Computer Speech and Language
journal homepage: www.elsevier.com/locate/csl

Exploiting Automatic Speech Recognition Errors to Enhance Partial and Synchronized Caption for Facilitating Second Language Listening

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

[a]*Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo Ward, Kyoto, 606–8501 JAPAN*

Abstract

This paper addresses the viability of using Automatic Speech Recognition (ASR) errors as the predictor of difficulties in speech segments, thereby exploiting them to improve Partial and Synchronized Caption (PSC), which we have proposed to train second language (L2) listening skill by encouraging listening over reading. The system uses ASR technology to make word-level text-to-speech synchronization and generates a partial caption. The baseline system determines difficult words based on three features: speech rate, word frequency and specificity. While it encompasses most of the difficult words, it does not cover a wide range of features that hinder L2 listening. Therefore, we propose the use of ASR systems as a model of L2 listeners and hypothesize that ASR errors can predict challenging speech segments for these learners. Among different cases of ASR errors, annotation results suggest the usefulness of four categories of homophones, minimal pairs, negatives, and breached boundaries for L2 listeners. A preliminary experiment with L2 learners focusing on these four categories of the ASR errors revealed that these cases highlight the problematic speech regions for L2 listeners. Based on the findings, the PSC system is enhanced to incorporate these kinds of useful ASR errors. An experiment with L2 learners demonstrated that the enhanced version of PSC is not only preferable, but also more helpful to facilitate the L2 listening process.

Keywords: Computer-Assisted Language Learning; Second Language Listening Skill; Automatic Speech Recognition; Partial and Synchronized Caption

1. Introduction

The advancement of Information and Communication Technology (ICT) has formed new avenues of research and promoted further opportunities in different domains. The application of these technologies in language learning and teaching is known as computer-assisted language learning - CALL (Levy, 1997), which is quickly changing the teaching materials and the learning environment. CALL systems provide the materials that meet the requirements of different language learners and foster exposure to the contextualized and authentic resources including multimedia presentations, web-based distribution of print-media, radio, and TV programs (Amaral and Meurers, 2011).

While the effectiveness of using these authentic materials is undeniable, the fact that these resources are often highly challenging for L2 learners is equally evident (Gilmore, 2007). To overcome the difficulties of authentic materials, which may cause a stage of frustration and demotivation, captioning can be used (Danan, 2004). Captioning

*Corresponding author: Tel.: +81-75-753-4952;
e-mail: maryam@sap.ist.i.kyoto-u.ac.jp (Maryam Sadat Mirzaei)
URL: http://sap.ist.i.kyoto-u.ac.jp/members/maryam/ (Maryam Sadat Mirzaei)

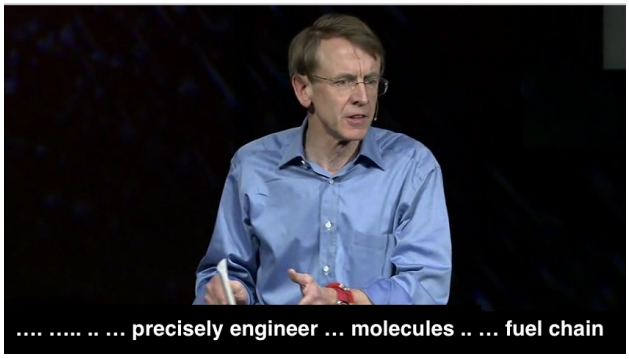


Figure 1. Screenshot of the PSC System: The caption text is presented incrementally in synch with the speech. The original transcript was: "That means we can precisely engineer the molecules in the fuel chain." TED talk by John Doerr: Salvation (and profit) in greentech.

provides the textual clues and phonological visualization of what is being heard and hence allows the use of reading while listening to comprehend the audio. Nevertheless, many learners prioritize reading the caption text over listening to the audio (Osada, 2004). These strategies assist learners in comprehending the audio but apparently do not promote the use of listening skill if not hinder it (Pujolà, 2002; Vandergrift, 2004).

In order to overcome the shortcomings of conventional captioning, we have proposed a novel captioning system called PSC (Mirzaei et al., 2014; Mirzaei and Kawahara, 2015), which automatically detects difficult words and presents them on the screen to scaffold the L2 listeners, while hiding easy words to encourage more listening than reading. Figure 1 shows a screenshot of the system. PSC synchronizes the text to speech in word-level using ASR technology. As a baseline, the detection of difficult words is realized based on three defined features: speech rate, word frequency, and word specificity. The level of difficulty and the amount of shown words in PSC is tailored to the requirement of different learners at different levels.

Studies on L2 listening difficulties indicate that learners may encounter a miscellaneous collection of factors that impede their listening (Bloomfield et al., 2010). Among those, the above-mentioned features are of special importance for the main causes of listening difficulties (Griffiths, 1992; Révész and Brunfaut, 2013). However, not all listening challenges could be explained by these features. As a result, PSC's selected words sometimes include several easy to recognize words and occasionally exclude difficult words or phrases, which highlights the importance of exploring other features. One main source of difficulties for many L2 listeners is the wrong boundary detection (Field, 2008). For many language learners, finding the right boundaries between the words in connected speech is often difficult, thus many L2 learners end up being confused with breached boundaries. Such difficulties severely hinder listening, but are not easy to detect without analyzing the nature of the speech and hence are missing in baseline PSC's selected words.

To decipher listening challenges, in this paper, we propose the use of ASR errors as a source to predict difficulties for L2 listening. ASR systems process the speech signal to generate a transcript of the audio file. This process, however, often involves some errors, which can be the product of some intrinsic speech difficulties. In this view, the performance of ASR systems is similar to L2 listeners when it comes to the transcription task. In other words, ASR errors in transcribing speech may derive from the same sources that lead to L2 misrecognition. Therefore, these errors can provide useful clues for the enhancement of PSC.

In this paper, we focus on finding useful patterns or features in the ASR errors to detect problematic speech segments for L2 listeners. The discovered patterns are tested in actual language learning environment to ensure that they cause difficulties for L2 listeners as they impede ASR performance. Then, useful errors are incorporated to the baseline PSC to provide better assistance. Finally, through an experiment, the enhanced version of PSC is compared with the baseline PSC by assessing L2 listeners' preferences and performance on using each version.

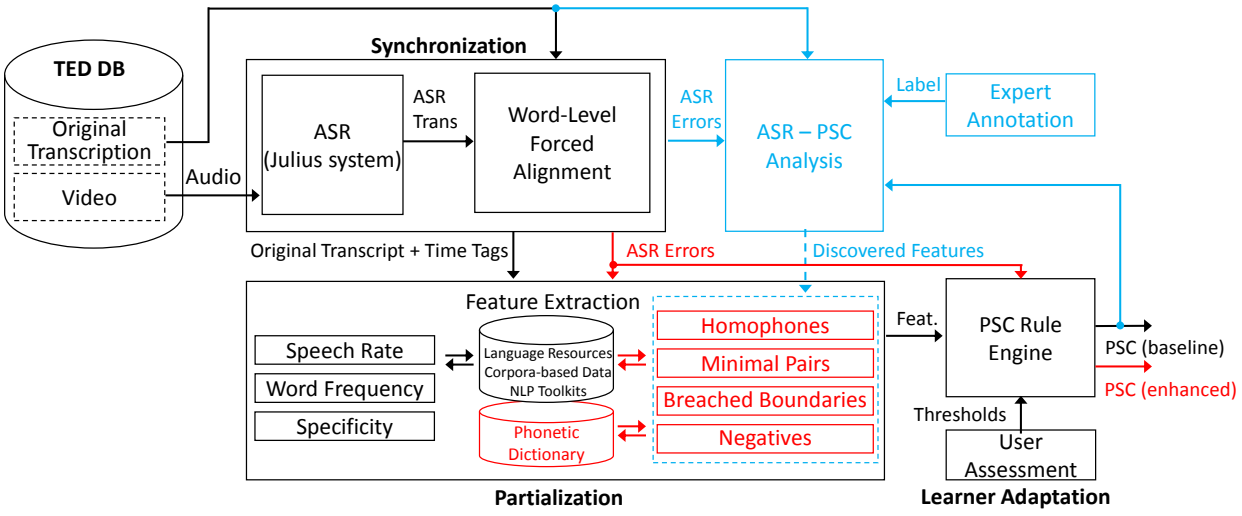


Figure 2. Schematic of PSC System. Baseline PSC (depicted in black) employs ASR system to synchronize words with their speech segments, then partialize the text based on its features. Via a root-cause analysis, ASR-PSC analysis (blue) examines several features to be incorporated into PSC's feature extraction. Once the features are identified, they are added to the feature pool of the system. These features (in red) enables the system to detect potentially difficult speech segments to be included in the caption in the enhanced PSC.

2. PSC: A Novel Tool to Train L2 Listening Skill

Full captioning has long been used as a means to facilitate L2 listening and promoting text-to-speech mapping (Danan, 2004). However, there are some criticisms against the use of full captions, which can be conceptualized around several key factors: encouraging a word-by-word decoding strategy and promoting the use of bottom-up skill (Osada, 2004), allowing comprehension of audio by just reading the text without listening (Pujolà, 2002), imposing a high level of cognitive load by providing a large amount of textual clues together with the audio (Sydorenko, 2010). We have proposed a new method of captioning in which a selected number of difficult words are shown in the caption and the rest are hidden in order to encourage listening over reading and decrease the cognitive load by providing limited, but helpful words (Mirzaei et al., 2014; Mirzaei and Kawahara, 2015). This system, which is called PSC, not only partializes the caption but also synchronizes each word to the corresponding speech segment to avoid the salient appearance of the words on the screen and obviate distraction (Figure 1).

2.1. Baseline PSC: System Overview

The system, as shown in Figure 2 (in black), uses TED talks as its database and consists of three main modules, synchronization, partialization, and learner adaptation. TED talks form the database of the system because they include the human annotated transcripts. Moreover, the talks encompass a wide range of topics delivered by trained speakers and are freely available. Accordingly, these videos can meet different interests and immerse L2 listeners in listening to inspiring talks, while being exposed to the authentic material.

To make PSC, TED talks are transcribed by our Julius ASR system (Lee and Kawahara, 2009)[1]. The acoustic and language models were trained with the TED corpus using 780 talks (180 hours) through the lightly-supervised learning method (Naptali and Kawahara, 2012). ASR transcripts are then aligned with the human-annotated transcripts through the force-alignment process to eliminate the ASR errors. This process realizes word-level synchronization and specifies the onset of each word, which in turn enables the calculation of each word's duration. The next step is the partialization stage, in which the system should detect the difficult words (in terms of listening) and decide on the inclusion or exclusion of each word in the caption based on three features: *(i)* speech rate as a dominant factor that

[1] https://github.com/julius-speech/julius

hampers L2 listening according to many studies (Griffiths, 1992; Rost, 2005), *(ii)* word frequency, which is known as an important factor influencing learners' comprehension (Schmitt and McCarthy, 1997; Bloomfield et al., 2010), and *(iii)* word specificity, which refers to the words or phrases that can be related to specific categories such as academic words, terminologies, etc. The latter factor also affects listening in a sense that many language learners are not familiar with these specific words (Révész and Brunfaut, 2013).

The final step is to tailor the caption to adjust for different language learners at different levels. At this stage, the system runs several tests to estimate the learners' current level of proficiency. These include a vocabulary size test (Nation and Beglar, 2007) to determine the learners' vocabulary reservoir and a speech rate test, based on the TOEIC samples with altered speed in order to detect the tolerable rate of speech for individual learners. The results of these tests are further consulted by L2 studies to determine thresholds on the features, hence select the words that suit the level of the learners.

2.2. Baseline PSC: Feature Calculation

The system first calculates the speech rate of the speaker, $sr(w_i)$, when delivering each individual word, w_i, where $i \in \{1, \ldots, N\}$. There are different units of measurement for speech rate including word per minute (WPM), phoneme per second (PPS) and syllables per second (SPS). WPM is not always recommended as it may be affected by pauses and changes of speech rate within a minute due to the speaker's excitement, anger, etc. (Griffiths, 1992). PPS has its own limitations as the relation between phonemes and speech rate is neither linear nor simple (Siegler, 1995). SPS, on the other hand, has fairly uniform distribution over speech rate and is more robust against variations in speech (Wang and Narayanan, 2005), thereby used as a unit of measurement in PSC. To estimate the speech rate of each word in SPS, the system calculates the duration of the word obtained from the force-alignment procedure and uses Knuth-Liang hyphenation algorithm to syllabify each word (Liang, 1983). To set the speech rate threshold, the system relies on the learner's result of the speech rate test and uses the standard rates of speech in (Tauroza and Allison, 1990).

To estimate the frequency of each word, $fr(w_i)$, the system refers to two comprehensive corpora: British National Corpus – BNC, which include 100 million words from spoken and written context, and the Corpus of Contemporary American English – COCA (Davies, 2008), which comprises 520+ million words and is the largest corpus of English based on spoken and written contexts. Along with these corpora, the system uses 25 word family lists (Nation and Webb, 2011), derived from BNC and COCA. These lists categorize all derivations of a word under a headword. Therefore words such as "*works*", "*working*" and "*worked*" are all categorized under the headword "*work*". To determine the thresholds on the word frequency, the results of the vocabulary size test (Nation and Beglar, 2007), which are compatible with the word family list (Nation and Webb, 2011), are used.

The next feature is specificity, $sp(w_i)$, i.e. if the word w_i can be categorized as an academic terminology. We referred to the academic word list of Coxhead (Coxhead, 2000), which includes 3000 academic words. Furthermore, we examined the word with COCA academic list, which is more comprehensive and up-to-date (Gardner and Davies, 2013). $sp(w_i)$ becomes 1 when w_i matches any of the entries in these lists.

Finally, the system checks for other instances of the words using corpora-based knowledge. Proper nouns (ppn), abbreviations (abb), and difficult compounds (dcp) are detected and shown in PSC because they are likely to be unfamiliar for L2 listeners. On the other hand, easy compounds (ecp), interjections (itj) and stop words (stp) (e.g. *an*, *the*, *by*) are assumed not to impose too much difficulty on L2 listeners, hence removed from PSC. These categories are detected by referring to the list of proper names, abbreviations, easy and difficult compounds in (Nation and Webb, 2011), and stop list. Accordingly, it is possible to decide about the special instances by categorizing them into keep or hide categories:

$$keep(w_i) = \mathbb{1}(w_i \in ppn \cup abb \cup dcp) \tag{1}$$

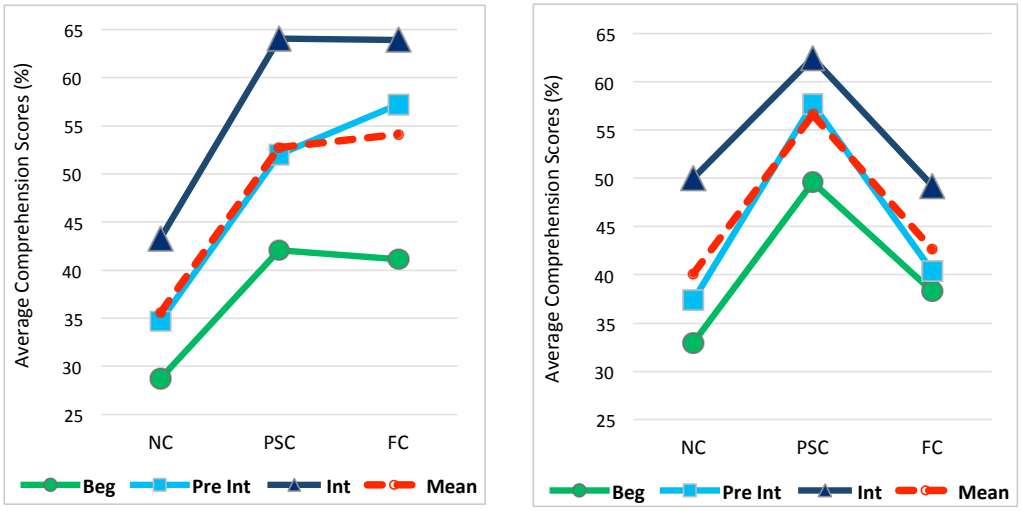
and

$$hide(w_i) = \mathbb{1}(w_i \in itj \cup stp \cup ecp) \tag{2}$$

where indicator function $\mathbb{1}(.)$ outputs 1 only if its argument is `TRUE` or positive, 0 otherwise.

The system determines to show a word in PSC if one or more features indicate that the word is difficult for the user. The user-centered features are compared with the thresholds obtained from user test results, whereas the corpora-based features are applied directly on the word.

$$show(w_i) = \mathbb{1}\Big(\mathbb{1}\big(fr(w_i) - \theta_{fr}\big) + \mathbb{1}\big(sr(w_i) - \theta_{sr}\big) + sp(w_i) + keep(w_i)\Big) \times (1 - hide(w_i)) \tag{3}$$



(a) Comprehension scores of participants under NC, PSC, and FC conditions.

(b) Comprehension scores participants on watching the remaining of the video without caption after watching the first part of it with NC, PSC, or FC.

Figure 3. Comprehension Scores of Beg (beginners), Pre Int (pre-intermediates) and Int (Intermediates) under NC (no-caption), PSC (partial and synchronized caption), and FC (full-caption) conditions.

Drawing on these features, the PSC system shows different amount of words to the learners at different proficiency levels. Meanwhile, the overall amount of shown words in PSC does not exceed 30% of the total words for any proficiency levels. In this view, PSC strives to provide the learners with a new means that allows them to rely more on their own listening skill and scaffolds them only when necessary.

2.3. Baseline PSC: System Evaluation

In an experiment with L2 learners, the baseline PSC system was compared with the full caption and no-caption conditions. The participants of the experiment were 58 Japanese students divided into three groups based on their levels of proficiency: beginners, pre-intermediate and intermediates. The TOEIC score of each group ranged from 560 ~ 599, 600~ 759, and 760~ 850 respectively. The materials were TED talks delivered by American native speakers of English to eliminate the effect of other accents. The participants watched short segments of videos under three conditions: (1) with the full caption, (2) with PSC (adjusted for each proficiency level) and (3) with no caption. The videos were rotated among participants to avoid watching the same video more than once. Learners were asked to answer several comprehension questions after watching the videos. The aim was to compare the effect of different captioning conditions on the comprehension scores.

Results are shown in Figure 3(a). As the graph shows, participants' comprehensions significantly increased when they received caption (either the full caption or PSC) as compared to the no-caption condition. ANOVA analysis of the results revealed that the scores gained under full caption condition (M=54.25, SD= 17.33) and PSC condition (M=52.89, SD=19.39) were not statistically different [F (1, 57)=.25, p =.62]. This finding suggests that PSC with less than 30% of the text can lead to a similar level of comprehension as the full caption. The same tendency is observed across different proficiency levels, indicating that PSC can adjust its content to the level of the learners and provide adequate assistance for each proficiency level.

Next, the participants were asked to watch the rest of the same clip without any caption and answer some comprehension questions. This part was intended to check whether watching the video with the full caption, PSC or no caption first would have any effect on watching the rest of video without any caption. The results are illustrated in Figure 3(b). As can be seen, upon receiving the full caption first and no-caption next, the scores had an abrupt de-

crease (M = 42.65, SD = 13.37), as compared with receiving PSC first and no-caption next (M = 56.59, SD = 17.34). Although indicating a short-term adaptation, this finding suggests the usefulness of PSC in preparing the learners for real-world listening situations without using textual clues.

While the effectiveness of the baseline PSC system is confirmed, there are still a number of problems that should be addressed. For instance, PSC often includes too easy or insignificant words (e.g. *one*, *look*, *every*, etc.). On the other hand, there are some instances of difficult words or phrases that PSC fails to show in the caption (e.g. *avian flu*). In order to address these issues, we investigate the use of ASR errors as a potential source to provide hints on difficulties of speech for L2 learners.

3. ASR Errors and L2 learners' Recognition Difficulties

A number of studies have investigated the relationship between ASR errors and native or non-native recognition errors, which are known as ASR-HSR (human speech recognition) research. However, to date, the comparison between ASR errors and L2 learners' recognition errors, the term we coined as ASR-L2SR, has not been closely examined. Accordingly, the objectives of this paper are to perform such comparison and determine whether ASR errors can highlight challenging speech segments and signal recognition difficulties for L2 learners, hence provide insights for PSC enhancement.

3.1. ASR - HSR

Many studies have investigated the ASR errors and HSR difficulties with the purpose of bridging the gap between the two and incorporating HSR findings to improve ASR performance (Moore and Cutler, 2001; Scharenborg et al., 2003; Meyer et al., 2006; Scharenborg, 2007; Vasilescu et al., 2012). The subjects of these studies are either native speakers of the target language or non-native speakers with no knowledge of the target language (e.g. Chinese with no knowledge of French tested with French audio, which includes words with the maximum phonetic similarity between the two languages). Some studies have emphasized the importance of conducting fair HSR-ASR comparisons by restricting the influence of background information, using logatomes/pseudowords (Meyer et al., 2006). Findings of these studies revealed that the intrinsic variation of speech such as speaking rate, pitch, style, speaker physiology, age, dialect, and accent has a significant influence on the overall recognition of both HSR and ASR (Meyer et al., 2011). Through these studies, researchers attempt to unfold solutions for improving the ASR systems (Shen et al., 2008). Inspired by such comparisons, we investigate the similarity or difference between ASR and L2SR in order to identify L2 listeners' difficulties.

3.2. ASR - L2SR

In ASR-HSR studies, ASR errors are counted as the negative product of the systems and the comparison is used to shed light on possible improvement to decrease the number of ASR errors. The erroneous output of the ASR system deteriorates the quality of the ASR-generated transcript, which is why such transcripts are not appropriate for L2 learners (Felps et al., 2012). In the context of L2 learning, there is low tolerance for the errors and even error rates below 5% are considered too high for the intended users (Vasilescu et al., 2011). In this study, however, when comparing ASR with L2SR, the ASR errors are viewed as a prospective predictor of speech difficulties and yield a model to elucidate L2 listening difficulties. In general, ASR errors arise either when there is an intrinsic difficulty in the speech (language bias) or when there is a limitation in the acoustic or language model of the system (model bias) (Vasilescu et al., 2011). Accordingly, some ASR errors may indicate to the sources of difficulty that hinder L2 listening. To investigate this hypothesis, we first perform a comparative analysis between the research on L2 listening difficulties and the studies focusing on ASR errors.

There are a number of factors accounted for L2 listening difficulties, some of which have already been explained and covered by PSC features. For instance, speech rate, whether too fast or too slow, is the main source of difficulty for many L2 learners (Griffiths, 1992). This argument also holds for ASR systems, which are largely influenced by variations in speech rate (Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001). The frequency of the words, as another factor considered by PSC, affects L2 listening indicating that low-frequency words often confine learner's attention, preventing them from following the rest of the audio (Bloomfield et al., 2010). Findings on the

ASR error analysis emphasize the importance of this factor in the performance of the system (Shinozaki and Furui, 2001). Similarly, a number of other factors such as co-articulation, pronunciation, speaking style, age, physiology, and emotions lead to ASR difficulties (Benzeghiba et al., 2007), which also affect L2 listening (Bloomfield et al., 2010). For instance, pronunciation can be unclear due to assimilation, reduction, etc, which in turn causes a lot of recognition difficulties for language learners. Moreover, stress, intonation patterns, and accent affect not only L1 but also L2 listening comprehension (Osada, 2004; Bloomfield et al., 2010). Word length has also been found to be a useful predictor of higher error rates in ASR systems (Shinozaki and Furui, 2001). Comparably, the length of a word has a strong effect on its recognition when it comes to L2 listening (Field, 2008). The class of the word is another influential factor. Recognition of open class words (e.g. noun and verbs) result in a lower ASR error rate compared to closed class words (e.g. prepositions and articles) (Goldwater et al., 2010). Similarly, recognition of content words is easier than function words for L2 listeners such that nouns predominate over prepositions (Field, 2008).

Overall, there are so many possible factors affecting L2 listening difficulty (Bloomfield et al., 2010), which may be correlated and some of them are not so certain to be modeled. In this view, the use of ASR errors as an indicative of listening difficulties can provide important insights for discovering such factors. However, the performance of the ASR is important in this analysis and it should be comparable to the level of L2 learners or L2SR.

4. ASR Error Analysis for Extracting Features

4.1. ASR Error Extraction

To perform a root-cause analysis on the ASR errors, 70 TED talk, approximately 21 hours, were transcribed by our Julius ASR system and the output transcripts were aligned with human-annotated transcripts to detect the mismatches. In Table 1, the errors are categorized into substitution, deletion and insertion categories. As the table indicates, ASR error rate is 21.34% and the majority of errors are in substitution category (17.53%). For ASR to be used for such a purpose it should have "reasonable" performance as ASRs with very high accuracy will not provide us with enough errors and those with very poor performance may not provide useful errors for the learners.

4.2. Trend Analysis of Baseline PSC Features in ASR Errors

To begin the analysis, ASR errors are examined to discover the underlying trends. This analysis is performed based on PSC's baseline features (speech rate and word frequency).

The speech rate of the ASR errors was calculated in SPS and its trend was explored in four bins: slow (~ 3.83 SPS) , moderate (3.83 ~ 5.33 SPS), fast (5.33 ~ 8 SPS) and too fast (above 8 SPS) based on the standard rates of speech (Tauroza and Allison, 1990). Figure 4 (right) illustrates that the ASR error rate increases when the speech rate becomes too fast. The trend is in line with those reported in L2 studies (Nitta et al., 2010). With increasing speech rate, L2 learners are more prone to make listening mistakes (Rost, 2005). Furthermore, misrecognition increases among L2 listeners when listening to audios with too slow speech rate (Griffiths, 1992).

Similar trend analysis is performed on ASR errors considering the word frequency feature. The frequency of words in ASR errors is calculated by referring to Nation's family lists (Nation and Webb, 2011) along with BNC and COCA. The frequency is partitioned into 3 bins - high frequency (~3000 word families), mid-frequency (3000~ 6000 word families) and low-frequency (above 6000 word families) according to (Schmitt and Schmitt, 2014). Figure 4

Table 1. ASR Error Analysis on TED Talks

Categories	**Frequency**
Total Words	206,469
Correct	162,407 (78.66%)
Errors	44,062 (21.34%)
Substitution	36,193 (17.53%)
Insertion	4,139 (2.00%)
Deletion	3,730 (1.81%)

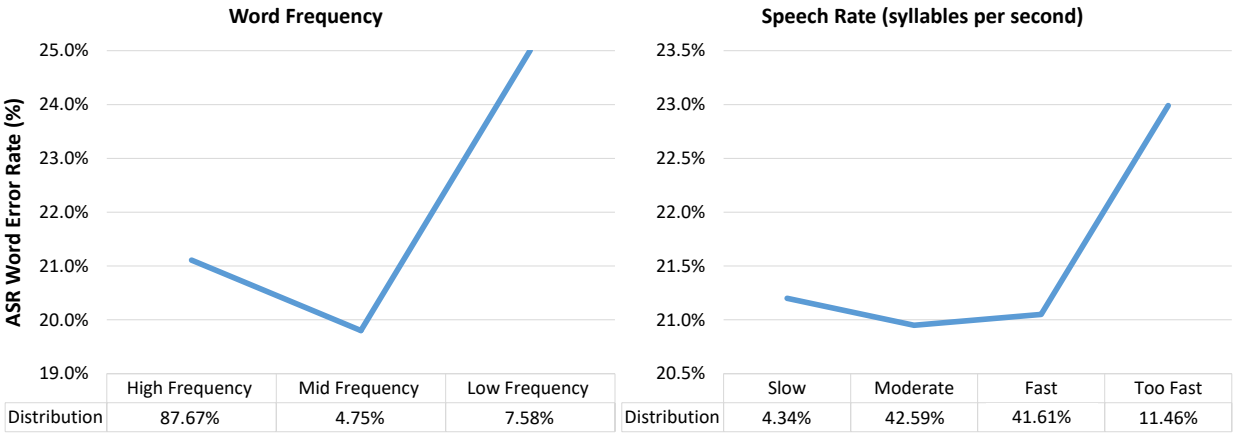


Figure 4. Trend Analysis on ASR Errors

(left) demonstrates that ASR generates more errors when encountering low-frequency words. This is in line with L2 studies noting that low-frequency words lead to L2 listening difficulties, while high-frequency words are generally accurately recognized (Bloomfield et al., 2010). However, ASR performs the best when receiving mid-frequency words considering that high-frequency words include many function words with short length and pronunciation variations.

The analysis revealed that similar trends are discovered on ASR errors and L2 listeners' misrecognition, considering PSC features (speech rate and frequency). Moreover, the trends we extracted from our ASR system are in line with those reported in previous studies on ASR error analysis using other ASR systems.

4.3. Comparison of ASR Output and PSC Selection

Findings of the ASR trend analysis suggested similar recognition difficulties for both ASR and L2 listeners regarding speech rate and word frequency. These two features are used by PSC to detect difficult words in listening materials. In this view, ASR errors and PSC selected words are both considering difficulties in speech and hence may share some similarities. To investigate any plausible similarities, PSC was generated for all 70 TED videos, controlling for high speech rate, low frequency, and specific or academic words. The selected words by PSC were then compared against ASR errors to find the degree of overlap.

Table 3 demonstrates the result of this comparison and indicates that 22% of the cases are common between ASR errors and PSC shown words (difficult cases), while many of ASR errors (78%) could not be covered by PSC's features. Furthermore, the table indicates that 83% of ASR correct cases were regarded as trivial for L2 listeners and not shown by PSC. Nevertheless, 17% of these ASR correct cases are still shown in PSC, implying that these should be hidden in PSC. This finding highlights the importance of investigating these categories to discover the underlying features.

Table 2. ASR versus Baseline PSC Comparison (70 TED Talks)

ASR vs. Baseline PSC	ASR Correct (78.66%)	ASR Errors (21.34%)
Baseline PSC Shown Words (17.80%)	13.13%	4.67%
Baseline PSC Hidden Words (82.20%)	65.53%	16.67%

Table 3. ASR versus Baseline PSC Comparison (70 TED Talks)

ASR vs. Baseline PSC	ASR Correct (78.66%)	ASR Errors (21.34%)
Baseline PSC Shown Words (17.80%)	13.13%	4.67%
Baseline PSC Hidden Words (82.20%)	65.53%	16.67%

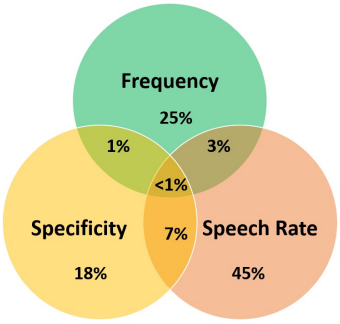


Figure 5. Feature analysis in ASR correct & PSC shown cases

4.4. Analysis on ASR Correct & PSC Shown Cases

Analysis on ASR correct & PSC shown cases identified the reasons for PSC's decision to include these words. As Figure 5 shows, speech rate is the primary factor that explains for the appearance of these words in PSC (45%). However, analysis of this category revealed that majority of them are unnecessary or not useful in terms of comprehension or recognition (e.g. *every*, *who*, etc.). It can be suggested that ASR correct cases can provide insightful clues on refining the speech rate threshold. While a default threshold is set for PSC based on user's tolerance and literature standards (θ_{sr}), a secondary threshold can be introduced to apply a strict margin on ASR correct cases in order to exclude easy words. Therefore, the primary threshold remains for ASR erroneous cases ($\theta_{sr}^{ASRcor(w_i)=0} = \theta_{sr}$), and the secondary threshold acts above the primary one in ASR correct cases, $\theta_{sr}^{ASRcor(w_i)=1} = \theta_{sr} + \Delta_{sr}$. $ASRcor(w_i)$ is a binary flag indicating the correctness of ASR output for word w_i according to forced-alignment unit ($ASRcor(w_i) = 0$ signals the ASR error status), and Δ_{sr} is an added margin for ASR correct cases.

The second factor that led to the inclusion of easy cases into PSC corresponds to the word frequency feature (25%). Examining this group revealed that the frequency feature generally votes for useful and essential words to appear in PSC, implying that the feature is very effective.

The third feature is word specificity, which brings the academic words in PSC (18%). Investigating this category clarified that many of the academic words in this group are too frequent to be unfamiliar for L2 learners. Examples include words such as *science, research, etc*. This finding suggests introducing a frequency threshold for specific words ($\theta_{sp}^{ASRcor(w_i)=1}$) to decide on their appearance in PSC rather than simply presenting them all in the caption. In the ASR error cases, however, such words should be presented, $\theta_{sp}^{ASRcor(w_i)=0} = 0$.

Through a comprehensive comparison between ASR correct & PSC shown category, it was found that (1) PSC's speech rate threshold should be tuned based on ASR clues, (2) the word frequency feature should be prioritized and (3) a frequency threshold for specific words and proper nouns should be taken into account based on ASR erroneous and correct cases. These measures will foster discarding the impotent cases from PSC and provide space for encompassing more useful cases.

Considering these findings, equation (3) will be changed to:

$$show(w_i) = \mathbb{1}\Big(\mathbb{1}\Big(fr(w_i) - \theta_{fr}\Big) + \mathbb{1}\Big(sr(w_i) - \theta_{sr}^{ASRcor(w_i)}\Big) + \mathbb{1}\Big(fr(w_i) - \theta_{sp}^{ASRcor(w_i)}\Big) \times sp(w_i) + keep(w_i)\Big) \times \Big(1 - hide(w_i)\Big) \tag{4}$$

4.5. Analysis on ASR Error & PSC Hidden Cases

The next comparison deals with analysis on ASR erroneous & PSC hidden cases in order to discover the useful candidates for PSC. In this view, we conducted a root-cause analysis on the ASR errors not shown by PSC, which are classified into the following categories (Table 4):

Table 4. Distribution of patterns and their usefulness in ASR error & PSC hidden category for substitution errors (12.54% of all words). The usefulness is calculated for each category considering the number of useful labels to all words of the category.

Category	Occurrence Ratio (%)	Usefulness (%)
(1) Homophones	0.20%	82.34%
(2) Minimal Pairs	0.34%	86.18%
(3) Negatives	0.20%	71.92%
(4) Breached Boundaries	3.75%	63.69%
(5) Verb Inflections	0.62%	22.19%
(6) Noun Inflections	0.71%	26.33%
(7) Determiners	1.83%	0.90%
(8) Interjections	0.21%	4.15%
(9) Derivational Suffixes	0.59%	29.47%
(10) Stop List	3.62%	18.22%
(11) Unknown Sources	0.47%	36.99%

1. Homophones: words with same pronunciation, but different spelling and meaning (e.g. *see* instead of *sea*, *pail* instead of *pale*, *feet* instead of *feat*). Homophones can deteriorate L2 listening by activating several candidates and imposing a high-level semantic analysis for learners to make a distinction (Field, 2003; Weber and Cutler, 2004).
2. Minimal pairs: words that differ only in one phonological element (e.g. *fund* instead of *fun*, *think* instead of *sink*, *park* instead of *bark*). Recognition of these pairs is reported to be difficult for language learners according to L2 studies (Weber and Cutler, 2004).
3. Negatives: cases in which the use of prefixes, suffixes or negative particle changes an affirmative word into a negative one (e.g. *can't* instead of *can*, *atheism* instead of *theism*, *illegal* instead of *legal*). The difference between the negative and affirmative forms in such cases is subtle, making them difficult to distinguish. As a result, many L2 learners frequently misrecognize these cases and misunderstand the meaning (Field, 2003).
4. Breached boundaries: cases in which the boundaries are either converged or diverged from the right setting point (e.g. *in close* instead of *enclose*, *it was an eagle* instead of *it was illegal*, *very ability* instead of *variability*, *thus he sent his drill in* instead of *dusty senseless drilling*). Breached boundaries are among the most problematic and common mistakes that impede L2 listening (Field, 2003), but are difficult to predict.
5. Verb inflections: cases in which the verb is modified to express different grammatical categories such as tense (e.g. *played* instead of *play*), voice (e.g. *played* instead of *was played*), person (e.g. *he play* instead of *he plays*), etc. The inflection of verbs is also called conjugation. These cases are generally easier to perceive if the contextual information is taken into account. While ASR systems generate plenty of such errors, these cases do not hinder comprehension.
6. Noun inflections: nouns are inflected to make a plural form (e.g. *books* instead of *book* and *women* instead of *woman*) and to show possession (e.g. *girls'* instead of *girls* and *Mary's* instead of *Mary*). This is another common category of ASR errors that is not necessarily the case for L2 learners.
7. Determiners: this category includes articles (*a, an, the*), possessives (e.g. *her, their*), demonstratives (e.g. *this, these*), interrogatives (e.g. *who, whose*) and quantifiers (e.g. *any, many*). The majority of these cases are included in the stop list, which explains why the words in this category are hidden from PSC. While L2 studies suggest that learners are often prone to make recognition mistakes on this category due to being inattentive to function words, these cases are normally easy to disambiguate.
8. Interjections: words or expressions used to signify the speaker's strong feeling, spontaneous emotion or reaction and includes fillers (e.g. *uh, em*), exclamations (e.g. *wow!*), etc. This category is of special importance when it comes to speaking, but the use of video along with the audio provides enough visual information to recognize these expressions.
9. Derivational suffixes: suffixes added to the word end to make a new word. Suffixes can attach to nouns to make an adjective, generate a verb or create another noun (e.g. *beauty, beautiful*, *beauty, beautify* and *bag, baggage*).

They can also attach to a verb to create a noun or adjective (e.g. *depart, departure* and *compare, comparable*) or be added to an adjective to make an adverb or a noun (e.g. *clear, clearly* and *faithful, faithfulness*), etc. Since the root of these words is in most cases similar, it is easy to switch between them while listening, Hence this category does not seem to hinder comprehension.

10. Stop List: cases which are usually the most common words in a language and include short function words, such as propositions (e.g. *at*, *on*, *up*). This category also includes "to be" verbs, "WH" questions, etc.
11. Unknown sources: there is no straightforward explanation for these errors. Examples include: *call of ice time* instead of *Albert Einstein* and *in Italy on and off* instead of *at least long enough*.

While some of these categories seem to have strong potential to cause L2 listening difficulties, others are assumed to be impotent factors because their inclusion in PSC will barely make any improvement. We annotated the ASR substitution errors on 70 TED talks (36193 words) to distinguish between useful and useless ASR erroneous cases, regardless of their categories. The annotator watched each video and labeled all ASR substitution errors as either useful or not useful, i.e., to examine *(i)* if a similar misrecognition can be expected by L2 listeners on ASR errors, and *(ii)* if the inclusion of such cases into PSC will provide L2 learners with useful information, which in turn facilitate listening. A subset of the videos including 7 TED talks with 2812 words in ASR substitution errors is annotated by another annotator to compare the agreement level between the two annotations. Given that both annotators had linguistic backgrounds and received a set of clear instructions and objectives, the comparison showed 91.8% of inter-annotation agreement with Cohen's $\kappa = 0.81$, which indicates a very high-level agreement.

The results of annotation on the usefulness of each category of ASR errors to appear in PSC are presented in Table 4. As the table presents the annotation results show that the first four categories of ASR errors include the majority of the useful cases and can explain 68.78% of the useful ASR errors & PSC hidden category. Minimal pairs have the largest ratio of usefulness with 86%, followed by homophones (82%), negatives (72%), and breached boundaries (64%), respectively. Interestingly, these cases were identified to be particularly challenging for language learners according to L2 studies (Field, 2003; Weber and Cutler, 2004).

On the other hand, cases such as verb inflections, interjections, determiners lack convincing amount of useful cases to be embedded in the PSC and their inclusion would contaminate the caption with many trivial words. Table 4 shows that in spite of involving more than 31% of useful words, the ratio of useful to all words in each of these categories is relatively low. As a result, we regard them as impotent factors that are not to be incorporated into PSC.

4.6. Experimental Evaluation of Additional Features

We conducted an experiment with L2 listeners in order to confirm the usefulness of the four features of ASR errors (homophones, minimal pairs, negatives and breached boundaries) for detecting problematic speech segments.

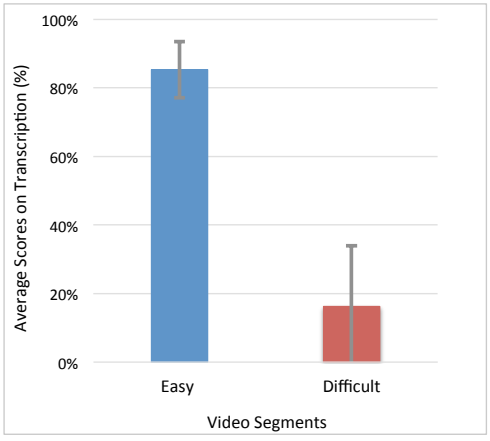
4.6.1. Participants

The participants of this study were 11 Japanese and 10 Chinese students (8 females and 13 males), who were undergraduate and graduate students at our university, majoring in different fields such as engineering, law, science, etc. All participants had TOEIC scores (or equivalents) of above 750 and their proficiency level was considered as intermediate.

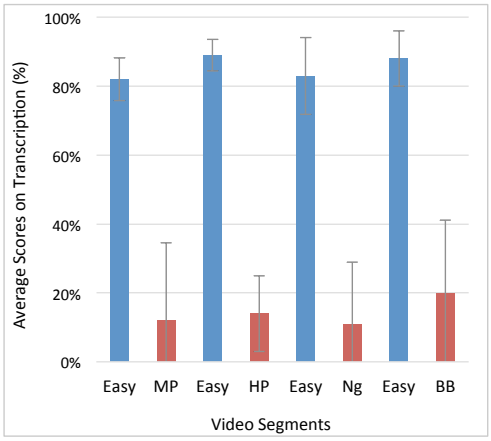
4.6.2. Materials

We selected 20 TED talks, opting for talks delivered by American native speakers in order to eliminate the effect of other accents. All talks were delivered by single speakers. From each video, two short segments (25~35 seconds) were selected based on the following criteria:

1. A segment including one category of ASR errors i.e. homophone, minimal pairs, negatives or breached boundaries that the baseline PSC failed to detect ("difficult cases" that may cause problem for L2 listeners);
2. A segment devoided of ASR errors, which PSC also determined to exclude from the caption for being too easy or impotent ("easy cases" as a control case).



(a) Participants' scores on transcribing easy (ASR correct) versus difficult (ASR error) video segments



(b) The distribution of participants' scores on transcribing difficult segments including ASR errors: MP (minimal pair), HP (homophone), NG (negatives), and BB (breached boundaries) along with the respective easy segments of the same video.

Figure 6. Transcription scores on segments of ASR errors vs. ASR correct

The former was selected from those parts of the video, in which the ASR failed to generate a correct transcription due to the presence of minimal pairs, homophones, negative forms, and breached boundaries. The latter cases were chosen as a control factor to make sure that there is a difference on the performance of L2 listeners for transcribing easy versus difficult speech segments. We randomly selected one sample from each criterion for each video and randomized the order of all 40 samples.

4.6.3. Procedure

The participants were asked to listen to these pieces of connected speech until the video was paused. Upon encountering a pause, the participants were asked to immediately transcribe the last few words they have just heard. To control for short-term memory span, learners were expected to provide the transcriptions of 4∼6 words, which included the target word(s). The videos were automatically paused at an irregular interval. The participants were neither aware of the time of pauses nor aware of the target word(s). They could watch each video only once. At each pause, blanks appeared on the screen in order to notify the participants to input the words they have heard. A timer was set for answering each question to avoid the participants from overthinking and analyzing, thereby allowing them to immediately input what they have recognized. Spelling errors were ignored unless affected the meaning. The test was launched online and took 40 minutes to complete.

Through this experiment we aimed to answer the following research questions:

1. Do learners easily transcribe those parts of the video that ASR correctly transcribed and PSC hid for being too trivial?
2. Do learners have difficulty in transcribing those parts of the video that ASR system failed to recognize?

4.6.4. Results

Figure 6(a) shows the statistics of participants' scores on transcribing (1) easy segments of the videos i.e., words correctly transcribed by ASR and (2) difficult segments of the videos i.e. the words including ASR errors. As the figure shows, participants' scores on transcribing the easy segments ($M = 0.85, SD = 0.08$) is significantly higher than their score on transcribing difficult segments ($M = 0.16, SD = 0.18$).

Figure 6(b) illustrates the distribution of participant's scores on each category of difficult segments against the corresponding easy segment selected from each video. The analysis on participants' scores showed a significant

difference on all categories of homophones, minimal pairs, negatives and breached boundaries as compared with their respective easy segments. The results provide a positive answer to our first and second research questions, suggesting that (*i*) easy segments caused substantially fewer problems for L2 learners, (*ii*) the participants share difficulty with ASR systems in transcribing homophones, minimal pairs, negatives and breached boundaries. The findings of this experiment confirm the usefulness of ASR errors in detecting problematic speech segments for L2 listeners.

4.7. Feature Extraction from ASR Errors

Figure 2 (red) depicts the extension on the PSC system by extracting the four categories of features derived from ASR errors (homophones, minimal pairs, negatives and breached boundaries). In this enhanced system, the forced-alignment unit not only synchronizes the ASR output with the original transcript but also highlights the erroneous segments of ASR transcript, which in turn, is used to extract the new set of features.

A given word w_i in the original transcript is aligned with an erroneous phrase $\hat{w}_i$ generated by the ASR system. We define four feature extraction functions: homophones $hp(w_i, \hat{w}_i)$, minimal pairs $mp(w_i, \hat{w}_i)$, negatives $ng(w_i, \hat{w}_i)$ and breached boundaries $bb(w_{i-1:i+1}, \hat{w}_i)$. The first three functions mark the word w_i if a homophone, minimal pair, or negative instance of this word exists in $\hat{w}_i$. The last function, $bb(w_{i-1:i+1}, \hat{w}_i)$, marks the word w_i if a breached boundary instance between word w_i and its predecessor (w_{i-1})/successor (w_{i+1}) is detected in $\hat{w}_i$. These functions output a binary value and adding them to eq(4) is straightforward:

$$\begin{aligned} show(w_i) =& \mathbb{1}\Big(\mathbb{1}\Big(fr(w_i) - \theta_{fr}\Big) + \mathbb{1}\Big(sr(w_i) - \theta_{sr}^{ASRcor(w_i)}\Big) + \mathbb{1}\Big(fr(w_i) - \theta_{sp}^{ASRcor(w_i)}\Big) \times sp(w_i) + keep(w_i)\Big) \times \Big(1 - hide(w_i)\Big) \\ &+ \mathbb{1}\Big(hp(w_i, \hat{w}_i) + mp(w_i, \hat{w}_i) + bb(w_{i-1:i+1}, \hat{w}_i) + ng(w_i, \hat{w}_i)\Big) \times \Big(1 - ASRcor(w_i)\Big) \end{aligned} \tag{5}$$

It should be mentioned that words added to PSC by the new features should not be suppressed by the $hide(w_i)$ feature.

It would be possible to formulate a discriminant function, such as logistic regression model, using these features with some weights and optimize them using the annotated data, but these new features (derived from ASR errors) are basically binary and mutually exclusive, therefore a weighted combination of them would not be effective.

5. Using ASR Errors to Enhance Baseline PSC

Findings from the previous experiment showed the usefulness of four categories of ASR errors indicating that these cases can be embedded into the PSC system to scaffold the learners on difficult speech segments. Accordingly, we enhanced the baseline PSC system to provide better assistance for L2 listeners.

5.1. Augmenting Baseline PSC with ASR Error-derived Features

The main idea is to view an ASR system as a model of L2 listener; thereby developing the enhanced PSC by:

1. Treating ASR correct cases as easy speech segments, which PSC can disregard;
2. Considering ASR errors as challenging speech segments, which PSC should encompass to better scaffold the learners.

To this end, similar to the baseline PSC, the videos are transcribed using our Julius ASR system (v4.3.1), which was trained on TED corpus. The ASR transcript is then aligned with the original transcript to make a word-level correspondence between the two, and detect erroneous segments in ASR output. Meanwhile, the matched words of the ASR transcript lend their time tag to their counterpart in the original transcript to enable the calculation of the speech rate as described in Section 2.2. Using the available language-based and corpora-based resources and NLP tools in the Feature Extraction unit, the word frequency, and specificity features are also extracted.

Next, the erroneous segments of ASR transcript along with its corresponding original transcript are sent to the Feature Extraction unit to automatically extract the new features. The Feature Extraction unit uses a phonetic dictionary on top of language models, corpora-based lists, and NLP tools. The pair of phrases is then scanned for possible

matches of homophones, minimal pairs, and negative cases. Also, the ASR output and the transcript are compared to find possible breached boundaries.

At this stage, the procedure starts with detecting homophones and minimal pairs. To this end, the phone sequence of ASR hypothesized output is compared with the phone sequence of the transcript word(s). We extract these phone sequences from CMU dictionary, selecting the closest entry in case several phonetics are available for one word. Then, the Levenshtein distance between the phone sequences of each word in ASR transcript and the human transcript is calculated. This distance is the number of deletions, insertions, or substitutions required to transform the first phone sequence to the second one. We mark a word in original transcript as homophone or minimal pair case, if a word with a distance of zero or one exists in the erroneous ASR transcript. Detection of breached boundaries is relatively difficult since there is no one-to-one correspondence between the pairs (the ASR-hypothesized output and the original transcript). In such cases, ASR errors are often "bursty" (Chen et al., 2013) and include a number of words forming an erroneous phrase, which is aligned with a phrase in the original transcript through the force-alignment procedure. The distance between these two pairs is not determined a priori, which renders breached boundary detection difficult. Thus, every possible combination should be considered.

Accordingly, the system detects these features based on the following procedure:

1. Two words are considered as *homophone* if they have identical phonetic transcript i.e. with Levenshtein distance of zero, but different writings (e.g., *rain* `/R EY N/` and *reign* `/R EY N/`). Special cases such as different possible pronunciations of the same word, or American and British spelling of a word are excluded.
2. Two words were categorized as *minimal pairs* if their phonetics have a Levenshtein distance of one. This enables detecting different types of minimal pairs: initial consonant (e.g., *pin* `/P IH N/`, *bin* `/B IH N/`), vowels (e.g., *bin* `/B IH N/`, *bean* `/B IY N/`), and final consonant (e.g., *hat* `/HH AE T/`, *had* `/HH AE D/`). This category also includes the third person (*work* an *works*) in the present tense and past tense for regular verbs (*work* and *worked*), which were disregarded and added to the impotent factors.
3. *Negative* cases are detected by considering the negative particle "*not*" and attending to the syntax of the word, looking for prefixes and suffixes that form negation. Furthermore, negative short form, i.e., words with "n't" are considered. Different types of negative occurrences are handled: *(i)* ASR transcript includes a negative word, whose affirmative form appeared in the original transcript (e.g. "*shouldn't*" in ASR and "*should've*" in transcript) or vice versa, and *(ii)* the original transcript includes a negative word whose affirmative form or the equivalent form is missing from the ASR output (e.g. *can't* in transcript missing in the ASR output).
4. To detect breached boundaries, every boundary in the original transcript phrase and the ASR error sequence is checked based on the following rules. In other words, we generate possible candidates for insertion, deletion, and relocation of boundaries in the original transcript, apply the rules and check if the modified boundaries can be found in the ASR error sequence. To begin with, every pair of the words excluding those in homophones or minimal pair categories were examined to check if any breached boundaries could be detected. To this end, the phone sequence of the ASR phrase is concatenated and compared against the phone sequence of the phrase in the original transcript. In the simplest case of breached boundaries, the phonetic sequences are identical while the corresponding words themselves are different. However, such boundary cases are very rare. To address this issue, we draw on L2 studies to find the prominent breached boundary patterns discovered by examining L2 listeners' transcription corpora. These cases have been analyzed by psycholinguists and are known as the slips of the ear, which include many word-boundary misrecognition (Cutler, 1990). The followings were known as the most dominant and common patterns to predict listeners' segmentation strategies:

    - Strong-syllable strategy (Cutler, 1990): Learners tend to insert word boundaries when they encounter a strong syllable so that the stressed syllable is set as the beginning of the word (e.g. "*disguise*" heard as "*the skies*"). Also, learners tend to delete the boundary before a weak syllable and thus merge the words (e.g. "*ten-to-two*" heard as "*twenty to*"). The CMU dictionary is consulted to look up the stress patterns of the words in order to detect this kind of breached boundaries.
    - Assimilation rule (Field, 2003): Learners have difficulty in setting the right word boundaries due to the common phonological process, which alters a word ending sound in expectation of the following sound (e.g. "*this shirt*" as "*thi–shirt*"). The assimilation rule is realized using Gimson's English assimilation standards (Cruttenden, 2014), which are quite systematic and follow restricted patterns.

Table 5. Baseline PSC versus Enhanced PSC (70 TED Talks)

ASR vs. Baseline PSC	ASR Correct: Easy Cases (78.66%)	ASR Errors: Difficult Cases (21.34%)
Baseline PSC Shown Words (17.80%)	13.13%	4.67%
Baseline PSC Hidden Words (82.20%)	65.53%	16.67%
Enhanced PSC Shown Words (17.77%)	8.95%	8.82%
Enhanced PSC Hidden Words (82.23%)	69.71%	12.52%

- Frequency rule (Cutler, 1990): Learners have a general tendency to insert word boundaries in order to perceive more frequent words than the actual target word. They scan continuous speech for matches between sequences of sounds and items of known vocabulary, which may cause word boundary misperception (e.g. "*achieve her way*" heard as "*a cheaper way*"). This is in line with the studies on ASR errors indicating that out-of-vocabulary words are broken into multiple in-vocabulary words causing insertion errors and false boundaries (Chen et al., 2013). COCA is used to extract the frequency of the words and check for the occurrence of the frequency rule. However, the frequency of function words is ignored for being dominantly high, following the argument in (Cutler and Butterfield, 1992) on frequency analysis of a sequence including content and function words.
- Resyllabification (Field, 2008): Learners may receive false boundary cues because of resyllabification, in which the final consonant of a word attaches to the following syllable (e.g. "*made out*" heard as "*may doubt*"). Resyllabification is detected based on word sequence structure, considering the occurrence of consonants in the final syllable of a candidate word, attached to the onset syllable of the following word.

Each word w_i and its features ($fr(w_i)$, $sr(w_i)$, $sp(w_i)$, $keep(w_i)$, $hide(w_i)$, $hp(w_i, \hat{w}_i)$, $mp(w_i, \hat{w}_i)$, $ng(w_i, \hat{w}_i)$ and $bb(w_{i-1:i+1}, \hat{w}_i)$) are then sent to PSC Rule Engine to determine whether it should be shown or not. Based on ASR correctness flag (ASR) for word w_i, this unit selects the appropriate procedure and thresholds to make a decision, that is summarized in eq (5). If the ASR transcribes the word correctly, the word frequency, speech rate, and the frequency of specific words are compared with the strict thresholds, and the show-decision is then filtered out if the hide list chooses to hide the word. On the other hand, if the ASR contains an error, more moderate thresholds (the ones obtained from user assessments) is used. After this primary stage, if the word is detected as a homophone, minimal pair, negative or breached boundary candidate, it will be included in the PSC regardless of being on the hide list.

5.2. Statistics of Baseline PSC versus Enhanced PSC

Table 5 indicates the statistical comparison between the baseline PSC and the enhanced PSC with regard to ASR correct and erroneous cases in 70 TED talks. As the table presents, the enhanced PSC version includes 41% of ASR errors, compared with the baseline PSC, which includes 22% of ASR errors, yet the enhanced PSC shows 18% of the total words, which is comparable to the percentage of words shown in the baseline (18%). The comparable quantity of the shown words in both versions can be explained by the reduction seen in ASR correct & PSC shown category. Applying a frequency threshold for academic words based on the ASR output along with a similar adjustment in the speech rate threshold led to the reduction by 4.55% in the amount of shown words.

5.3. Experimental Evaluation of Enhanced PSC

While the baseline PSC was compared with full captioning in terms of comprehension, the enhanced PSC is compared with the baseline focusing on recognition of specific modified parts. When learners' listening is evaluated on a particular phrase, overall comprehension is no more suitable as it applies to a broader scope. Thus, we designed an experiment including a transcription test and a paraphrase test. The former is similar to our previous experiment and the latter is a test that focuses on the recognition of a specific part of listening material (Buck, 2001).

5.3.1. Participants

In this experiment 36 Japanese and 2 Chinese undergraduate students, mostly from engineering fields, participated. The participants' TOEIC scores ranged from 450 to 560. All participants were enrolled in a CALL class, where the experiment was held.

5.3.2. Material

The material of this experiment, same as the previous one, consisted of TED talks given by American speakers. Only those segments of the videos in which there was a difference between the baseline PSC and the enhanced PSC (i.e. segments including homophone, minimal pair, negatives and breached boundaries) were selected. However, to make the comparison fair, we ensured that the number of shown words in the target phrase were equal in the baseline PSC and the enhanced PSC, while the choices of the words were different. In this view, we circumvent a situation where learners prioritize a version over another because of the larger quantity of shown words.

5.3.3. Procedure

The experiment consisted of two parts:

In Part I, the participants were supposed to watch a series of videos without any caption (each lasted for 25~35 seconds) until paused. After each unexpected pause, the participants were asked to transcribe the last few words they had heard. It was assumed that through transcription, learners would realize which word(s) were more difficult for them to recognize. Therefore, immediately after the transcription, the learners received the baseline PSC and the enhanced PSC each including a segment of target words they had to transcribe. The participants then were asked to choose between two versions of PSC deeming for the one that included better words i.e. more of the words they misrecognized or had difficulty to recognize. Given that the number of shown words was equal in two versions of PSC, learners' selected caption would indicate its superiority in the choice of shown words as compared to the other version. It should be noted that learners were uninformed about which choice is the baseline PSC or the enhanced PSC.

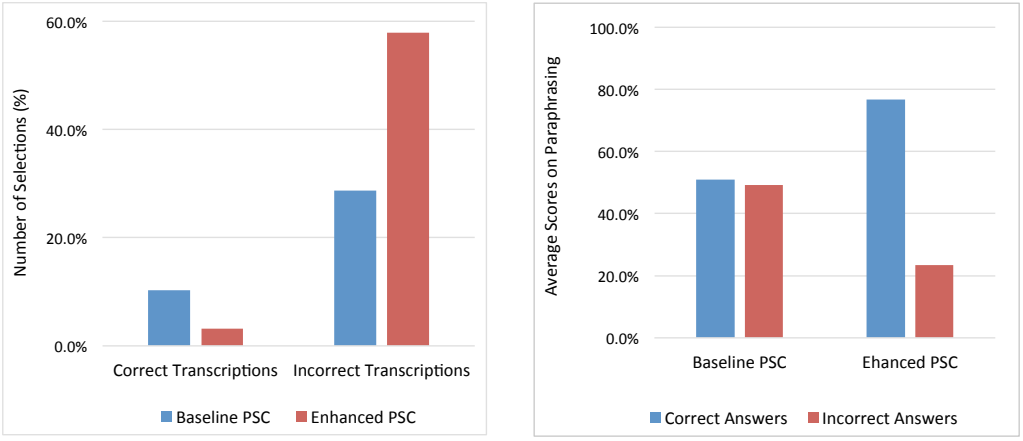
To evaluate the enhanced PSC over the baseline PSC with a more quantitative approach, we designed a paraphrasing test to complement our experiment. Accordingly, in Part II of our experiment, the learners were divided into two groups: (a) those who received the baseline PSC along with the videos and (b) those who received the enhanced PSC along with the video. In both groups, the learners were asked to watch a series of videos (each lasted 25~35 seconds) with the assigned caption (baseline PSC vs. enhanced PSC) until paused. Upon each pause, the learners were given two paraphrasing sentences on the last heard sentence. They had to select a paraphrasing choice that had the closest meaning to the last heard sentence. Since each group received a different PSC, comparison of their paraphrasing score could identify which PSC, baseline or enhanced, provided better clues to disambiguate and recognize the target phrase, hence select the best paraphrasing choice. The results of this part of the experiment provide us with quantitative data on evaluation of the baseline PSC and the enhanced PSC based on the learners' scores.

5.3.4. Results

Figure 7(a) shows the results of the experiment for Part I, in which the participants selected between the baseline and the enhanced PSC based on their preference. This was done immediately after the participants dealt with transcription and identified their recognition difficulties. It is shown that 61% of the times the participants opted for the enhanced PSC compared to the baseline PSC (39% of the times). It can also be seen that only a small number of transcriptions (13.4%) were correct. The correct transcription indicates that the learners did not require any caption to recognize the target sentence, thus we do not draw any conclusion on captions selected after correct transcriptions.

However, as the figure shows, the majority of the participants had difficulty in transcribing the ASR erroneous segments, which led to 86.6% of incorrect answers. In this case, a large majority of the participants could find the required clues in the enhanced PSC as opposed to the baseline PSC. This result indicates that significant improvement in the enhanced PSC makes it more preferable.

Figure 7(b) illustrates the paraphrasing scores of the baseline PSC group compared against the enhanced PSC group (Part II of the experiment). The results indicate that participants in the baseline PSC group answered the



(a) Part I: Participants' preferences on choosing between baseline PSC and enhanced PSC after transcription.

(b) Part II: Parapharasing scores of the participants in baseline PSC group vs. enhanced PSC group.

Figure 7. Experimental evaluation of Baseline PSC and Enhanced PSC

questions more or less by chance: 50.9% correct versus 49.1% incorrect answers. However, the performance of the learners in the group with the enhanced PSC is significantly better, gaining 76% correct answers as opposed to the 24% of incorrect responses. Findings of this experiment, which is based on quantitative data derived from the participants' scores, demonstrates that the enhanced version provides more appropriate assistance to the learners and is more successful in fostering L2 listening.

6. Discussions on the Performance of the ASR System

The performance of the ASR system used to conduct this study is critical to our proposed method. If the ASR accuracy is too high like over 90%, there would be few errors useful for this method. On the other hand, if the ASR accuracy is very low like below 80%, it would be difficult to conduct an alignment between the ASR result and the oracle transcript, thus effective breached boundaries cannot be extracted. Therefore, we assume the accuracy around 80% would be preferred. When the accuracy is much better, we may need to exploit N-best candidates of the ASR result.

The state-of-the-art ASR systems are built using a huge amount of training data like thousands of hours of speech, preferably matched with the test data. Some state-of-the-art ASR systems built for TED talks within the IWSLT evaluation campaign (Cettolo et al., 2015) reported the accuracy over 90%, but they are very complex involving a number of ASR processes. In this study, however, we adopted a *standard* setup to get a target accuracy of around 80% and to maintain an acceptable efficiency. On the other hand, it is reported that ASR performance of general YouTube videos are still low (50-60%) even with a number of techniques used to enhance the accuracy (Liao et al., 2013). This suggests that we should focus on prepared and fluent speech and we need to build a dedicated ASR system with a matched acoustic and language model, which requires data of a hundred-hour scale.

There are many factors that affect speech recognition. Therefore, we turned to ASR errors as well as statistics of individual factors. Speech rate, for instance, is one of the main factors that affect speech recognition, as we pointed out in this work. We previously investigated the effect of recognition frame rate on ASR performance and proposed several methods to adaptively change the frame rate (Okuda et al., 2002; Nanjo and Kawahara, 2004). Although there was some improvement with them, we kept a standard setting in this study to extract difficult segments effectively and to maintain the simplicity of the system. While the ASR performance of 80% is preferred for the purpose of this study, in some domains it might not be easy to build an ASR system with this target accuracy. However, even in that case, it would not be difficult to choose speech segments with this accuracy range from a pool of test data.

7. Conclusions

We have investigated the use of Partial and Synchronized Caption (PSC) for L2 listeners and proposed a new approach, exploiting the ASR cues, to detect difficult speech segments in order to improve the baseline PSC. The baseline PSC detects difficult words based on the speech rate of the words, their frequency, and specificity. Through calculating these features, PSC generates a caption, which presents the difficult words on the screen and hides the easy ones to promote more listening and less reading. With a considerably lower amount of shown words (less than 30%), it can provoke a similar level of comprehension as the full caption. However, given that difficult words are not bounded to PSC's defined criteria, the system yet anticipates for improvement to cover other difficult cases.

To address this issue we proposed the use of an ASR system as a model epitomizing L2 listeners, where ASR errors can be viewed as problematic speech segments for learners and ASR correct cases can be seen as easy to recognize segments. To attest our hypothesis on the usefulness of ASR errors in predicting difficulties of the audio, we had TED talks transcribed by an ASR system and analyzed the ASR errors to discover the underlying factors. Annotation of these errors distinguished four categories among many possible features that deemed to be useful for embedding them into the PSC: homophones, minimal pairs, negatives, and breached boundaries.

To confirm the usefulness of the features derived from ASR erroneous cases, an experiment was conducted with L2 listeners asking them to transcribe two segments on each video: one segment including homophones, minimal pairs, negatives, and breached boundaries, and another one including trivial words as a control case. The results showed that these four categories of ASR errors were problematic for L2 listeners, whereas learners hardly faced difficulties in transcribing easy (control) cases. Following the findings, PSC was enhanced by leveraging ASR errors and was compared against the baseline PSC in another experiment. The results of the latter experiment revealed that L2 listeners noticeably preferred the enhanced PSC to the baseline and gained better recognition and paraphrasing scores with the enhanced PSC.

This work opens a new avenue on the use of ASR errors to explore difficult speech segments for L2 listeners and hence provide them with useful means to overcome listening difficulties. However, as long as our statistics revealed, not all ASR errors are useful in this regard. While some of the ASR errors have unknown root-causes that cannot be determined easily, hence discarded, some can be ineffective because of the contextual clue. In this view, not all ASR errors are good predictors of learners' difficulty in listening, but some of them (e.g. breached boundaries) are indeed worth investigating.

These findings shed light on future advances of the PSC system by using ASR as a model of a language learner where through degrading the ASR, its errors can provide more useful instances for PSC on language learners with different proficiency levels. This process can be done by degrading the acoustic model or the language model by reducing the training data. Finally, another area that should be explored is learner adaptation, which is essential for the PSC system to encompass a wide range of learners with different requirements and interests.

Acknowledgments

We would like to express our gratitude to professors Masatake Dantsuji and Yuya Akita for sharing their knowledge and professional experience with us and for their generous support during the course of this research.

References

Amaral, L.A., Meurers, D., 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. ReCALL 23, 4–24.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., et al., 2007. Automatic speech recognition and speech variability: A review. Speech Communication 49, 763–786.

Bloomfield, A., Wayland, S.C., Rhoades, E., Blodgett, A., Linck, J., Ross, S., 2010. What makes listening difficult? Factors affecting second language listening comprehension. Technical Report. College Park, MD: University of Maryland Center for Advanced Study of Language.

Buck, G., 2001. Assessing listening. Cambridge University Press.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., Federico, M., 2015. The IWSLT 2015 evaluation campaign. Proc. of IWSLT, Da Nang, Vietnam .

Chen, W., Ananthakrishnan, S., Kumar, R., Prasad, R., Natarajan, P., 2013. Asr error detection in a conversational spoken language translation system, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE. pp. 7418–7422.

Coxhead, A., 2000. A new academic word list. TESOL quarterly 34, 213–238.
Cruttenden, A., 2014. Gimson's pronunciation of English. Routledge.
Cutler, A., 1990. Exploiting prosodic probabilities in speech segmentation, in: Cognitive models of speech processing: Psycholinguistic and computational perspectives, Cambridge MA: MIT Press. pp. 105–121.
Cutler, A., Butterfield, S., 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. Journal of Memory and Language 31, 218–236.
Danan, M., 2004. Captioning and subtitling: Undervalued language learning strategies. Meta: Journal des traducteursMeta:/Translators' Journal 49, 67–77.
Davies, M., 2008. The corpus of contemporary American English. BYE, Brigham Young University.
Felps, D., Geng, C., Gutierrez-Osuna, R., 2012. Foreign accent conversion through concatenative synthesis in the articulatory domain. IEEE Transactions on Audio, Speech, and Language Processing 20, 2301–2312.
Field, J., 2003. Promoting perception: Lexical segmentation in l2 listening. ELT journal 57, 325–334.
Field, J., 2008. Bricks or mortar: which parts of the input does a second language listener rely on? TESOL quarterly 42, 411–432.
Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in convertional speech. Speech Communication 29, 137–158.
Gardner, D., Davies, M., 2013. A new academic vocabulary list. Applied Linguistics 35, 305–327.
Gilmore, A., 2007. Authentic materials and authenticity in foreign language learning. Language teaching 40, 97–118.
Goldwater, S., Jurafsky, D., Manning, C.D., 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52, 181–200.
Griffiths, R., 1992. Speech rate and listening comprehension: Further evidence of the relationship. TESOL quarterly 26, 385–390.
Lee, A., Kawahara, T., 2009. Recent development of open-source speech recognition engine julius, in: Proceedings of (APSIPA ASC) Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee. pp. 131–137.
Levy, M., 1997. Computer-assisted language learning: Context and conceptualization. Oxford University Press.
Liang, F.M., 1983. Word Hy-phen-a-tion by Com-put-er. Citeseer.
Liao, H., McDermott, E., Senior, A., 2013. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription, in: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, IEEE. pp. 368–373.
Meyer, B., Wesker, T., Brand, Thomas sand Mertins, A., Kollmeier, B., 2006. A human-machine comparison in speech recognition based on a logatome corpus, in: Speech Recognition and Intrinsic Variation Workshop, pp. 95–100.
Meyer, B.T., Brand, T., Kollmeier, B., 2011. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. The Journal of the Acoustical Society of America 129, 388–403.
Mirzaei, M.S., Akita, Y., Kawahara, T., 2014. Partial and synchronized captioning: A new tool for second language listening development, in: CALL Design: Principles and Practice-Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands, Research-publishing. net. pp. 230–236.
Mirzaei, M.S., Kawahara, T., 2015. Asr technology to empower partial and synchronized caption for l2 listening development, in: Workshop on Speech & Language Technology for Education (SLaTE), Leipzig, Germany, pp. 65–70.
Moore, R.K., Cutler, A., 2001. Constraints on theories of human vs. machine recognition of speech, in: Workshop on Speech Recognition as Pattern Classification (SPRAAC), Max Planck Institute for Psycholinguistics. pp. 145–150.
Nanjo, H., Kawahara, T., 2004. Language model and speaking rate adaptation for spontaneous presentation speech recognition. IEEE Transactions on Speech and Audio Processing 12, 391–400.
Naptali, W., Kawahara, T., 2012. Automatic speech recognition for ted talks, 6th Spoken Document Processing Workshop, Toyohashi, Japan.
Nation, I., Beglar, D., 2007. A vocabulary size test. The language teacher 31, 9–13.
Nation, I.S., Webb, S.A., 2011. Researching and analyzing vocabulary. Heinle, Cengage Learning.
Nitta, H., Okazaki, H., Klinger, W., 2010. Missed word rates at increasing listening speeds of high-level japanese speakers of english. Studies In The Humanities: The Journal of the Senshu University Research Society 87, 171–198.
Okuda, K., Kawahara, T., Nakamura, S., 2002. Speaking rate compensation based on likelihood criterion in acoustic model training and decoding., in: INTERSPEECH.
Osada, N., 2004. Listening comprehension research: A brief review of the past thirty years. Dialogue 3, 53–66.
Pujolà, J.T., 2002. Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program. ReCALL 14, 235–262.
Révész, A., Brunfaut, T., 2013. Text characteristics of task input and difficulty in second language listening comprehension. Studies in Second Language Acquisition 35, 31–65.
Rost, M., 2005. L2 listening. Handbook of research in second language teaching and learning , 503–527.
Scharenborg, O., 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. Speech Communication 49, 336–347.
Scharenborg, O., ten Bosch, L., Boves, L., Norris, D., 2003. Bridging automatic speech recognition and psycholinguistics: Extending shortlist to an end-to-end model of human speech recognition. The Journal of the Acoustical Society of America 114, 3032–3035.
Schmitt, N., McCarthy, M., 1997. Vocabulary: Description, acquisition and pedagogy. volume 2035. Cambridge university press Cambridge.
Schmitt, N., Schmitt, D., 2014. A reassessment of frequency and vocabulary size in l2 vocabulary teaching. Language Teaching 47, 484–503.
Shen, W., Olive, J., Jones, D., 2008. Two protocols comparing human and machine phonetic discrimination performance in conversational speech, in: Proceedings of Interspeech, pp. 1630–1633.
Shinozaki, T., Furui, S., 2001. Error analysis using decision trees in spontaneous presentation speech recognition, in: Automatic Speech Recognition and Understanding, 2001. ASRU01. IEEE Workshop on, IEEE. pp. 198–201.
Siegler, M.A., 1995. Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition. Ph.D. thesis. Carnegie Mellon University Pittsburgh.

Sydorenko, T., 2010. Modality of input and vocabulary acquisition. Language Learning & Technology 14, 50–73.

Tauroza, S., Allison, D., 1990. Speech rates in british english. Applied linguistics 11, 90–105.

Vandergrift, L., 2004. 1. listening to learn or learning to listen? Annual Review of Applied Linguistics 24, 3–25.

Vasilescu, I., Adda-Decker, M., Lamel, L., 2012. Cross-lingual studies of asr errors: paradigms for perceptual evaluations., in: LREC, pp. 3511–3518.

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

Wang, D., Narayanan, S., 2005. An unsupervised quantitative measure for word prominence in spontaneous speech, in: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., IEEE. pp. 377–380.

Weber, A., Cutler, A., 2004. Lexical competition in non-native spoken-word recognition. Journal of Memory and Language 50, 1–25.