7%

Table 3: ASR performance versus Enhanced PSC's choice of words

4 Experimental Evaluation

While some improvement could be anticipated based on these enhancements, an experiment was conducted in order to confirm our hypothesis that the enhanced PSC is more helpful to the learners than the baseline.

4.1 Participants

The participants of this study were 38 Japanese and Chinese students who enrolled in CALL courses at Kyoto University. They were undergraduates, majoring in different fields such as engineering, law, science, etc. All participants had TOEIC ITP scores between 450 to 560, notifying that they were beginners to pre-intermediates. There were 8 females and 30 male students.

4.2 Material

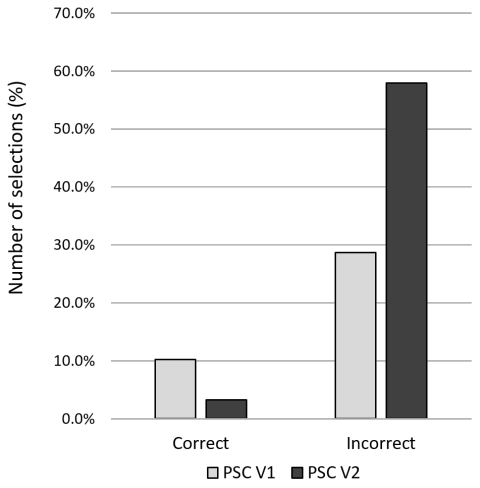
20 TED videos were selected for this experiment. All talks were delivered by native American English speakers. We excluded the effect of other accents such as British English in our experiment. From these videos, those segments in which there was a difference between the baseline and the enhanced PSC were extracted. These segments included ASR error – PSC hide cases i.e. they involved one of the four categories of minimal pair, homophone, breached boundary, and negative.

4.3 Procedure

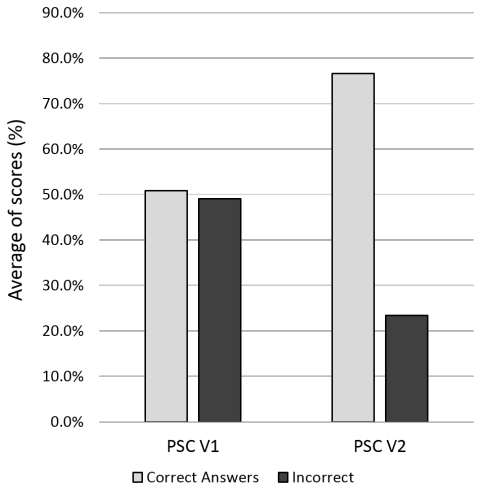
Two type of questions were designed for this experiment:

Transcribing: The participants were asked to watch a series of video segments, each segment lasting from 25 to 35 seconds. Once a video paused, the participants were supposed to transcribe the last few words that they have heard. Each segment ended with 4∼6 words, including the target word(s), all replaced by a blank. Videos were paused at irregular intervals and the participants were not aware of the exact time of the pause in order to simulate real-life listening. Moreover, the target word(s) was among the words to be transcribed and the participants had no clue about it. A timer was set for each question and the participants were supposed to transcribe the words right after the pause without having excess time to overthink or analyze, but to type down exactly what they have recognized. It was anticipated that immediately after the transcription, the participants could recognize their difficulties and misrecognitions, therefore immediately after transcribing the audio, the participants received two types of captions (baseline and enhanced PSC) to choose from. They were supposed to select the caption that included more of their misrecognized words, i.e., the ones that can better assist them to overcome their listening difficulties. Both the baseline and the enhanced versions included the same number of words, but different choices to make a fair comparison.

Paraphrasing: To make a more quantitative analysis, students were randomly assigned into two groups. One group received the baseline PSC, whereas the other one received the enhanced version.



(a) Part 1: Transcribing and choosing from baseline and enhanced PSC – the graph shows the number of participants who chose baseline PSC (PSC V1) versus Enhanced PSC (PSC V2).



(b) Part 2: Paraphrasing based on the type of caption received in each group: baseline (PSC V1) vs. enhanced (PSC V2) – The graph shows the paraphrasing scores in the two groups.

Figure 4: Experimental results

They were asked to watch the videos with the assigned type of caption and paraphrase the last sentence of each segment whenever the video was paused. Paraphrase test focuses on the recognition of a specific part of listening material thus the participants' answers on paraphrasing the last segment are based on the caption clues they have received in each group. It was assumed that the group who received the enhanced PSC had better hints to disambiguate the sentence and select the best paraphrasing choice.

5 Results

The results of this experiments should be explained in two parts: (1) the number of times the enhanced PSC was chosen over the baseline PSC after the transcription task (qualitative analysis) and (2) the scores of the participants on paraphrasing the last sentence based on the type of the caption they had received (quantitative analysis).

Figure 4(a) shows the analysis of the results on participants' preferences regarding the selection between the baseline and the enhanced PSC. As the figure shows the number of times the participants preferred the enhanced PSC is significantly higher than the baseline version (61% versus 39%).

Our quantitative analysis on the participants' paraphrasing scores illustrated in figure 4(b) demonstrates that the students in the enhanced PSC group gained statistically higher scores than their peers in the baseline group (based on a t-test analysis). While the baseline PSC group answered the questions chance-like (50.9% correct versus 49.1% incorrect), the enhanced PSC group could choose the correct answer 76% of the time.

The results reveal that some of the ASR errors signal problematic speech segments and ASR clues could be used for facilitating recognition and comprehension of the input. Moreover, our findings identified that some improvements are realized in the enhanced PSC, which makes it more preferable to the L2 listeners. Furthermore, the enhanced version, which encompasses useful ASR errors, can better assist the L2 listeners compared to the baseline.

6 Conclusion

The study investigated the use of ASR errors in detecting problematic speech segments and improving the word selection criteria in PSC. Following a thorough analysis, it was found that several categories in the ASR errors signal the difficulties for the L2 listeners. These categories include homophones, minimal pairs, negative, and breached boundaries. Our baseline PSC system was extended to detect these cases

and generate the enhanced PSC drawing upon these clues. On the other hand, analysis of these errors provided some insights on how to refine PSC's selection by omitting too easy cases.

Experiment with the L2 listeners confirmed this hypothesis that some of these errors can predict L2 listening difficulties; hence the enhanced PSC, which includes these cases, can effectively assist the L2 listeners. To conclude our in-depth analysis on the ASR errors revealed that ASR systems epitomize a model of an L2 listener, shedding light on both easy and difficult words and phrases in the input, which can be directly used to enhance the PSC system in order to foster L2 listening.

References

Amber Bloomfield, Sarah C Wayland, Elizabeth Rhoades, Allison Blodgett, Jared Linck, and Steven Ross. 2010. What makes listening difficult? factors affecting second language listening comprehension. Technical report, DTIC Document.

Wei Chen, Sankaranarayanan Ananthakrishnan, Rohit Kumar, Rohit Prasad, and Prem Natarajan. 2013. Asr error detection in a conversational spoken language translation system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7418–7422. IEEE.

Averil Coxhead. 2000. A new academic word list. *TESOL quarterly*, 34(2):213–238.

Alan Cruttenden. 2014. *Gimson's pronunciation of English*. Routledge.

Anne Cutler. 1990. Exploiting prosodic probabilities in speech segmentation.

Martine Danan. 2004. Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal des traducteursMeta:/Translators'Journal*, 49(1):67–77.

John Field. 1998. Skills and strategies: Towards a new methodology for listening. *ELT journal*, 52(2):110–118.

John Field. 2003. Promoting perception: Lexical segmentation in l2 listening. *ELT journal*, 57(4):325–334.

John Field. 2008. Bricks or mortar: which parts of the input does a second language listener rely on? *TESOL quarterly*, 42(3):411–432.

Dee Gardner and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, page amt015.

Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Roger Griffiths. 1992. Speech rate and listening comprehension: Further evidence of the relationship. *TESOL quarterly*, 26(2):385–390.

Akinobu Lee and Tatsuya Kawahara. 2009. Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pages 131–137. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee.

Bernd Meyer, Thorsten Wesker, Thomas Brand, Alfred Mertins, and Birger Kollmeier. 2006. A human-machine comparison in speech recognition based on a logatome corpus. In *Speech Recognition and Intrinsic Variation Workshop*.

Maryam Sadat Mirzaei, Kourosh Meshgi, Yuya Akita, and Tatsuya Kawahara. 2015. Errors in automatic speech recognition versus difficulties in second language listening. In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, page 410. Research-publishing.net.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Tatsuya Kawahara. 2016a. Leveraging automatic speech recognition errors to detect challenging speech segments in ted talks. In *CALL Communities and Culture-Proceedings of the 2016 EUROCALL Conference, Limmasol, Cyprus*. Research-publishing.net.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Tatsuya Kawahara. 2016b. Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill. *ReCALL (in press)*.

Roger K Moore and Anne Cutler. 2001. Constraints on theories of human vs. machine recognition of speech. In *Workshop on Speech Recognition as Pattern Classification (SPRAAC)*, pages 145–150. Max Planck Institute for Psycholinguistics.

Welly Naptali and Tatsuya Kawahara. 2012. Automatic transcription of ted talks. IWSLT.

Ambra Neri, Catia Cucchiarini, and Wilhelmus Strik. 2003. Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS*, pages 1157–1160.

Joan-Tomàs Pujolà. 2002. Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, 14(02):235–262.

V Radha and C Vimala. 2012. A review on speech recognition challenges and approaches. *doaj. org*, 2(1):1–7.

Andrea Révész and Tineke Brunfaut. 2013. Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35(01):31–65.

Odette Scharenborg. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Ron I Thomson and Tracey M Derwing. 2014. The effectiveness of l2 pronunciation instruction: A narrative review. *Applied Linguistics*, page amu076.

Larry Vandergrift. 2011. Second language listening. *Handbook of research in second language teaching and learning*, 2:455.

Ioana Vasilescu, Dahbia Yahia, Natalie D Snoeren, Martine Adda-Decker, and Lori Lamel. 2011. Cross-lingual study of asr errors: On the role of the context in human perception of near-homophones. In *INTERSPEECH*, pages 1949–1952.

Andrea Weber and Anne Cutler. 2004. Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1):1–25.

Paula Winke, Susan Gass, and Tetyana Sydorenko. 2010. The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, 14(1):65–86.

Silke M Witt. 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. IS ADEPT*, 6.