INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Adaptive Listening Difficulty Detection for L2 Learners Through Moderating ASR Resources

Maryam Sadat Mirzaei, Kourosh Meshgi

RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
maryam.mirzaei@riken.jp, kourosh.meshgi@riken.jp

Abstract

Teaching listening skills to those learning a second language (L2) is one of the most challenging tasks mainly because predicting L2 listening difficulties is not always straightforward. Complex processes are involved in decoding connected speech, constructing meaning, and comprehending the audio material. Many studies have attempted to identify the significant factors leading to listening difficulties, yet, a comprehensive model is to be constructed. We argue that an automatic speech recognition (ASR) system with limited training can be viewed as a rough model for an L2 listener with particular language proficiency. We proposed a method to select the training samples for the ASR system to match the mistakes of L2 listeners when listening to the authentic listening materials. This model can predict the learners' listening difficulties, thus allowing for generating tailored captions to assist them with L2 listening.

Index Terms: speech recognition, L2 listening, partial and synchronized caption, error analysis

1. Introduction

While listening to authentic material is one of the best ways to teach listening to second language (L2) learners, it often includes several cases of misrecognition by the learners. Such listening mistakes happen due to learners' limited vocabulary, scarce encounters with advanced grammars or complex sentence structures, and limited exposure to listening materials obtained in the wild. So far, language teachers rely on their experience to predict the listening difficulty of the learners. Yet, there is a growing need for automatic tools that can predict L2 listening difficulties and address them.

Mirzaei et al. [1] posit that errors of automatic speech recognition (ASR) systems can emulate the L2 listeners' difficulties in listening. This idea is based on the contrastive analysis between the sources of ASR and L2 listening difficulties (Table 1). In this sense, ASR errors can be used to signal the potential L2 listening difficulties in a listening task. These errors not only involve cases with phonetic similarities, which indicate the words whose phonetics are similar in a target language but also include the cases of perceptual difficulties, which are the focus of this study. Words could be uttered in a way that makes recognition/ perception difficult and could be misrecognized with another word/words due to the way the speaker adjusts pitch, uses liaison, etc., thus leading to confusion and breached boundaries (e.g., "*it was an eagle*" heard instead of "*it was illegal*"). Moreover, for both ASR systems and L2 learners with limited proficiency, out of vocabulary words are not only miss-recognized as in-vocabulary words with similar phonetics, but such errors also cause further problems in recognizing nearby words. In such cases, ASR could be very helpful in indicating the problematic regions.

Table 1: *Sources of difficulties for ASR and L2 listeners*

⋆ **Lexical**
ASR: Infrequent words are more likely to be misrecognized. [2]
L2: The occurrence of infrequent words is correlated to complexity [3].
ASR: Word length is a useful predictor of higher error rates. [2]
L2: The length of a word has a strong effect on its recognition for the learners. [4]
ASR: Open class words (N. and V.) cause fewer errors compared to the closed class (Prep., articles). [5]
L2: Recognition of content words is easier than function words and nouns predominate over predicates/verbs. [6, 7]
⋆ **Acoustic, Speech, Perceptual**
ASR: Fast or very slow speech rate raises the ASR errors [2, 8].
L2: Whether it is too fast or too slow, speech rate can deteriorate L2 listening and increases difficulties. [9]
ASR: Co-articulation, pronunciation, speaking style, disfluencies, accent, age, physiology, and emotions of speakers lead to the ASR difficulties. [10]
L2: Pronunciation can be unclear due to assimilation, reduction, etc. Stress, intonation patterns, and accent affect L1 and L2 listening. [3, 11]
ASR: Male speakers cause more problems for ASR systems than female speakers [12].
L2: Understanding male speakers who generally have faster articulatory rates is more difficult for L2 listeners. [13]
ASR: Ambiguity in the speech, such as the occurrence of homophones or ambiguous word boundaries are the factors that lead to recognition difficulties for the ASR systems. [14]
L2: Phonological neighbors and words with identical pronunciation make L2 recognition hard [15].
Assimilation, reduction, etc., leads to breached boundaries and attenuate L2 listening [16].

Many efforts have been dedicated to detecting ASR errors automatically. Decoder-based error detectors [17, 18] use features generated from the ASR decoder, such as confidence scores, linguistic information, confusion networks, phonetic acoustic model (AM) deviation, part-of-speech, and homophone indicator. On the other hand, non-decoder-based approaches [19, 20] employ features from other sources such as binary match with another ASR output detected topic, bigram hits on different corpora, or emergent lexical and acoustic features in recurrent neural networks [20]. ASR error rate is also predicted from phonemic confusion [21], acoustic word embedding [22], and speech intelligibility [23].

Bridging the gap between ASR errors and human errors is also another challenge. Word error rate (WER) as the primary metric for assessing ASR performance is not a satisfactory proxy for human intelligibility [24, 25]. Therefore other

http://dx.doi.org/10.21437/Interspeech.2021-372

metrics (e.g., Goodness of Pronunciation [26], e-WER [27]) and systems (e.g., Quality-Net [28] and IntelliNet [23]) have been proposed to infer a more general sense of intelligibility directly from the performance of an ASR system. Despite all of these efforts, Moore et al. [23] enumerate some reasons that current state-of-the-art is not suitable to predict commonalities between ASR and human recognition errors, such as: linguistic context, variation in phoneme pronunciations (due to distortion or accent), complicated interactions of AM and language model (LM), and different resolutions required to process speech acoustics (e.g., frame-level, word-level). Apart from these, the language proficiency of the listener in this prediction further complicates the detection of listening difficulties. Additionally, since not all ASR errors are indicators of speech difficulty for L2 learners, such errors should be further sifted to enhance their correlation with L2 learners' mistakes [1].

Here, we take another approach to the problem. Instead of detecting errors caused by acoustic factors, we adapt the ASR resources to match the target L2 learner/target accuracy. In this regard, we consider filtering out the training data for AM, LM, and lexicon as a training data selection problem. Conventionally, this problem is formulated as selecting the training data for which a favorable property is granted to the ASR, like minimizing the need for labeled training data for AM [29] and LM [30], or selecting data that is both informative and representative to increase the generalization [31]. However, to our knowledge, no study has been done to match the ASR errors with that of non-native human listeners.

In summary, we propose to limit the resources of ASR to match the errors caused by this *"impaired"* ASR with the mistakes of an L2 learner, within a target proficiency level, when processing the same listening material. To show that focusing on these predicted difficult segments is, in fact, useful for the L2 learners, we conduct an experiment where only the designated segments are shown in the captions to the L2 listeners (in the style of Partial and Synchronized Caption (PSC) [1]). Preliminary results revealed that such a caption can significantly enhance the recognition in L2 learners and can be used as an assistive tool to teach listening.

2. Method

In this section, we explore how impairing ASR resources affects its performance, propose a way to match ASR errors with those of L2 learners with particular language proficiency, and introduce an effective way to use such ASR errors for training L2 listening skill.

2.1. ASR Resource Ablation

Conventionally, an ASR system employs three main resources to recognize and transcribe the input stream: lexicon, acoustic model (AM), and language model (LM). Here, we methodically impair these resources to study the type of problems that they cause for the ASR system. To this end, we filter out the training data that may include the data above a certain level of complexity. Such knowledge is later used to create a resource profile that generates the desired set of ASR errors.

Impaired Lexicon: To match the vocabulary size of the L2 listener and ASR, the filtering strategy involves removing all parts of training data that have vocabulary out of the learner's reservoir. To this end, the vocabulary is determined by considering the level of the target listener. Limited lexicon leads to confusion in the ASR in different ways. For instance, changes in word boundary detection occur (Table 2) as the cases of out-of-vocabulary increases when ASR is processing normal speeches [17]. As posited in [32], L2 learners face more difficulty when listening to low-frequency words compared to high-frequency ones. The learners' lack of knowledge leads to misperception of these words, and they often tend to substitute the difficult words with those that are more familiar to them [33], which is also the case for impaired ASR with a limited lexicon. Moreover, learners have a general tendency to add word boundaries in order to perceive more frequent words than the actual word. They scan continuous speech to detect the similarities between sound sequences and known vocabulary, which leads to word boundary misperception.

The frequency of the words can be calculated considering the notion of word families [34] from the BNC and COCA corpora, and learners with higher L2 proficiency are expected to cover more families.

Table 2: *ASR recognition of "celebrated with ticker tape parades in every city" with different lexicon sizes.*

low	celebrated with	*thicker taper aids*	in every city
med	celebrated with	*thicker tape parades*	in every city
high	celebrated with	*ticker tape parades*	in every city

Impairing Language Model Language models are usually trained on large corpora, capturing different patterns of language, and fine-tune how the words can be connected. Training on larger corpora typically results in better language models where the distribution of the language model is closer to the natural patterns occurring in the target language, i.e., the actual distribution (e.g., Table 3). Therefore, to emulate lower proficiency levels of learners, we sort the sentences based on their PCFG (Probabilistic context-free grammar) surprisal [35] to favor syntactically easier sentences for the impaired LM model. The term surprisal refers to the cases that are less probable or expected to appear in the text, considering the precedent context, thus require more complex processing to comprehend [36]. We used SRILM (The SRI Language Modeling Toolkit) to create a language model from our training corpus. It should be noted that with a larger vocabulary, sufficiently training the language models becomes increasingly harder. Therefore, we cap the vocabulary size for training LM. Without the loss of generalization, more advanced language models (e.g., SciBERT [37]) can be used here. However, training the impaired versions of such models and eliciting ASR errors using these models require more time and computational power.

Table 3: *ASR recognition of "concerns about AIDS and avian flu and we'll hear about that" with different impaired language models (training data size).*

low	AIDS and	*even flew here*	about that
high	AIDS and	*avian flu and we'll hear*	about that

Impairing Acoustic Model The acoustic model of an ASR system extracts acoustic features of the input, and similar to human speech recognition, it requires training to have better performance. Early in L2 learning, the learners tend to recognize more frequent words (e.g., selecting the word in minimal pairs with higher frequency [38] as in Table 4). Therefore it is assumed that their hypothetical acoustic model is trained on more

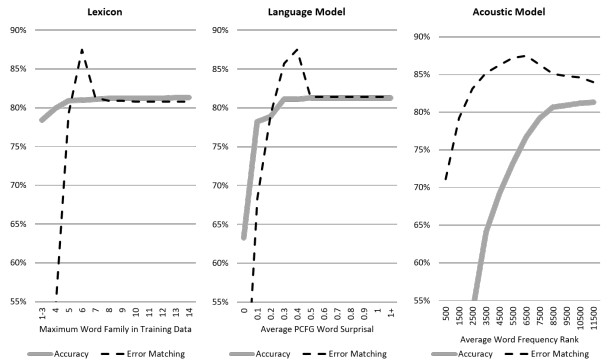

Figure 1: *The accuracy (on TEDLIUM 2 test set) and the error matching rate with a target proficiency (on the prepared dataset) of the impaired lexicon (left), impaired LM (mid) and impaired AM (right).*

frequent words. Aligned with this, we sort our training samples by the average frequency of the uttered words and train an impaired AM. We remove training samples that have less frequency on average. We employed a TDNN [39] with four hidden layers, each with 1536 units as the AM.

Table 4: *ASR recognition of "it's thick atmosphere, not it's distance" with different impaired acoustic models (training size).*

low	*it keeps that fear*	not it's distance
high	*it's thick atmosphere*	not it's distance

2.2. ASR to Emulate L2 Speech Recognition Mistakes

For ASR to emulate the mistakes of an L2 learner with particular language proficiency, we proposed to use simulated annealing to search for the optimal training size for ASR resources. We consider the TEDLIUM 2 [40] corpus that contains 1495 TED talks (452 hours) and their transcripts. For each training sample, we calculated the lexical difficulty using the word family list [34], word frequency using COCA corpus, and sentence complexity index using the PCFG parser [35]. The term word family refers to a base word and all its derived and inflected forms and includes 25 lists, each having 1000 word families [34]. To train the Impaired ASR, we need to consider the training data that satisfies three thresholds, on lexicon (based on average word family), LM (based on sentence complexity index), and AM (based on average word frequency).

In our method, an ASR engine is trained on the subset of data marked by these three thresholds, and its errors are identified using a forced alignment algorithm. These errors are compared with the difficult segments of the input audio, annotated for L2 listeners with a target proficiency. We use simulated annealing to update the thresholds to improve the correlations of these two sets (Figure 1). The initial threshold was set to be the middle point in the word family list range (for the lexicon), in PCFG word surprisal (for LM), and in word frequency rank (for AM). Here, we use Kaldi ASR [41] and Gentle force aligner, the dataset indices are normalized to 1. The simulated annealing optimizer fine-tunes the thresholds to maximize the F1-score of the ASR-detected segments compared to the annotations of difficult segments for the L2 learners of a target proficiency.

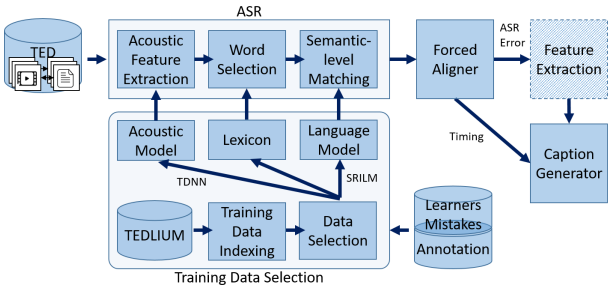

Figure 2: *Schematic of the system. ASR training data are selected to match its errors with the annotations/learner mistakes. Training on this data, the lexicon, language, and acoustic models for this impaired ASR are obtained. For an arbitrary input audio/video, the ASR transcription is given to forced aligner to detect the impaired ASR errors. These errors signal the listening material difficulties and, together with (optional) feature extraction, help to generate the final partial and synchronized caption.*

2.3. ASR Errors to Signal Listening Difficulty in Caption

Here, we used the Partial and Synchronized Caption (PSC) [1] and embed the ASR errors into this caption. In PSC, the audio is first fed to an ASR system. The ASR-generated and the original transcriptions are then passed to the forced aligner to obtain a word-level text-to-speech alignment. In PSC, a subset of words deemed difficult for the target proficiency are selected from the whole input transcript using features such as speech rate, word frequency, specificity, and perceptually challenging cases such as homophones and minimal pairs. The selected (difficult) words are then shown in sync with their utterance in the caption, while others are masked. Here, the difficult words are predicted by ASR errors (Figure 2).

3. Experiment

We first compare the performance of the proposed method to predict the difficult speech segments for L2 listeners using an annotated dataset. Then we experiment to examine the effectiveness of such caption in assisting L2 listeners when watching unseen TED videos.

3.1. Comparing ASR Errors with L2 Listeners' Mistakes

We collected and annotated a corpus of TED talk videos delivered by native speakers with a cumulative length of 93 minutes, including more than 10,000 words. We chose the videos of native American English speakers (standard accent) to eliminate the effect of accent as much as possible (although dialect might have had some effect). The reason is that accent is a factor that can largely affect listening comprehension for non-native speakers and thus should be analyzed individually. When the standard accent is used, listening difficulty could be a product of many other factors, most important of which is perceptual difficulty raised by the occurrence of the acoustical homophones, minimal pairs, breached boundaries, and speech disfluencies in the audio material [1].

The videos are selected from TED talks in 2017-18 and are not present in TEDLIUM 2 training dataset. The videos are forced aligned in word-level using Kaldi ASR, and the easy/difficult labels for each word are provided by two annotators targeting L2 intermediate proficiency. Our annotators were

Table 5: *Comparing the performance of proposed caption (based on errors made by Impared ASR) with feature-based PSC on the annotated dataset for intermediate learners (%)*

Caption	Easy words		Difficult words		Metrics			
	Show	Hide	Show	Hide	Prec.	Rec.	Acc.	F1
PSC-Rule	11.5	56.6	15.3	16.6	57.1	48.0	71.9	52.1
PSC-DT	9.2	60.9	10.3	19.6	52.8	34.4	71.2	41.7
PSC-NB	**2.4**	71.5	12.5	13.6	83.9	47.9	84.0	61.0
PSC-SVM	8.5	69.5	15.1	6.9	64.0	68.8	84.6	66.2
iASR	8.1	**73.1**	14.4	**4.4**	64.0	**76.6**	**87.5**	69.7
iASR+Feat	7.7	70.2	**16.1**	6.0	**67.6**	72.9	86.3	**70.2**

L2 instructors with clear annotation guidelines and identical instructions on the criteria to label the words. The resulting annotation has a κ=0.83 Cohen inter-annotation agreement.

To evaluate our model, we compare it with our baseline (PSC-Rule [1]). Additionally, we trained three different classifiers (SVM, Naïve Bayes, and Decision Tree) with a pool of features, including lexical (e.g., frequency, specificity, length, syllables), syntactic (e.g., part-of-speech, dependency parse relations), semantic (e.g., polysemic words, coreferences, idiomaticity), and acoustic or perceptual complexities (e.g., speech rate, breached boundaries, negatives) on the dataset with all these features and the labels annotators agreed upon. Additionally, we added the features of the baseline to the caption to select easy/difficult words (denoted by iASR+Feat). Table 5 compares the performance of these methods to detect the difficult words/phrases in the input audio for the target learner proficiency. Note that, in this classification task, *recall* indicates the portion of the difficult words that are shown. Therefore, it is considered more important than *precision*.

As shown in the table, iASR obtains the best accuracy compared to other methods and gained the best recall, as it focuses on the difficult cases. However, overall, iASR+Feat demonstrates a better F1-score by labeling around 3% more words as difficult. This shows that there are still some listening difficulties (e.g., those caused by moderately high speech rate) that the Impaired ASR model cannot capture.

3.2. Predicted Difficult Segments to Improve Listening

In this experiment, the proposed caption based on Impaired ASR is compared with the no-caption as well as baseline PSC (PSC-RULE) in experiments with L2 learners of English. We are using all factors here because previous experiments with our baseline using individual factors showed that including only one feature in the caption will not lead to statistically significant improvement in learners' comprehension. A possible explanation is that comprehension is a complex process involving decoding, and word recognition, syntax understanding, and meaning construction, and these factors are correlated [42].

In the first experiment (iASR vs. no-caption), 26 learners of English with TOEIC scores above 750 listened to TED talks of about three minutes long, either with iASR caption or without captions. After each video, the participants were asked to answer the comprehension questions and the listening cloze tests that followed (videos were rotated among participants). Figure 3 (left) shows the learners' scores on no-caption versus iASR caption condition. As shown in the figure, participants' scores when receiving iASR caption are significantly higher than those with no captions ($p<.001$).

In the next phase, these intermediate-level learners were divided into two groups and watched a series of short video segments either with baseline PSC or with iASR caption, followed by several paraphrasing questions. The numbers of shown words to both groups were controlled to be the same, while the choices of words in the baseline and enhanced versions were different. The results in Figure 3(right) show that participants who used the iASR caption obtained higher paraphrasing scores compared to the other group. This shows that the proposed caption enhances the recognition of the video and better assists with selecting words to appear in the caption.

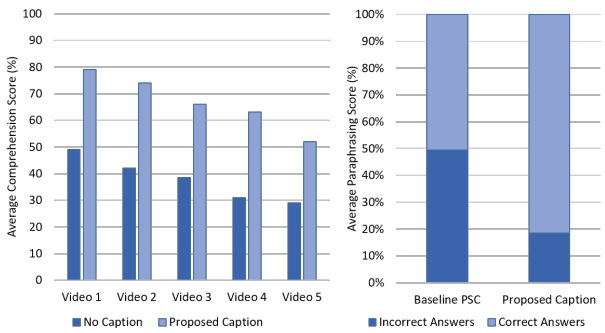

Figure 3: *(Left) Results of proposed caption versus no caption for five unseen videos, (Right) Results of proposed caption versus baseline PSC using paraphrasing test*

4. Conclusions

In this study, we proposed using ASR errors as the indicators of speech difficulties that can be adapted for different L2 learners with various proficiency levels. It does so by using a calibrated selection of ASR training samples. We considered the factors that affect L2 listening for different proficiency levels and tried to match the ASR level with the L2 learner's proficiency, considering those factors. In this regard, L2 learners' vocabulary reservoir, acquired grammar, and listening experience is modeled by ASR's lexicon, language model, and acoustic model. Using a guided selection of training samples, any arbitrary language proficiency can be modeled by the ASR. Our preliminary experiments with L2 learners of English revealed that such a model could predict the listening difficulties accurately and, when used with captioning, can act as an assistive tool for training L2 listening.

5. References

[1] M. S. Mirzaei, K. Meshgi, and T. Kawahara, "Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening," *Computer Speech & Language*, 2018.

[2] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *Automatic Speech Recognition and Understanding*, 2001.

[3] A. Bloomfield and S. C. e. a. Wayland, "What makes listening difficult? factors affecting second language listening comprehension," College Park, MD, Tech. Rep., 2010.

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

[5] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[6] D. Gentner, "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. technical report no. 257." 1982.

[7] H. Nitta, H. Okazaki, and W. Klinger, "An analysis of articulation rates in movies," *ATEM J.*, 2010.

[8] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, 1999.

[9] R. Griffiths, "Speech rate and listening comprehension: Further evidence of the relationship," *TESOL quarterly*, 1992.

[10] M. Benzeghiba, R. De Mori, O. Deroo, and S. e. a. Dupont, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[11] N. Osada, "Listening comprehension research: A brief review of the past thirty years," *Dialogue*, 2004.

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

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

[14] M. Forsberg, "Why is speech recognition difficult," *Chalmers University of Technology*, 2003.

[15] M. Broersma, "Increased lexical activation and reduced competition in second-language listening," *Language and cognitive processes*, vol. 27, no. 7-8, pp. 1205–1224, 2012.

[16] J. Field, "Promoting perception: Lexical segmentation in l2 listening," *ELT journal*, vol. 57, no. 4, pp. 325–334, 2003.

[17] W. Chen, S. Ananthakrishnan, R. Kumar, R. Prasad, and P. Natarajan, "Asr error detection in a conversational spoken language translation system," in *ICASSP'13*. IEEE, 2013, pp. 7418–7422.

[18] T. Pellegrini and I. Trancoso, "Error detection in broadcast news asr using markov chains," in *Language and Technology Conference*. Springer, 2009, pp. 59–69.

[19] T. Pellegrini and I. Trancoso, "Improving asr error detection with non-decoder based features," in *INTERSPEECH'10*, 2010.

[20] Y.-C. Tam, Y. Lei, J. Zheng, and W. Wang, "Asr error detection using recurrent neural network language model and complementary asr," in *ICASSP'14*, 2014.

[21] Y. Deng, M. Mahajan, and A. Acero, "Estimating speech recognition error rate without acoustic test data," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[22] S. Ghannay, Y. Estève, and N. Camelin, "A study of continuous space word and sentence representations applied to asr error detection," *Speech Comm.*, 2020.

[23] M. Moore, M. Saxon, H. Venkateswara, V. Berisha, and S. Panchanathan, "Say what? a dataset for exploring the error patterns that two asr engines make," in *INTERSPEECH*, 2019.

[24] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv*, 2016.

[25] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing asrs: Evaluating the need for more inclusive automatic speech recognition systems," in *INTERSPEECH*, 2018.

[26] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[27] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in *ACL*, 2018, pp. 20–24.

[28] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv*, 2018.

[29] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *CSL*, 2010.

[30] T. Drugman, J. Pylkkonen, and R. Kneser, "Active and semi-supervised learning in asr: Benefits on the acoustic and language models," *arXiv*, 2019.

[31] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *NeurIPS*, 2010, pp. 892–900.

[32] N. Schmitt and D. Schmitt, "A reassessment of frequency and vocabulary size in l2 vocabulary teaching," *Language Teaching*, vol. 47, no. 04, pp. 484–503, 2014.

[33] A. Cutler, "The lexical statistics of word recognition problems caused by l2 phonetic confusion." in *Interspeech*, 2005, pp. 413–416.

[34] I. S. Nation, "How large a vocabulary is needed for reading and listening?" *Canadian Modern Language Review*, vol. 63, no. 1, pp. 59–82, 2006.

[35] M. Van Schijndel, A. Exley, and W. Schuler, "A model of language processing as hierarchic sequential prediction," *Topics in Cognitive Science*, 2013.

[36] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[37] I. Beltagy, K. Lo, and A. Cohan, "Scibert: Pretrained language model for scientific text," in *EMNLP*, 2019.

[38] J. Field, "Bricks or mortar: which parts of the input does a second language listener rely on?" *TESOL quarterly*, 2008.

[39] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH'15*.

[40] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *Conference on Language Resources and Evaluation (LREC)*, 2014.

[41] D. Povey and A. e. a. Ghoshal, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[42] 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.