Detecting listening difficulty for second language learners using Automatic Speech Recognition errors

Maryam Sadat Mirzaei[1], *Kourosh Meshgi*[1], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Japan

mirzaei@sap.ist.i.kyoto-u.ac.jp, meshgi-k@sys.i.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Abstract

This paper introduces a new approach to detect difficulties in speech for second language (L2) listeners using automatic speech recognition (ASR) systems. In this study, the ASR systems are viewed as a model to predict L2 learners' listening difficulties and the ASR erroneous cases are analyzed to find useful categories of errors that can epitomize language learners' transcription mistakes. Annotation of the ASR errors revealed the usefulness of several categories in predicting learners' listening difficulties when watching TED videos delivered by American native speakers. Experiments with L2 learners of English confirmed that these categories lead to listening problems for the majority of the learners. One application to make use of these errors can be found in partial and synchronized captioning (PSC), in which only difficult words are selected and shown to facilitate listening, while easy words are hidden. Findings of the experiments attested that embedding the useful categories of the ASR errors into PSC improves learners' comprehension.

Index Terms: automatic speech recognition, second language listening, error analysis, partial and synchronized caption

1. Introduction

For majority of the L2 learners, listening to authentic contents, which are created by native speakers (not specially designed for language learning purposes) is a very demanding task. There are many different sources of listening difficulties involved varying from lexical, speech-related factors to linguistic-related complications [1]. Among these, some speech related factors, such as the speech rate and perceptual difficulties are known as the prominent sources of problem for many language learners [2, 3]. When it comes to the automatic recognition of speech, ASR systems are also subjected to some errors, some of which stems from similar factors [4]. While human listeners have little difficulties in dealing with recognition of spoken language in acoustically challenging situations, ASR systems often lack the same robustness that is achieved by the humans [5]. This observation has been the source of motivation for the studies that investigated the ASR errors and HSR (human speech recognition) difficulties with the purpose of bridging the gap between the two and incorporating HSR findings to improve ASR performance [6, 7, 8, 9, 10]. The subjects of these studies are either a native speaker of the target language or non-native speakers with no knowledge of the target language (e.g., Japanese with no knowledge of French tested with French audio, which includes words with the maximum phonetic similarity between the two languages). Through such studies, the researchers attempt to improve the ASR performance and eliminate the ASR errors [11]. In this paper, however, we assume that some of the speech related difficulties, which prevent L2 learners from recognizing a speech can lead to the emergence of the ASR errors.

There are numerous factors accounted for L2 listening difficulties, some of which are also observed when investigating ASR systems' errors. For instance, fast or too slow speech rate increases ASR error rate [12], similarly too fast or too slow speech rate can hinder L2 listening comprehension [2]. Infrequent words are likely to be misrecognized by ASR systems [13] and also cause perceptual complexity for L2 learners [14]. The length of the word serves as a useful predictor of ASR errors [13], while strongly affects L2 listening recognition [15]. Automatic recognition of male speakers are more challenging for ASR systems [16], and L2 learners also find it more difficult to recognize male utterances [17]. Finally, perceptual difficulties in speech impede the recognition of both ASR systems and L2 listeners [3, 4].

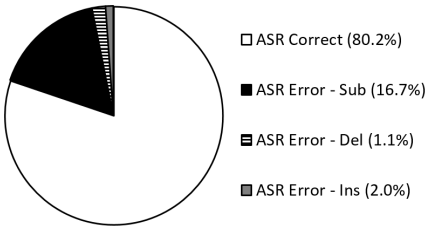

Figure 1: *ASR error statistics of 52 TED talks*

The effect of speech rate, word frequency and word length are investigated in [18], therefore these factors are not considered in this study. The gender of the speaker and class of the words are very broad predictors of the ASR errors, thus, excluded from this study. Therefore, we investigate the perceptual difficulties in the speech focusing on ASR error categories that signals L2 learners' listening difficulties.

2. ASR Predicts L2 Listening Difficulties

2.1. ASR Error Analysis

In this study, 52 TED Talks (∼15 hours) were annotated by *Julius ASR 4.3.1* [19], which is pre-trained on 780 TED talks using a lightly-supervised approach [20]. The ASR transcripts were compared with human annotations (available from TED website) using word-level alignment. The reason to select a trained ASR is to obtain a reasonable amount of ASR errors to analyze. For the same reason, only 1-best ASR hypothesis is used for error detection. Figure 1 shows that in total we had 16.7% ASR errors while the majority of the errors belongs to the substitution category. Since we are interested in misrecognition of words, our main focus will be on the substitution category.

2.2. Root-Cause Analysis

We performed a root-cause analysis on the ASR error substitution cases and found the following clusters: *(i)* homophones, *(ii)*

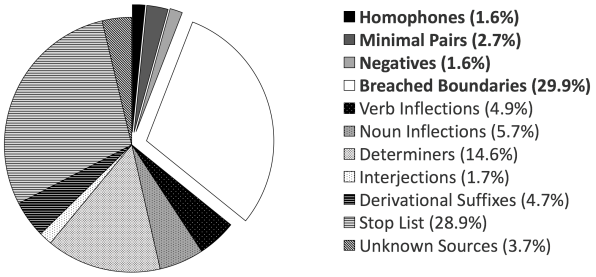

Figure 2: *Ratio of different categories in ASR substitution errors obtained from transcribing 52 TED videos.*

minimal pairs, *(iii)* negatives, *(iv)* breached boundaries, *(v)* verb inflection, *(vi)* noun inflections, *(vii)* determiners, *(viii)* interjections, *(ix)* derivational suffixes, *(x)* stop list, and *(xi)* unknown sources. The distribution of these errors are shown in Figure 2.

We labeled the ASR substitution error cases, as *useful (i)* if a similar misrecognition could be expected by L2 listeners and *(ii)* if providing the learners with these cases in the form of a caption can facilitate their recognition. Around 10% of the words are annotated by another annotator based on the same criteria to obtain the annotation agreement. We found a very high level of agreement (Cohen's $\kappa = 0.81$), which refutes the subjectivity of the annotation. Figure 3 presents the usefulness ratio of each category based on the annotation results. The figure suggests that minimal pairs, homophones, negatives, and breached boundaries are the most useful categories of the ASR errors. This is in line with the findings of the studies on L2 listening difficulties [3, 21] and makes these categories of ASR errors the potential predictors of L2 listening difficulties.

2.3. Automatic Categorization of ASR Errors

We developed an ASR error analysis unit, which uses syntactical analyzers, large-scale corpora, and phonetic dictionaries to determine the categories of ASR errors. For each ASR error case, the ASR transcript and the original transcript are aligned and checked for the word pairs that can be confused with each other. Word lemmatizers, language-specific grammar rules, and COCA corpus [22] were used to detect verb/noun inflections, determiners, interjections, and derivational suffixes.

Homophones and minimal pairs are detected by comparing the phone transcription of the utterance (by ASR) and the transcript (using CMU phonetic dictionary). Homophones are words with different writings, but identical phone sequence (e.g., *feet* and *feat* /F IY T/). An exception to this rule is cases such as American and British spelling mismatches, that are handled in our implementation. Minimal pairs are the words whose phone sequences differ only in one phonological element (e.g., *fund* /F AH N D/ and *fun* /F AH N/). To detect minimal pairs the words whose phone sequences have a Levenshtein distance of one are considered.

Breached boundaries are cases in which the boundaries of the perceived utterance are converged or diverged from the correct location when compared to the transcription, thus creating new word sequences (e.g., *in close* instead of *enclose*, *thick atmosphere* instead of *to keep this fear*). Many language learners cannot set the right boundaries between the words [23], and there is no comprehensive rule to detect such cases. In the rare case, the two phrases have identical phone sequences while the boundaries and the resultant words are different. The following four cases are derived from the linguistic studies that focused on language learners' boundary misrecognition by investigating many cases that were misrecognized by the language learners. We found these categories very useful in detecting the major breached boundary cases:

Higher Frequency: when the speaker uses less-frequent or out-of-vocabulary words, the listeners tend to associate the uttered words to high-frequency words, which are generally more familiar to them [24], similar to what happens in ASR systems when facing such words [25], e.g., *achieve her way* is heard as *a cheaper way*. To detect such instances of breached boundaries, the average of the frequency of the words in both the ASR and the original transcript are calculated and compared. In this calculation, function words –that have excessively high frequency– are excluded [24].

Stress Syllables: strong syllables typically appear at the beginning of the words, so L2 learners tend to believe that the words begin with strong syllables. Therefore in most of the cases, they insert a boundary before a strong syllable, e.g., *the skies* instead of *disguise*. On the other hand, learners tend to merge the word starting with a weak syllable to the previous or the next word, e.g., *twenty two* instead of *ten to two*. Based on these findings detailed in [24], potential breached boundaries can be detected between the sequence of words in the ASR and the original transcripts.

Resyllabification: learners tend to attach the final consonant of the word to the beginning of the next word [3] and create false boundaries, e.g., *made out* instead of *may doubt*.

Assimilation: in this phenomena a sound morphs into a similar/neighbor sound in special patterns [26]. In some languages such as English these patterns are regular, and can be easily encoded into the system, e.g., *Sam which* instead of *sandwich*.

Among other situations where acoustic and speech artifacts impede word recognition for ASR and L2 learners, negative forms have the most influence over the comprehension of the speech. Negative form of modals (e.g., *can* instead of *can't*) and negative prefixes (e.g., *legal* instead of *illegal*) are detected as negative cases to address the language learners' difficulties.

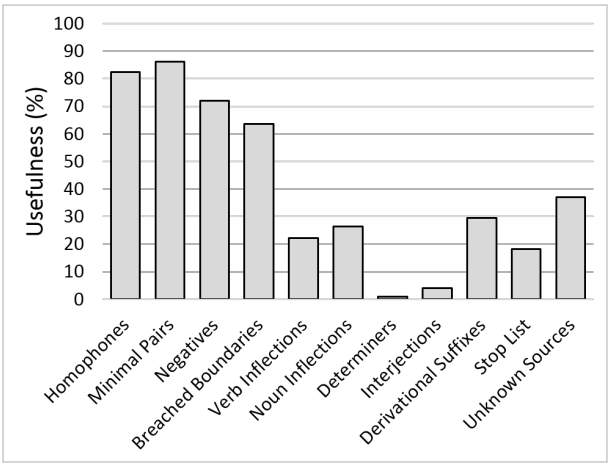

Figure 3: *The usefulness of patterns of ASR Error–PSC Hidden category for substitution errors. The usefulness is calculated for each category considering the number of words labeled as useful by the annotator to all words of the category.*

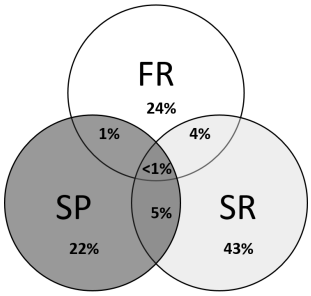

Figure 4: *Feature statistics of ASR Correct–PSC Shown. FR, SR and SP denotes word frequency, speech rate, and specificity, respectively.*

3. Addressing L2 Listening Difficulties

In the following section, an overview of the Partial and Synchronized Caption (PSC) is provided, its advantages in addressing L2 listening difficulties are discussed, and the target ASR categories are included in this system to provide an Enhanced PSC that better assists L2 learners in the listening task.

3.1. Partial and Synchronized Caption (PSC)

To facilitate training L2 listening skill, PSC was developed[1] to present the difficult words in the caption and hide easy ones [27]. In this caption, an ASR system is employed to align the transcripts with their respective speech segments (synchronization) and difficult words are selected from the transcript based on speech rate, word frequency, and specificity (partialization). This framework strives to find the most problematic factors for L2 listening by drawing upon studies on L2 listening difficulties. By evaluating individual learner's proficiency level, this system adjusts feature parameters to realize a personalized caption for the individual learners. In addition, a stop list (including marginal words, propositions, etc.) and a repetition counter is embedded into the PSC to improve the word selection process.

The synchronization feature of PSC aids word boundary detection and promotes speech-to-text mapping. On the other hand, partialization prevents the learners from over-reliance on reading the captions and encourages them to listen more and read less. By using proper features, PSC is capable of providing the right amount of scaffold for different learners. This characteristic of the system is further enhanced by its adaptation to the learners' proficiency. Another great advantage of the PSC system is that it is fully automated.

While word frequency and specificity accounts for lexical difficulties, speech rate is the only feature in the Baseline PSC system that represents acoustic and speech aspects of the listening material [2, 28]. However, there are a number of acoustic and speech factors that may cause difficulties for L2 listening such as hesitations [29], noise [30], speaker's variations [17], and perceptual difficulties in speech [3, 31]. Among them we focus on the four target categories of ASR errors capable of predicting L2 listening difficulties: minimal pairs, homophones, negatives, and breached boundaries.

3.2. Enhancing the PSC

To improve the word selection in the Baseline PSC, we extended this framework with the ASR error analysis unit. Instead of discarding ASR errors, this unit compares the original and ASR transcripts to identify the source of the errors. If the detected ASR error source falls into one of the target categories, the word is decided to be shown in the Enhanced PSC.

To maintain the desirable textual density of the final caption, some of the most trivial words of the *ASR Correct–PSC Shown* category should be removed. Figure 4 demonstrates that most of these words are included in the PSC by speech rate feature, therefore a switching mechanism is designed to set a more strict threshold for showing such words when ASR recognized them correctly. In addition, some specific (academic) words (e.g., *research*, *positive*) are frequent in the contemporary language, hence, they could be less challenging for the learners. Therefore, a threshold based on ASR correct or erroneous cases is introduced to this category. Upon correct recognition of the specific word by the ASR, the frequency of the word is checked and if the frequency exceeds the threshold, the word will be hidden in the caption. Figure 5 compares the distribution of the word categories in the Enhanced PSC against the Baseline.

4. Experiments

To evaluate the performance of designated ASR error categories as predictors of L2 learners' listening difficulties, three different tests were conducted. The participants were 38 Japanese and Chinese undergraduate students, with TOEIC ITP scores ranging from 450 to 560 implying that their proficiency level was pre-intermediate.

The test material is taken from annotated videos, filtered for native American speakers. From these videos, the "difficult" segments involving *ASR error–PSC hidden* cases (using Baseline PSC) were selected, which contained one of the four target categories (minimal pairs, homophones, negatives, and breached boundaries). The video segments were not repeated in the experiments.

4.1. Transcription Tasks

In this experiment, a short video clip (25 to 35 seconds long) was given to the participants. The video was suddenly paused, and the participants were supposed to transcribe the last 4 to 6 words (including the target word), that included the target words. This test was timed to prevent the participants from rethinking and reformulating and no clue was given about the tar-

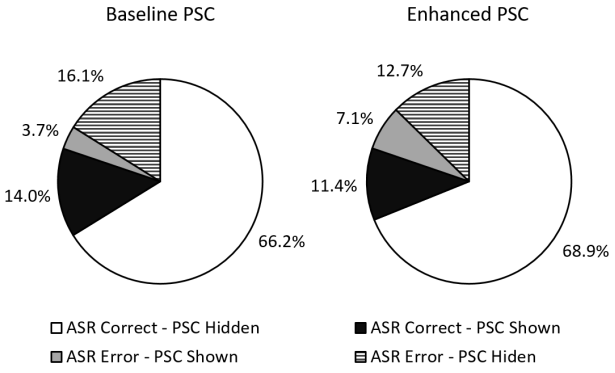

Figure 5: *ASR performance versus PSC's choice of words. Baseline PSC (left) shows 17.7% of the words. The goal of Enhanced PSC (right) is to show more of ASR Error–PSC Hidden cases and to hide more of ASR Correct–PSC Shown cases.*

[1] http://sap.ist.i.kyoto-u.ac.jp/psc/#DEMO

get word(s) or time of the pause. Through this process, we tried to check the participants' listening recognition when the video segment included an ASR erroneous case. As a control measure, some "easy" segments of the videos from *ASR correct–PSC hidden* is selected from the same video.

Table 1 demonstrates that participants' scores on the easy segments are significantly higher than the difficult segments, which included four target categories of ASR errors. As a result, the findings indicate that these four categories are challenging for the participants compared to the easy segments of the same video. It can be concluded that the participants share the difficulty with the ASR systems in transcribing homophones, minimal pairs, negatives, and breached boundaries.

Table 1: *Transcription Test: Transcription scores on difficult segments (ASR errors) vs. easy ones (ASR correct).*

Average Score in Transcription	Easy	Difficult
Homophone	81.9%	12.4%
Minimal Pairs	89.3%	14.3%
Negatives	83.3%	11.2%
Breached Boundaries	87.6%	20.0%
Total	85.2%	16.2%

4.2. Caption Selection

In this experiment, similar to the previous experiment, the participants were supposed to transcribe the last 4 to 6 heard words when the video paused. PSC was provided during the video playback, except for the last sentence. After the transcription, the participants receive both Baseline or Enhanced PSC for the final sentence, and they were asked to choose the caption which provided better clues to overcome their listening difficulties. The idea is that after transcription, the participants become aware of their difficulties and misrecognition, hence they can select the most informative choice between the Baseline and the Enhanced version. To conduct a fair comparison we ensured that both captions have a similar number of shown words.

Table 2 shows that upon encountering a problem in transcription, the participants preferred the Enhanced PSC, which includes the ASR errors. This again shows that the participants shared the difficulty in recognizing the target word(s) with ASR. In addition, the participants have selected the Enhanced PSC 61% of the times, indicating that the Enhanced version could better assist them with recognition difficulties and provided them with better choices of words.

Table 2: *Caption selection test: the preferred caption of participants with respect to their transcription correctness.*

Transcription	Baseline PSC	Enhanced PSC
Correct Transcription	10.2%	3.2%
Incorrect Transcription	28.7%	57.9%

4.3. Paraphrasing

Paraphrase tests emphasize the recognition of a specific part of the listening material. To perform this test we randomly divided the participants into two groups, one group received the Baseline PSC along with the video and the other received Enhanced PSC. They watched a short video clip (10 to 15 seconds long) and tried to paraphrase the last sentence they heard, once the video was paused (they were given two paraphrase options to choose from). The paraphrases pivoted on the target word(s), therefore selecting the wrong paraphrase choice conveys the misrecognition of the target word(s) that were chosen from the four categories of the ASR error.

Table 3 shows that given the Baseline PSC, the participants could not resolve the listening difficulty and they chose the correct and incorrect choices chance-like. On the other hand, given Enhanced PSC that included the target word(s), the participants performed significantly better. This emphasizes the role of the word selection in PSC and demonstrates that Enhanced PSC (including the ASR errors) realizes a better word selection to foster listening comprehension.

Table 3: *Paraphrasing test: the average scores of two groups of participants given Baseline vs. Enhanced PSC*

Group	Correct	Incorrect
Baseline PSC (G1)	50.9%	49.1%
Enhanced PSC (G2)	76.6%	23.4%

5. Conclusions

In this study, we introduced some categories of ASR errors as good predictors of problematic speech segments for L2 learners. An extensive analysis of the literature on the L2 listening and ASR errors indicated some similarities between the two. Furthermore, a careful investigation of the ASR substitution errors revealed that homophones, minimal pairs, negatives, and breached boundaries are among the most important categories to predict L2 learners' listening difficulties. Experimentally, we showed that the designated categories of ASR errors are able to predict some of the L2 listening difficulties. Additionally, our findings revealed that incorporating these categories into the PSC framework can lead to a significant improvement in the word selection of PSC.

The current study considers ASR as a simplified and general model of L2 learners, however, the next step would be to make ASR systems similar to L2 learners in term of listening proficiency and make them adaptive to the different levels of the learners. This can be done through degrading the ASR system so that its errors can provide more useful instances for PSC on language learners with different proficiency levels. Moreover, this framework can be extended to other languages to be used as a universal training tool for L2 listening development. This can be realized by substituting the word-frequency corpora, ASR models, and syntactic analyzers. Furthermore, with regards to the breached boundary category, the current framework focuses on the detection of most dominant cases, however, detecting all possible cases of breached boundaries in ASR errors (which are also misrecognized by L2 learners) requires more investigation.

6. References

[1] A. Gilmore, "Authentic materials and authenticity in foreign language learning," *Language teaching*, vol. 40, no. 02, pp. 97–118, 2007.

[2] R. Griffiths, "Speech rate and listening comprehension: Further evidence of the relationship," *TESOL quarterly*, vol. 26, no. 2, pp. 385–390, 1992.

[3] J. Field, "Promoting perception: Lexical segmentation in L2 listening," *ELT journal*, vol. 57, no. 4, pp. 325–334, 2003.

[4] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[5] B. T. Meyer, T. Brand, and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 388–403, 2011.

[6] R. K. Moore and A. Cutler, "Constraints on theories of human vs. machine recognition of speech," in *Workshop on Speech Recognition as Pattern Classification (SPRAAC)*. Max Planck Institute for Psycholinguistics, 2001, pp. 145–150.

[7] O. Scharenborg, L. ten Bosch, L. Boves, and D. Norris, "Bridging automatic speech recognition and psycholinguistics: Extending shortlist to an end-to-end model of human speech recognition (l)," *The Journal of the Acoustical Society of America*, vol. 114, no. 6, pp. 3032–3035, 2003.

[8] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *Speech Recognition and Intrinsic Variation Workshop*, 2006.

[9] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, 2007.

[10] I. Vasilescu, M. Adda-Decker, and L. Lamel, "Cross-lingual studies of ASR errors: paradigms for perceptual evaluations." in *LREC*, 2012, pp. 3511–3518.

[11] W. Shen, J. Olive, and D. Jones, "Two protocols comparing human and machine phonetic discrimination performance in conversational speech," in *INTERSPEECH*, 2008.

[12] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, vol. 29, no. 2, pp. 137–158, 1999.

[13] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 198–201.

[14] A. Bloomfield, S. C. Wayland, E. Rhoades, A. Blodgett, J. Linck, and S. Ross, "What makes listening difficult? Factors affecting second language listening comprehension," DTIC Document, Tech. Rep., 2010.

[15] B. Laufer, "Words you know: How they affect the words you learn," *Further insights into contrastive linguistics*, pp. 573–593, 1990.

[16] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *INTERSPEECH*, 2005, pp. 2205–2208.

[17] H. Quené, "On the just noticeable difference for tempo in speech," *Journal of Phonetics*, vol. 35, no. 3, pp. 353–362, 2007.

[18] M. S. Mirzaei and T. Kawahara, "ASR technology to empower partial and synchronized caption for L2 listening development." in *SLaTE*, 2015, pp. 65–70.

[19] A. Lee and T. Kawahara, "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*, 2009, pp. 131–137.

[20] W. Naptali and T. Kawahara, "Automatic speech recognition for ted talks."

[21] A. Weber and A. Cutler, "Lexical competition in non-native spoken-word recognition," *Journal of Memory and Language*, vol. 50, no. 1, pp. 1–25, 2004.

[22] M. Davies, "The corpus of contemporary American English: 520 million words, 1990–2015," http://corpus.byu.edu, 2008, accessed: 2013-07-04.

[23] J. Field, "Bricks or mortar: which parts of the input does a second language listener rely on?" *TESOL quarterly*, vol. 42, no. 3, pp. 411–432, 2008.

[24] A. Cutler, "Exploiting prosodic probabilities in speech segmentation." 1990.

[25] W. Chen, S. Ananthakrishnan, R. Kumar, R. Prasad, and P. Natarajan, "ASR error detection in a conversational spoken language translation system," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7418–7422.

[26] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[27] M. S. Mirzaei, K. Meshgi, Y. Akita, and T. Kawahara, "Partial and Synchronized Captioning: A new tool to assist learners in developing second language listening skill," *ReCALL*, vol. 29, no. 2, pp. 178–199, 2017.

[28] A. Révész and T. Brunfaut, "Text characteristics of task input and difficulty in second language listening comprehension," *Studies in Second Language Acquisition*, vol. 35, no. 01, pp. 31–65, 2013.

[29] G. Buck, *Assessing listening*. Cambridge University Press, 2001.

[30] J. Aydelott and E. Bates, "Effects of acoustic distortion and semantic context on lexical access," *Language and Cognitive Processes*, vol. 19, no. 1, pp. 29–56, 2004.

[31] A. Cutler, "The lexical statistics of word recognition problems caused by L2 phonetic confusion." in *INTERSPEECH*, 2005, pp. 413–416.