ASR Technology to Empower Partial and Synchronized Caption for L2 Listening Development

Maryam Sadat MIRZAEI[1], *Tatsuya KAWAHARA*[1]

[1,2]Graduate School of Informatics, Kyoto University
Sakyo, Kyoto, 606-8501 Japan
maryam@ar.media.kyoto-u.ac.jp, kawahara@ar.media.kyoto-u.ac.jp

Abstract

This study introduces a tool, partial and synchronized caption (PSC), for training second language (L2) listening skill. PSC uses an automatic speech recognition (ASR) system to realize word-level alignment between text and speech while it refers to the corpora to effectively select a subset of words for inclusion in the caption. The selection criteria are based on three features contributing to L2 listening difficulties: speech rate, word frequency and specificity. Our findings reveal that PSC in its current state leads to the same level of comprehension as the full caption condition. PSC, however, outperforms the full caption when it comes to preparing learners for listening without using any textual clues as in real-life situations. To enhance this system the incorporation of other features is a necessity. However, the relationship between these factors and their contribution to listening difficulty is complex. This study conducts a root cause analysis on the ASR errors to better understand the underlying features that make recognition difficult for such systems and compares these features with L2 listening influential factors. Our preliminary analysis revealed an interesting similarity between features leading to L2 difficulty and those resulting in ASR errors. Such insightful findings shed light on the future improvements for PSC.

Index Terms: listening skill, partial and synchronized caption, automatic speech recognition, speech rate, word frequency

1. Introduction

Captioning has been long used as a source of scaffold for L2 listeners when watching authentic videos [1, 2, 3]. This assistive tool provides textual clues to facilitate listening comprehension. However it also promotes significant amount of reading which raises the question whether listeners' comprehension is gained by merely reading the text instead of listening to the audio [4]. In order to encourage L2 listeners to listen more, meanwhile assisting them to overcome the difficulties of listening to authentic material, this study developed a system to generate a smart type of caption (PSC) that strives to improve L2 listening skill using two approaches: synchronization and partialization. Synchronization is to map each word to its corresponding speech signal using an ASR system, which makes a word-level alignment between the text and speech, thereby emulates the speech flow and allows for precise speech-to-text mapping. This feature better visualizes the word boundaries, however it develops word-by-word decoding strategy, which deteriorates effective listening [5]. Thus the latter approach, partalization, is introduced in order to promote listening to the audio by restricting the number of words presented in the caption and decreasing textual density. As a result, listeners can only refer to the text for difficult words/phrases and are obliged to listen in order to comprehend. Figure 1 demonstrates a screenshot of a video captioned with PSC.

The effectiveness of this method largely depends on the choice of words to appear in the caption. Previous studies have investigated the effect of partial captions in the form of "keyword captioning" where keywords are manually selected by some experts and presented in the caption while the rest of words are not visible [6, 7]. In PSC, however, keywords are not the selection criteria. Inspired by an overview of significant research on L2 listening, this study is based on the factors contributing to listening difficulties as prudent criteria for selecting words to appear in PSC. A number of factors in speech and language varying from acoustic level to lexical, syntactic and pragmatic level affect comprehension [8]. While each feature plays a role in listening difficulty, some are largely referred to as the dominant obstacles of L2 listening. Of these, fast rate of speech, infrequent words and specific terms are taken for granted as the primary factors for PSC word selection [9, 8]. These features are quantified and incorporated to the system in order to sift the words. However, central to this process are the learners' demands that vary according to their proficiencies. In this view, PSC should be adjusted to the learners' level and its parameters need to be tuned. Demonstrations of this type of caption are available at http://www.ar.media.kyoto-u.ac.jp/psc/.

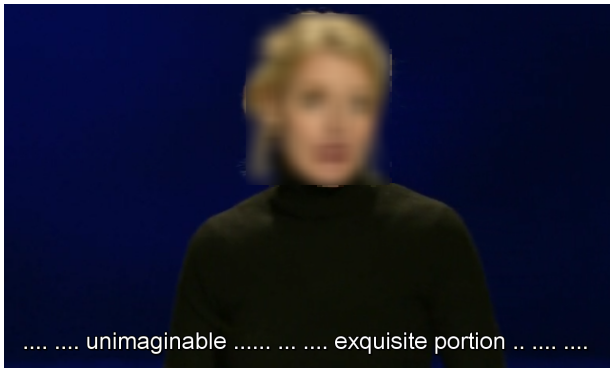

Figure 1: *Screenshot of PSC generated for a TED talk. The original transcript is: "[. . .] from some unimaginable source, from some exquisite portion of your life."*

To further improve the performance of the system, it is necessary to consider a wide range of features to be aggregated and act on PSC generation. However, the relationship between these factors and their significance in listening difficulties is complex. This study addresses this issue by investigating the

sources of ASR errors compared with the factors that lead to listening difficulties for language learners. This effort has been inspired, in part, by the comparable nature of the difficulties in transcription of spoken data by both the ASR system and L2 listeners. Performance of ASR systems in perceiving the speech has been compared against human beings on a transcription task: human speech recognition (HSR) [10, 11, 12]. However, in most studies HSR subjects are limited to either the native speakers or people who has no knowledge of the target language (e.g. Japanese with no knowledge of Italian listening to an Italian audio which includes words with maximum phonetic similarity between the two languages). Some studies have emphasized the importance of conducting equitable HSR-ASR comparisons by restricting the influence of background information, using logatomes/ pseudowords [10]. However, as both ASR and L2 listeners have limited knowledge/resources when transcribing authentic materials, comparison between ASR and L2 speech recognition (L2SR) seems to be a more appropriate choice. ASR errors indicate the problematic speech regions with respect to the system's configuration. L2 listeners' difficulties identify the problematic factors that attenuate effective comprehension for language learners. A comparison of the two highlights the joint errors, reveals the differences and specifies whether ASR errors can be epitomized as the sources of L2 listening difficulties. The findings of this study can be incorporated into PSC to improve the choice of words.

The manuscript is organized in two parts: first, a system is developed to generate PSC; next, ASR errors are analyzed and compared against L2 listeners' problems with the aim of enhancing PSC.

2. PSC System

The videos of this study were selected from the TED website, based on properties such as popularity and recentness. The selection was inclusive of American speakers in order to restrict the effect of accent. The pipeline of generating PSC is based on two main modules: the synchronization and the partialization. Figure 2 depicts the architecture of the system.

2.1. Synchronization

Synchronization in this method is done in word level, which pertains to aligning each word to its respective speech signal. Synchronized captioning presents the phonological visualization of the words and thus leads to improvement in word recognition skills as it allows for mapping between the speech stream and verbatim text. This process is realized by the word-level alignment of an ASR system, *Julius 4.3.1* [13]. To pull out this task, the audio is ripped from the video and sent to the ASR system. In order to increase the accuracy of captions and alignments, the system is trained with the TED dataset, using 780 hours of TED talks. This model training was done based on a lightly supervised training approach [14].

The ASR system outputs the generated caption, the timestamp for each word, and the confidence measure of each word. ASR transcripts are then compared with the original transcripts of TED talks which are available on the TED website. Finally, the two transcripts are aligned using a dynamic programming procedure in order to eliminate all the ASR errors. Synchronized captions, although in favor of many language learners, may bring too much assistance, making learners more and more dependent on the caption, encouraging them to follow each word (word-by-word decoding) and merely use their reading

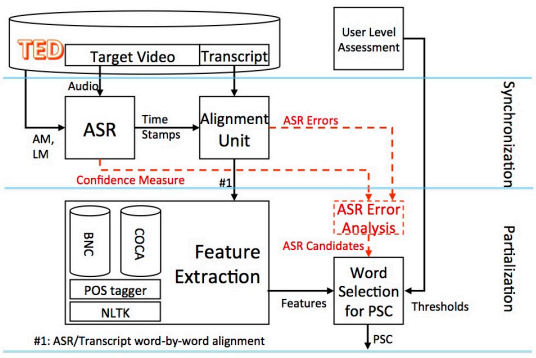

Figure 2: *System architecture of PSC generation.*

skill instead of listening [5]. In order to solve the disadvantages of this method, we propose partial captioning, which builds on synchronized caption. Therefore the outputs of this stage are used for the next tier: the partialization module.

2.2. Partialization

The aim of this process is to tailor the captions to the learner's needs based on a set of features that influence effective listening. Partialization here refers to keeping specific words/phrases in the verbatim captions intact, while replacing others with a masking character (here period is used). The goal is to promote listening over reading and hence assist learners to develop listening skills while effectively facilitating this process by allocating problematic words/phrases as a recourse to the learners.

The partialization module adopts feature extraction in order to determine the choice of words/phrases. The feature extraction module relies on different aspects that have impact on successful listening comprehension. *Speech rate*, *word frequency* and *specificity* are selected as the features of this study for being recognized as the major sources of frustration for L2 listeners, especially Japanese learners [15, 16]. Another beneficial reason is the feasibility of quantifying these features; the provision of timestamps throughout the synchronization phase enables precise speech rate calculation and the viability of referring to corpora enables word frequency determination and specificity detection.

In the partialization stage, words are considered individually and the features are calculated for each of the words. The following elaborates on feature calculation.

2.2.1. Speech Rate

Previous studies showed that high speech rate can negatively affect L2 listeners' comprehension and this is even true for native speakers. Nitta *et al.* [16] reported that even when familiar with all the vocabulary in the experiment, higher speech rates result in more missed or mistaken words. Therefore, calculating the speaker's speech rate is beneficial for effective word selection. To this end, we detect words with high speech rate and present them in PSC in order to facilitate comprehension.

Speech rate is calculated using different measures such as: words per minute, syllables per second, and phonemes per second. Of these, syllables per second is the most appropriate choice. Words per minute can be affected by pauses and variations in speech that are often caused by speaker's emotional fluctuations and excitement, hence is not recommended. Moreover, the relation between phonemes and speech rate is neither linear nor simple. In contrast, syllables per second (SPS), has

fairly uniform distribution over speech rate and is more robust against variations in speech [17]. Therefore, SPS is used as a measurement unit for speech rate calculation in this study. Computing speech rate is then realized by using the timestamps of the words derived in the first stage serving as duration of the word segments in seconds and syllabifying the words to count the number of syllables. The defacto standard to perform automatic syllabification is based on Knuth-Liang hyphenation algorithm for English language [18]. In this method Natural Language Toolkit was used to do this task. Speech rate is then simply calculated in SPS for each word.

2.2.2. *Word Frequency*

The frequency of word occurrence in large spoken/written corpora is often referred to as the frequency of a word in a language. Previous research has shown that the occurrence of infrequent words in speech confines the learner's attention, and prevents him/her from pursuing the subsequent parts of the audio. When encountering such words, the listener invests a lot of time to understand what he/she missed [19]. As a result, presenting infrequent words in PSC is beneficial for facilitating listening process.

To calculate the frequency of each word, we referred to two famous corpora and a series of word family lists:

- British National Corpus (BNC): includes 100 million word collection of written and spoken language from a wide range of sources, to represent a wide cross-section of British English.
- Corpus of Contemporary American English (COCA): contains more than 450 million words of text and is equally divided among spoken, newspapers, academic texts, etc. The corpus is also updated regularly [20].
- Word family list: The term word family here refers to a base word and all its derived and inflected forms that can be recognized by a learner without having to learn each form separately [21]. These consist of 29 word family lists, which were made based on BNC and COCA. Twenty-five of the lists contain word families based on frequency and range data. The four additional lists are: an ever-growing list of proper names, a list of marginal words including swear words, exclamations and alphabets, a list of transparent compounds, and a list of abbreviations.

Using these, the system automatically checks the frequency of each word in order to make a decision on its inclusion in PSC.

2.2.3. *Specificity*

According to Goh [19], limited vocabulary especially for academic words is often seen as a cause of L2 listening comprehension impair. Using TED talks in this study emphasizes the importance of considering academic jargon. This feature is handled by using the following resources:

- Academic Word List: a popular catalogue including 570 headwords and about 3000 academic words [22].
- COCA academic word list: uses a larger and newer corpus to provide a better coverage for academic words with higher specificity and word family information.

The final decision on selecting the words will include less than 30% of the total transcript and is based not only on the acting features, but also on the proficiency level of each learner.

Table 1: Standards rates of speech for L2 learners.

Speech Rate	Word per Minute	Syllables per Minute
Fast	Above 220	Above 320
Moderately Fast	190∼220	280∼320
Average	160∼190	230∼280
Moderately Slow	130∼160	290∼230
Slow	Below 130	Below 190

To this end, learners are asked to complete a series of tests, which reflect the appropriate level of PSC to be generated for them. The tests include listening to several questions, which were played with 4 different levels of speech rate: slow, moderate, fast and very fast. Listeners' results on this test allowed us to determine their tolerable rate of speech and hence thresholding the system for this feature. Besides, the standard rates of speech for L2 learners [23, 24] were also taken into account when defining the thresholds (Table 1).

In order to adjust the word frequency level, the learners were asked to take a vocabulary size test. The test used here is designed by Nation [25] and is based on the aforementioned word family lists and hence provides a fairly accurate borderline for determining the unknown/unfamiliar words for the learners. Finally, specific words were decided to be included in PSC intact for all of the learners. Along with specificity, other instances of the words such as proper nouns, abbreviation and difficult compound words were also always presented in PSC to assist the listeners. However, interjection and easy compound categories are constantly hidden.

The adjustable nature of PSC makes it an effective tool to serve a wide range of learners with different proficiency levels. Table 2 compares the PSC method with other captioning methods and highlights its advantages.

Table 2: Advantages of PSC.

Caption type / Advantages	Full	Keyword	Proposed Partial	Synchronized	PSC
Aid word boundary detection	✓			✓	✓
Speech-to-text mapping				✓	✓
Avoid over-reliance on reading		✓	✓		✓
Avoid being distractive	✓			✓	✓
Automatic	✓		✓	✓	✓
Adjustable to learners' knowledge			✓		✓
Adjustable to the content		✓	✓		✓

2.3. PSC Evaluation

PSC has been compared against other methods by conducting an experiment in two CALL classes with 58 Japanese learners of English. The participants were undergraduate students from 19 to 22 years old who enrolled in a CALL course.

These students were divided into three proficiency groups (beginners, pre-intermediate and intermediate) based on their scores of a CASEC™or TOEIC™ test. The subjects were asked to take the speech rate tolerance test and the vocabulary size test, as was noted in the previous section. The results allowed us to generate appropriate PSC for each of the three proficiency groups. Throughout the experiment the participant watched a series of TED talks under one of these conditions: no caption

(NC), full caption (FC) and PSC. They then answered several multiple choice comprehension questions based on the content of the videos to assess their comprehension. In order to evaluate the effectiveness of each condition on preparing learners for listening in real-life situations (i.e., when assistive tools such as captions are not available), the experiment was designed as the following: When watching videos under NC, FC or PSC condition, only 70% of the video (from the outset) was played to the learners. This was followed by the comprehension questions and formed the 1st part of the experiment. The remaining 30% of the videos, however, were preserved for the 2nd part of the experiment and used without any caption, as in real-life situations. After watching the remaining part, again the learners were asked to answer several comprehension questions. This experiment distinguishes the impact of NC, FC or PSC on preparation for listening without any captions.

Table 3 reports the scores of the participants with different proficiency levels on the comprehension tests for the 1st part of the experiment, i.e., 70% of video with NC, FC or PSC.

Table 3: Mean scores and standard deviations on listening comprehension - Part 1

Caption	Proficiency Level	N	Mean	SD
NC	Beginner	19	28.67	13.56
	Pre-intermediate	19	34.71	11.85
	Intermediate	20	43.27	15.11
	Total	58	**35.69**	14.68
PSC	Beginner	19	42.04	16.70
	Pre-intermediate	19	52.00	17.50
	Intermediate	20	64.05	17.99
	Total	58	**52.89**	19.39
FC	Beginner	19	41.10	12.35
	Pre-intermediate	19	57.20	14.85
	Intermediate	20	63.93	16.38
	Total	58	**54.25**	17.33

The results of repeated-measure ANOVA test on the overall performance of participants on the first part of the experiment revealed statistically significant differences between NC condition ($M = 35.7, SD = 14.7$) and PSC condition ($M = 52.9, SD = 19.4$) or FC condition ($M = 54.2, SD = 17.3$) at $p < .05$. However, a pairwise comparison between the scores of the PSC and FC conditions in this part revealed no statistically significant difference $[F(1, 57) = 25, p = .62]$ between these two conditions. The results suggest that PSC, while presenting less than 30% of the transcript, leads to the same level of comprehension as FC, which includes 100% of the text.

As presented in Table 4, in the second part of the experiment (30% without caption), the best performance is associated with the condition in which the learners first watched the video with PSC $[F(2, 118) = 20.5, p < .05]$ and then without caption, as compared to watching video with FC and NC first. The results indicate the effectiveness of PSC on preparing the learner for real-life situation as compared to NC and FC. Although this is a short-term enhancement partly because of adaptation to speaker, this finding is still valuable.

3. ASR Errors vs. L2 Listening Problems

Generally, the errors of ASR systems are evaluated in terms of their alignment-timing accuracy and their correctness. Here we are not dealing with the timing errors, but the recognition errors in lexical level. Such kind of errors have been consistently viewed as a negative product of the ASR system, which explains why ASR generated transcripts are not beneficial to be used for L2 learners. Such errors are known to be misleading and confusing for L2 listeners since they interrupt the comprehension process, interferes with dual coding of the input data and impedes text-to-speech mapping. As a result, ASR transcripts to be used for L2 learners have low tolerance to the errors. Even below 5% word error rate (WER) is too high for the end-users [12]. While this assumption is true, ASR errors can be viewed as an information source for problematic speech regions not only for the system itself, but also for L2 listeners who, similar to the ASR, have limited background knowledge and resources of the target language.

Table 4: Mean scores and standard deviations on listening comprehension - Part 2

Caption	Proficiency Level	N	Mean	SD
NC	Beginner	19	32.95	16.03
	Pre-intermediate	19	37.37	16.57
	Intermediate	20	50.05	15.56
	Total	58	**40.12**	17.39
PSC	Beginner	19	49.60	15.74
	Pre-intermediate	19	57.67	17.15
	Intermediate	20	62.51	17.37
	Total	58	**56.59**	17.34
FC	Beginner	19	38.31	13.48
	Pre-intermediate	19	40.39	11.86
	Intermediate	20	49.26	12.71
	Total	58	**42.65**	13.37

Establishing a meaningful relation between different extracted features and the type of ASR errors requires a careful investigation, which is the topic of several studies such as [26, 27]. In this study we try to confirm this background knowledge using our system, and discover new relations in order to compare the findings on ASR error analysis with L2SR problems.

The findings from the former will be used to enhance the quality of PSC. We analyze the correctness of generated transcript by aligning the ASR output with the human transcript word-by-word in order to detect different types of errors. The errors are then grouped into three main categories: insertion, substitution and deletion. In the next phase, the errors were further analyzed in order to identify the underlying features that led to their occurrence. The selection of these features is inspired by the factors that makes L2 listening difficult for the learners such as: duration, speech rate, word frequency, word length and part of speech.

3.1. ASR Errors Analysis

Eleven TED talks were selected for this experiment and the ASR errors on the transcription task were detected and analyzed. In total there were 29391 word tokens, of which 7017 words were mistakenly recognized. Word error rate (WER) averaged 23.87%. These errors are categorized into:

Substitution Errors: ASR transcript and ground truth are sometimes different in one or more words. These mismatches can be categorized into several subcategories:

1. Basic mismatch: This can be the beginning of a mismatch sequence. (36.24% of all errors)
2. Long mismatch: This is part of a chain of mis-recognition by the ASR. (44.02%)
3. End of Sentence: Happens due to the mis-recognition of the end of the sentence by ASR. (3.97%)

4. Hyphenation: The transcribed word is hyphenated while the ground truth has it as two separated words, and vice versa. (1.48%)
5. Numbers: The numbers are sometimes spelled out in the ASR transcript, while they are always in the numerical format in the ground truth. (2.37%)
6. Abbreviations: The abbreviations are not always properly dotted in the training corpus or in the ground truth. (0.08%)

Deletion Errors: Instances where ASR did not transcribe something, which is present in the ground truth. (5.18%)

Insertion Errors: In this case, ASR transcribed something, which is not present in the ground truth. (6.62%)

3.2. ASR-L2SR Comparison

Table 5 demonstrates a comparative study we performed through extensive investigation of background studies to compare the factors that affect the performance of ASR and those that influence L2 listening comprehension.

Table 5: ASR vs. L2SR

ASR Difficulties	L2 Listening Difficulties
Co-articulation, pronunciation, speaking style, disfluencies, accent, age, physiology and emotions lead to ASR difficulties [28]	Pronunciation can be unclear due to assimilation, reduction, etc. Stress, intonation patterns and accent affect L1/L2 listening comprehension [8]
Infrequent words are more likely to be misrecognized [26]	The occurrence of infrequent words in speech is correlated to complexity [9]
Fast speech / very slow speech increases error rates [29]	Whether too fast or too slow, speech rate can act as a barrier for listening [30]
Word length has also been found to be a useful predictor of higher error rates [26, 27]	The length of a word has strong effect on word recognition and word learning. Studies have reported mixed results on this effect [31]
Open class (N. and V.) has lower error rate compared to closed class (Prep., articles) [32]	Recognition of content words is easier than function words [16]

Based on this table we extracted various features in order to detect possible trends of ASR errors. In this study, word duration, word length, number of syllables and speech rate were calculated for each word according to the same procedure used in PSC generation. Word frequency was calculated by referring to COCA and word family lists. Moreover, Stanford POS tagger [27] was used to identify the part of speech.

As Figure 3 suggests, too fast speech rates deteriorate the performance of ASR systems in a significant way. In line with this result, studies on L2 listening skill have emphasized the role of the fast speech rates in L2 listening comprehension impair. Nitta et al [16] reported that at 4 sps, L2 learners missed or mistook 4.2% of the words, of which 2.7% were function words and 1.5% were content words. At 5 sps, this number jumped to 12.6%: 10.5% function words and 2.1% content words. At 8 sps, the errors were 40.6%; 30.1% for function words and 10.5% for content words. They also indicated that at 7 sps and 8 sps, the native speaker subjects also began to miss the words.

Secondly, in order to determine the effect of word frequency, we grouped the words into three frequency groups. These groups were adopted from Nation [21] categorization of English vocabulary: High-frequency (the most frequent 2000∼3000 word families), Mid-frequency (3000∼9000 word families), and Low-frequency (beyond 9000 frequency band).

Figure 3 demonstrates that WER significantly increased for those words that belonged to the low-frequency categories. The ASR system could successfully select the words that were in mid-frequency band and its performance was even better than the case the words were in the high frequency group. The fact that many of high-frequency words are functional words can explain the reason to this phenomena. In accordance with this result, studies on L2 listening comprehension agree that infrequent words are often more difficult for L2 learners and the emergence of these words highly affect successful comprehension [9].

Thirdly, the figure presents that in most cases, the longer the words are the better the ASR can recognize them. Longer words have longer duration, which makes it easier for ASR system to identify them [29, 30]. For language learners, however, studies have reported controversial results. While some studies showed that longer words are easier to be recognized when listening to audio materials, others have reported that shorter words are easier to learn and hence easy to recognize [31].

Finally, in line with background studies the ASR system tends to have lower WER in noun recognition. Similarly, for L2 learners, such words are easier to learn and also easier to recognize. Nouns predominate over predicates/verbs in most of languages [16, 31]. In case of verbs, however, we found out that most of the ASR errors belong to the past participle form of the verbs. There were the cases when ASR has actually correctly transcribed the audio, but the human annotator preferred to use the formal written form of the verbs (e.g. “gonna” by ASR vs. “going to” by human), which led to a mismatch between the

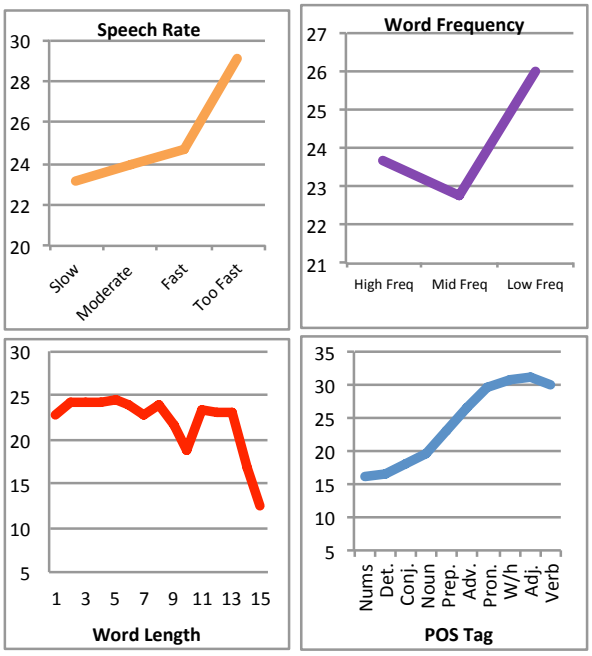

Figure 3: ASR error analysis based on features. The Y-axis shows the percentage of ASR WER.

two. The results excluding this case indicate less WER on the verb categories as well as nouns.

By analyzing ASR errors and comparing them with L2 listening difficulties, we found interesting similarities. However, apart from the factors that contribute to both ASR errors and L2SR difficulties, there are some instances that ASR errors cannot be attributed to any of these categories. Some of these errors indicate the words which are generally difficult to recognize even for language learners. These instances confirm the usefulness of ASR errors as a source of predicting L2 listening difficulties and suggest the potential importance of adding these words to PSC. On the contrary, ASR can sometimes recognize fast and/or infrequent words correctly. If these words are actually easy to recognize, they can be simply excluded from PSC with the hints of ASR. Nevertheless, these assumptions should be confirmed by conducting some experiments with L2 learners, using similar samples. This step is an ongoing process of this study.

4. Conclusions and Future work

Using TED Talks as appropriate authentic medium, we made a captioning method that strives to provide adequate support, decrease dependence on captions and prepare learners for real-world situations. The results confirmed the effectiveness of this method in preparing learners for listening without reading.

We also conducted ASR-L2SR comparison to diagnose the difficulties of L2 listening since ASR can serve as a simplified model of a language learner. The complex architecture of ASR is an invaluable resource to indicate possible barriers in the listening process. Modeling L2 learner with ASR introduces new trends to adapt the system to learners' need. In this regard, as a future work we can degrade the ASR models to the learners' levels. The first and foremost idea is to train ASR acoustic model on the learners' L1 corpora, to emphasize the role of phonetic differences between L1 and L2 in listening impediment. It is also possible to degrade language model by reducing the training data or omitting low-frequency words from dictionary. The ASR error analysis unit is then provided with the transcript of these three attenuated ASRs to find new candidates for inclusion in PSC.

5. References

[1] T. J. Garza, "Evaluating the use of captioned video materials in advanced foreign language learning," *Foreign Language Annals*, vol. 24, no. 3, pp. 239–258, 1991.

[2] M. Danan, "Captioning and subtitling: Undervalued language learning strategies," *Meta*, vol. 49, pp. 67–77, 2004.

[3] P. Winke, S. Gass, and T. Sydorenko, "The effects of captioning videos used for foreign language listening activities," *Language Learning & Technology*, vol. 14, no. 1, pp. 65–86, 2010.

[4] J.-T. Pujolà, "Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program," *ReCALL*, vol. 14, no. 02, pp. 235–262, 2002.

[5] L. Vandergrift, "1. listening to learn or learning to listen?" *Annual Review of Applied Linguistics*, vol. 24, pp. 3–25, 2004.

[6] H. G. Guillory, "The effects of keyword captions to authentic french video on learner comprehension," *Calico Journal*, 1998.

[7] M. Montero Perez, E. Peters, and P. Desmet, "Is less more? effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension," *ReCALL*, vol. 26, 2014.

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

[9] N. Osada, "Listening comprehension research: A brief review of the past thirty years," *Dialogue*, vol. 3, no. 1, pp. 53–66, 2004.

[10] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *SRIV'06 Workshop*, 2006.

[11] 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.

[12] I. Vasilescu, D. Yahia, N. D. Snoeren, M. Adda-Decker, and L. Lamel, "Cross-lingual study of ASR errors: On the role of the context in human perception of near-homophones." in *INTERSPEECH*, 2011, pp. 1949–1952.

[13] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *APSIPA ASC'09*, 2009.

[14] W. Naptali and T. Kawahara, "Automatic transcription of TED talks," in *IWSLT'12*, 2012.

[15] N. Osuka, "What factors affect Japanese EFL learners/listening comprehension," *JALT'07*, pp. 337–344, 2008.

[16] H. Nitta, H. Okazaki, and W. Klinger, "An analysis of articulation rates in movies," *ATEM Journal*, vol. 15, pp. 41–56, 2010.

[17] D. Wang and S. Narayanan, "An unsupervised quantitative measure for word prominence in spontaneous speech." in *ICASSP'05*, 2005.

[18] F. Liang, "Word hy-phen-a-tion by com-pu-ter." Ph.D. dissertation, Stanford University, 1983.

[19] C. C. Goh, "A cognitive perspective on language learners' listening comprehension problems," *System*, vol. 28, pp. 55–75, 2000.

[20] M. Davies, "The Corpus of Contemporary American English: 450 million words,1990-present(http://corpus.byu.edu/coca/)," 2008-.

[21] I. S. Nation and S. A. Webb, *Researching and analyzing vocabulary*. Heinle, Cengage Learning, 2011.

[22] A. Coxhead, "A new academic word list," *TESOL quarterly*, vol. 34, no. 2, pp. 213–238, 2000.

[23] P. Pimsleur, C. Hancock, and P. Furey, "Speech rate and listening comprehension," *Viewpoints on English as a second language*, 1977.

[24] S. Tauroza and D. Allison, "Speech rates in British English," *Applied linguistics*, vol. 11, no. 1, pp. 90–105, 1990.

[25] I. Nation and D. Beglar, "A vocabulary size test," *The language teacher*, vol. 31, no. 7, pp. 9–13, 2007.

[26] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *ASRU*, 2001.

[27] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *HLT-NAACL*, 2003, pp. 173–180.

[28] M. Benzeguiba, 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 intrinsic speech variation," in *ICASSP'06*, vol. 5. IEEE, 2006.

[29] 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.

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

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

[32] 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.