

**Partial and Synchronized Caption
to Foster Second Language Listening
based on
Automatic Speech Recognition Clues**



Maryam Sadat Mirzaei

**GRADUATE SCHOOL OF INFORMATICS
KYOTO UNIVERSITY**

© Copyright by Maryam Sadat Mirzaei, 2017.
All rights reserved.

Abstract

Central to the development of second language (L2) is the ability to perceive, process and comprehend the speech in the target language, which forms the bedrock of L2 listening skill. Listening is indeed a fundamental skill in L2 acquisition, which comes before speaking. It is the least explicit and essentially a transient and invisible process, hence the most sophisticated skill to master. To advance L2 listening skill, exposure to the authentic materials plays a crucial role. The advancement of ICT has promoted further opportunities for the application of contextualized and authentic materials, making them the mainstays of contemporary L2 learning education. Nevertheless, these materials, which are originally intended for native speakers of the target language, are often too difficult for L2 learners even at advanced levels. To facilitate the comprehension of such resources, captioning is widely used as an assistive tool for providing the text along with the speech. However, through the use of captions, learners tend to rely more on their reading skills, hence neglect the goal of training the listening skill.

This thesis attempts to solve this problem by introducing a novel captioning method, partial and synchronized captioning (PSC), as a tool for developing L2 listening skill. In this method, an ASR system is employed to align the words in precise timing with their respective speech signals in order to enable text-to-speech mapping and the caption is partialized by presenting words and phrases which are likely to hinder learner's listening comprehension. Since there are various factors that lead to L2 listening difficulty, this study investigates the viability of using ASR errors as the predictor of difficulties in speech segments, thereby exploiting them to improve the baseline PSC system. Note that the human-annotated transcript is aligned with the ASR-generated transcript to realize synchronization and ASR error detection.

Chapter 1 provides an overview of the general topics of second language listening and the use of captioning to facilitate this process. It goes on to discuss the problems with the existing approaches and lays out the general motivation for and techniques used in the work presented in the following chapters.

Chapter 2 describes the overview of computer-assisted language learning (CALL) systems, as well as the use of ASR technology. It also focuses on the use of technologies in training L2 listening skill and discusses the different factors affecting L2 listening. This chapter provides a thorough introduction of different captioning methods and elaborates on the limitations of each method. Within this framework, this thesis presents the importance of adopting partialization and synchronization for creating a new system of captioning.

Chapter 3 presents the idea of partial and synchronized caption as a tool that strives to mandate the shortcoming of previous methods. A trained ASR system allows for precise mapping between the text and the speech. With regards to partialization, the system relies on factors impeding the L2 listening process. In the baseline PSC system, partialization is performed based on well-known factors of speech rate, word frequency and specificity. This makes it straightforward to select difficult words and allows for caption adjustment through taking into account learners' vocabulary size and their tolerable rate of speech. Experiments demonstrate that the proposed method is able to realize comparable comprehension as the full caption while reducing the textual clues to less than 30%. The method is also able to address the requirement of L2 learners at different proficiency levels and can prepare them for listening in real-life situations.

Chapter 4 presents a comparative analysis of ASR errors and L2 learners' problems so that ASR errors can epitomize learners' listening difficulties with a particular audio. Given an erroneous output of the ASR system, we look for useful instances that can signal challenging speech segments for L2 listeners. The chapter reports on the analysis of ASR errors and the baseline PSC shown and hidden cases. This analysis provides hints for detection and inclusion of effective ASR errors into the PSC system, which is essential for achieving high accuracy in word selection to scaffold the learners. Annotation of ASR erroneous output led to the discovery of four effective categories of errors: homophones, minimal pairs, negatives, and breached boundaries to improve the choice of words in PSC. Experiments with L2 learners show that these categories are able to detect problematic speech segments and can be useful for enhancing the PSC system.

Chapter 5 extends the baseline PSC framework to encompass the features derived from the ASR errors and to enhance the word selection. In this view, two enhancements are made to improve the baseline system. The first improvement is removing easy cases by defining a secondary threshold for speech rate and word specificity features, taking into account the ASR correct or erroneous output, which allows for more effective pruning of the words coming to PSC. The second improvement is based on aggregating the four useful categories discovered in ASR errors, which lead to providing better choices of words to disambiguate the speech while listening. An experimental evaluation finds that the enhanced version of PSC is able to provide better clues for language learners while addressing most of their problems and being selected as a preferable caption.

Chapter 6 concludes the thesis with an overview of the baseline and the enhanced system of PSC and directions for future research.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my advisor, Professor Tatsuya Kawahara who has trusted me the first time I walked into his office and expressed my interest in studying under his supervision. He encouraged me and welcomed me into his lab, provided me with tremendous academic support, and allowed me to pursue a research project of my interest. Professor Kawahara was my primary resource for getting all my questions answered. He has always made himself available to guide me despite his busy schedule and generously shared his exceptional research expertise to walk me through the procedure of research and publication and to show me what it takes to be an excellent researcher. His dedicated support, inspiration, encouragement, high-quality work and integral view of research have made a deep impression on me. I owe him lots of gratitude and I feel privileged to be his student.

I would also like to extend my special regards to members of my Ph.D. committee Professor Sadao Kurohashi and Professor Masatake Dantsuji for taking their time to review this thesis and to share their remarkable expertise for providing excellent advice and insightful comments. Their valuable suggestions have greatly improved the thesis.

I owe my deepest gratitude to my wonderful collaborator and true friend, Dr. Kourosh Meshgi, who has greatly contributed to this work and genuinely supported me not only by his immense and active cooperation at every stage of this research but also by his precious moral support through the rough road to finish this thesis. This work would not have been possible without his help and support. His constant guidance, cooperation, and motivation have always kept me going ahead. I always counted on him whenever I needed help and I will forever be indebted to him.

I would like to express my deepest thankfulness and heartfelt gratitude to Professor Mark Peterson, my former advisor and forever mentor, for being both a remarkable teacher and a splendid friend. Not many students have a chance to call their professor their best and closest friend and I am the most honored to have this chance. He has guided me with his outstanding research expertise through all these years and supported me in every possible way. His family members shared a great relationship as compassionate friends and I am always indebted to their warmth and kindness.

His tremendous support and unconditional help have deeply impressed me and I feel honored to say that I have learned my most important life lessons from him.

I would like to especially thank Professor Toyooki Nishida, for his continuous support, motivation and kind interest in my research. In his classes, I have learned extensively from him not only academic lessons but also moral lessons. Professor Nishida is and will be my role model for his excellent academic achievements and his exemplary cordial relationships with students.

My special regards to former and current members of Kawahara lab, particularly Professor Yuya Akita, whose assistance and support played a crucial role in the completion of this work. Professor Shinsuke Mori for his insightful comments and support. Professors Kazuyoshi Yoshii, Katsutoshi Itoyama and Koichiro Yoshino, Dr. Li Sheng and Mr. Koji Inoue for their timely help and friendship and all other members who made my experience in Japan wonderful.

I gratefully acknowledge the Ministry of Education, Culture, Sports, Science and Technology of Japan for granting a full scholarship that made my Ph.D. work possible. I would like to extend my thankfulness to institutions that supported this research, Kyoto University and ECC corp. whose collaboration has been a source of pride.

A special mention of thanks to my friends Sara Owj, Somayeh Nilouyal, Behnam Ghalei, Alireza Jalili and many more for their eternal kindness, support, and motivation, which encouraged me to accomplish this work.

Last but not least, I would like to dedicate this thesis to the most wonderful parents in the world, Mohammad Mirzaei and Sorour Abazari for being my source of inspiration, forming my vision, teaching me to make impossible possible, growing me with distinguished values, filling me with joy and happiness, cherishing me with love and supporting me through the borders in a way that nobody else could ever do.

My wholehearted gratitude goes to my brothers Saeed and Masoud who always stood on my side to give me strength and to assure me that I always have their support and to my dearest sister Sima for her infallible love, care, and affection that I can never reciprocate. I feel a deep sense of gratitude for my brother and sister-in-laws Farnaz, Marjan, Hamidreza and the angels of family Sourosh, Sepeher, Toranj, Nelie, and Yana.

I would like to dedicate my thesis to the best parents in the world

Mohammad Mirzaei and Sorour Abazari

*To my Dad, to whom I always looked upon and of whom I am always proud
For all those days he stood beside me and made me feel strong
For all those years he raised me with love and gave me more than I could ask*

*To my Mom whose love is endless and who's the role-model of my life
For all those days she embraced me even when I did wrong
For all those years she dedicated her life to show me what an angel is like*

Dad, Mom I just can't thank you enough for what you have done

Contents

Abstract	iii
Acknowledgements	v
List of Tables	xii
List of Figures	xiii
1 Introduction	1
1.1 Second Language Listening Skill	1
1.2 Captioning as Assistive Tool for L2 Learners	2
1.3 Problems with Existing Captioning Methods	3
1.3.1 Advocating Hindering Listening Strategies	3
1.3.2 Reading instead of Listening	4
1.3.3 Overlooking Different Learners' Needs	5
1.4 Partial and Synchronized Captioning for L2 Listening	5
1.5 Challenges and Approaches	7
1.5.1 Defining and Selecting Difficult Words	7
1.5.2 Addressing Different Learners' Needs	8
1.5.3 Discovering Listening Difficulties in Speech in Different Videos	8
1.6 Organization of the Thesis	10
2 Literature Review	13
2.1 Computer-Assisted Language Learning (CALL)	13
2.2 ASR Systems in Second Language Learning	15
2.2.1 ASR Framework	16
2.2.2 Factors Causing ASR Errors	19
2.2.3 ASR Applications in Second Language Learning	22
2.3 Factors affecting L2 Listening Skill	24
2.3.1 Listening Strategies	25

2.3.2	Listening Materials	25
2.3.3	Lexical Factors	26
2.3.4	Acoustic, Speech and Perceptual Factors	28
2.4	Captioning Methods	31
2.4.1	Full Captions	32
2.4.2	Synchronized Caption	33
2.4.3	Keyword Caption	35
2.4.4	Limitations of Captioning Methods	36
3	Baseline Partial and Synchronized Caption	39
3.1	Concept of Partial and Synchronized Caption	40
3.2	Feature Extraction	43
3.2.1	Speech Rate	44
3.2.2	Word Frequency	45
3.2.3	Word Specificity	46
3.3	System Implementation	46
3.3.1	System Input	47
3.3.2	Synchronization	48
3.3.3	Partialization	49
3.3.4	Learner Adaptation	51
3.3.5	Caption Generation	52
3.4	Experimental Evaluation of Baseline PSC	53
3.4.1	Participants	53
3.4.2	Material	54
3.4.3	Data Collection Instruments	55
3.4.4	Procedure	56
3.4.5	Results	58
3.5	Discussions	62
3.5.1	Overall Effect of Different Captioning Methods	62
3.5.2	Effectiveness of PSC Compared to FC	62
3.5.3	Effectiveness of PSC across Proficiency Levels	63
3.5.4	Effectiveness of PSC to Prepare Learners for Real-life Situations	64
3.6	Conclusion	65

4	ASR Errors to Predict L2 Listening Difficulties	67
4.1	ASR versus Human Speech Recognition	67
4.2	Reviews on ASR versus L2 Speech Recognition	69
4.3	ASR Error Analysis	73
4.3.1	ASR Error Statistics	73
4.3.2	ASR Error Trends	74
4.4	Comparison of ASR Output and PSC Selection	76
4.4.1	Analysis on ASR Error and PSC Shown Cases	77
4.4.2	Analysis on ASR Correct and PSC Shown Cases	78
4.4.3	Analysis on ASR Error and PSC Hidden Cases	80
4.5	Experimental Evaluation of Additional Features	84
4.5.1	Participants	84
4.5.2	Material	84
4.5.3	Procedure	85
4.5.4	Results	86
4.6	Conclusion	87
5	Enhanced Partial and Synchronized Caption	89
5.1	Using ASR Clues to Enhance Baseline PSC	90
5.1.1	Improving Baseline PSC with ASR Correct Cases	91
5.1.2	Augmenting Baseline PSC with ASR Erroneous Cases	92
5.2	Feature Extraction from ASR Errors	95
5.3	Enhanced PSC System Realization	97
5.3.1	Extended System Overview	97
5.3.2	Statistics of Baseline PSC versus Enhanced PSC	98
5.4	Experimental Evaluation of Enhanced PSC	98
5.4.1	Participants	99
5.4.2	Material	100
5.4.3	Procedure	100
5.4.4	Results	101
5.5	Conclusion	103
6	Conclusions	105
6.1	Contributions and Summary	105

6.2	Future Work	108
6.2.1	Data-driven PSC	108
6.2.2	ASR as a Model of Language Learner	108
6.2.3	Learner Adaptation in PSC	109
6.2.4	PSC-Integrated CALL System Development	109
Appendix I List of TED Talks		111
Appendix II Comprehension Questions Sample		113
Appendix III Questionnaire on Baseline PSC		115
Appendix IV Publications by the Author		117
Bibliography		119

List of Tables

3.1	Comparison of different captioning methods	42
3.2	Standard Rates of Speech	44
3.3	Descriptive statistics of comprehension scores – Part I	59
3.4	ANOVA analysis on the effect of different captions – Part I	59
3.5	Posthoc comparisons on scores of different conditions - Part I	60
3.6	T-test analysis on scores of PSC vs. FC across proficiencies - Part I	60
3.7	Descriptive statistics of comprehension scores – Part II	61
3.8	Posthoc comparisons on scores of different conditions – Part II	61
4.1	ASR-L2SR Comparison	72
4.2	ASR Error Analysis on TED Talks	74
4.3	ASR Cases versus Baseline PSC Comparison	77
4.4	Patterns and their usefulness in ASR Error & PSC Hide	83
5.1	Baseline PSC vs. Enhanced PSC	98
III.1	5-point Likert-Scale Survey Results	115

List of Figures

1.1	Screenshot of the PSC System	6
1.2	Overview of the thesis	10
2.1	Typical Framework of ASR.	17
2.2	Screenshot of Synchronized Caption vs. Full Caption	34
3.1	Screenshot of the Full Caption vs. PSC	41
3.2	Schematic of Baseline PSC System	47
3.3	Statistics on the speech rate and word frequency of the video clips	54
3.4	Percentage of words shown in PSC for pre-intermediate group	55
3.5	Experimental procedure	57
3.6	Experimental design for Part I	58
4.1	Trend Analysis on ASR Errors	75
4.2	Feature analysis in ASR Error and PSC shown cases	78
4.3	Feature analysis in ASR correct and PSC shown cases	79
4.4	Experimental procedure of a transcription task	86
4.5	Transcription scores on segments of ASR errors vs. ASR correct	87
5.1	Enhanced PSC Process Flow	97
5.2	Screenshot of baseline and enhanced PSC	99
5.3	Evaluation of Baseline PSC and Enhanced PSC – Part I	102
5.4	Evaluation of Baseline PSC and Enhanced PSC – Part II	103

Chapter 1

Introduction

1.1 Second Language Listening Skill

Learning a new language is a challenge that involves mastering different skills of listening, speaking, reading and writing. Of these, the listening skill is viewed as the very first skill required not only for learning second language (L2), but also for acquiring first language (L1) (Krashen, 1981; Lightbown & Spada, 2006).

Children learn to listen in L1 by spending a great deal of time listening to others while being abundantly exposed to L1 input in the immediate environment (Rost, 2013). However, learning to listen in an L2 is different in many ways especially because it concerns access to useful sources of input. Moreover, listening skill, whether it is listening in the first or the second language, involves an attentive combination of phonological, morphemic, syntactic and semantic rules of the language and the ability, in practice, to apply such knowledge rapidly and automatically (Buck, 1988).

While listening, we have a difficult task to utilize all the necessary knowledge and skill simultaneously. In this view, L2 listening is distinct from the other three skills of speaking, reading and writing where the learner controls both the speed and the content (Leveridge & Yang, 2013). As a listener, however, our role is rather passive and it is very likely that the speaker proceeds before we can sort out what we have heard (Osada, 2004).

Given the importance of developing listening skill in mastering a foreign language, investigating effective tools and pedagogical methods to promote this skill is a necessary consideration. However, direct instruction of L2 listening skills had been neglected for a long period of time (Oxford, 1993). While speaking, reading and writing were at the heart of L2 instructions, instructors frequently considered listening as a receptive skill and expected the language learners to advance this skill by osmosis and with no assistance (Mendelsohn, 1984). Nevertheless, listening can no more be overlooked as learners and instructors grew to understand the unique characteristic of this skill and its direct benefit to effective communication.

Central to the development of L2 listening skills is the demand of exposure to authentic and comprehensible input (Danan, 2004; Vanderplank, 2010). Such materials represent the natural speech, and help the learners become familiar with real cadences of the target language (Field, 1998). With the advances of technologies, authentic audio and visual materials such as broadcast news, movies, and academic lectures have become easily accessible and increasingly embedded into language learning classrooms (Vanderplank, 2010; Vandergrift, 2011). While these resources provide rich content and reflect real-life language, they often entail complex listening comprehension skills (Gilmore, 2007). To assist L2 listeners in comprehending these materials, facilitative tools such as visual or textual clues (Danan, 2004) and speed controllers (Zhao, 1997) have emerged to enhance the pedagogical effect of this medium (Vandergrift, 2011).

1.2 Captioning as Assistive Tool for L2 Learners

One of the assistive tools used to facilitate the comprehension of authentic materials is captioning. Captioning was first used to help hearing impaired people when watching TV, but later became popular as a medium for L2 instruction (Garza, 1991). Following this, many researchers investigated whether captioning could improve language processing ability of L2 learners, who are “hard of listening” as well as hearing impaired people, “hard of hearing” (Vanderplank, 1988, p. 272). Captioning tex-

tualizes the verbatim speech and makes it more recognizable by demonstrating the word boundaries neatly, without being affected by accent, pronunciation and audio deficiencies (Vanderplank, 1988). It allows the learners to parse the speech stream into meaningful chunks, which is an essential process for learning (Ellis, 2003).

Many studies have highlighted the potential of captioning not only to facilitate comprehension but also to promote vocabulary acquisition (Winke et al., 2010; Montero Perez et al., 2013). The conventional captioning method, hereinafter full captioning, has long been used as a means to facilitate L2 listening and promote text-to-speech mapping (Garza, 1991; Danan, 2004; Winke et al., 2010). However, there are some critics with the use of full captions especially when the purpose is to train L2 listening skill.

1.3 Problems with Existing Captioning Methods

Problems regarding the use of full captions can be conceptualized around several key factors: encouraging a word-by-word decoding strategy and promoting the overuse of bottom-up skill (Osada, 2004), allowing comprehension of audio by just reading the text without listening (Pujolà, 2002), and imposing a high level of cognitive load by providing a large amount of textual clues together with the audio (Sydorenko, 2010). These problems which concern the use of existing captioning methods for promoting L2 listening skill raises some questions: how can we prevent learners from adopting hindering listening strategies when they use captions? How are we able to encourage the learners to rely on their listening skill and decrease their dependence on reading the text when the full caption is given? And how can we reduce the cognitive load and textual density in the caption (like in keyword captioning), but be sure that the selected words are actually proper for different levels of learners?

1.3.1 Advocating Hindering Listening Strategies

Many learners adopt the wrong strategies when using the caption along with the video; they often go through word-by-word decoding and over-emphasize on the use

of bottom-up strategies (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).

For effective listening, learners should make a balance between the use of top-down and bottom-up strategies (Rost, 2005). Instead of continuously decoding each word in the caption, learners should be able to use their background knowledge to disambiguate what they have heard. This, in turn, helps the listeners to build on larger meaningful segments rather than depending on individual words (Mayor, 2009).

On one hand, the learners require a sort of assistance to overcome the difficulties of listening to authentic materials, on the other hand, their misuse of captions impedes L2 listening development. Consequently, it is crucial to find a better tool that can mandate the limitation of conventional captioning and avoid the use of word-by-word decoding strategy, while effectively scaffolding the learners when necessary.

1.3.2 Reading instead of Listening

The conventional full-captioning method has also received critical attention for bringing too much textual assistance and promoting reading over listening (King, 2002; Pujolà, 2002; Vandergrift, 2004). This assistive tool facilitates listening comprehension by providing the text along with the audio, but it also promotes a significant amount of reading which raises the question if listeners' comprehension is gained by merely reading the text (Pujolà, 2002) rather than listening to the audio.

Hence, when it comes to using captions for training listening skill, both language learners and teachers face a dilemma. In fact, when reading captions is part of watching a video, learners often rely on their reading skill to compensate for their listening skill deficiencies, whereas in a real-world communication, learners should solely use their listening skill as no assistive tools are available. As a result, while the usefulness of caption is confirmed, it is important to consider the necessity to decrease the dependence on the caption at least through the course of training.

1.3.3 Overlooking Different Learners' Needs

In most cases, when authentic videos are accompanied with full captions, learners' differences and limitations of cognitive capacity may impose difficulty in attending to the three sorts of input i.e., audio, video, and text (Taylor, 2005; Sydorenko, 2010). As such, some learners may use their attention selectively and prioritize reading over listening (Lund, 1991; Pujolà, 2002), since it may be hard to divide their attention equally over both skills (Chang, 2009).

In the light of such limitations of full captioning, keyword captioning emerged as an alternative solution (Guillory, 1998; Montero Perez et al., 2014b). In this method, only keywords, which are manually selected by some experts, are presented in the caption. While keywords may reflect important points thereby foster listening comprehension, the use of keywords does not necessarily lead to listening skill development. This is partly due to the limitation of such captioning method, in which the caption is not tailored to fit the learners' requirements. More precisely, the manual selection of keywords is not only costly and time-consuming but also subjective, which may not be beneficial when having a wide range of learners at different levels and with different needs. In this method, the keywords appear on the screen abruptly, which distracts the learners and makes it inappropriate for those who easily lose their focus (Montero Perez et al., 2014b). While reducing the textual density in the caption is beneficial in many ways, assuring that the remaining textual clues are sufficient for different learners and can rectify their listening problems is of paramount importance.

1.4 Partial and Synchronized Captioning for L2 Listening

In order to overcome the shortcomings of conventional captioning, this work proposes a novel captioning system called Partial and Synchronized Caption (PSC), which unlike conventional captions, automatically detects difficult words and presents them on the screen to foster listening, but hides easy words to encourage more listening than reading. Figure 1.1 shows a screenshot of the PSC caption.

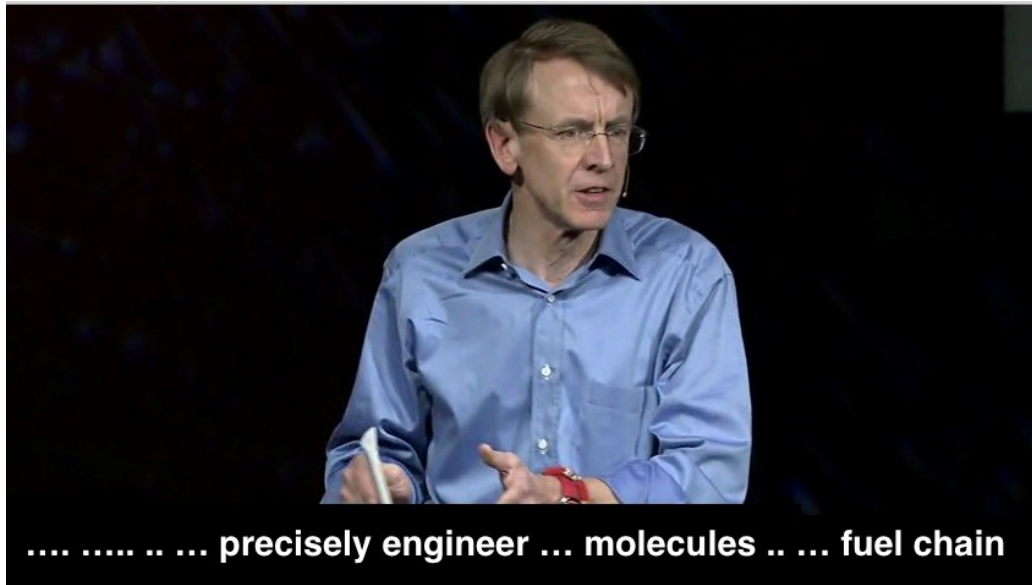


Figure 1.1: Screenshot of the PSC System: The caption text is presented incrementally in synch with the speech. The original transcript was: “That means we can precisely engineer the molecules in the fuel chain.” ©TED talk by John Doerr: Salvation (and profit) in green tech.

In this method, L2 learners should rely on their listening skill while being assisted by PSC to overcome the breakdowns in the listening process i.e., upon encountering a difficult word, learners can find it in PSC as a textual clue.

This method has two features: “partialization” and “word-level synchronization”. The former is to automatically reduce the amount of text and make a principled selection of words to appear on the screen, while the latter is to force each word to appear on the screen based on its timestamp (i.e., one-by-one, from left to right and in synchronous with the speaker’s utterance). Partialization is done in order to promote listening to the audio and referring to the caption only when encountering a problem. Meanwhile, synchronization is done to emulate the speech flow, allow for text-to-speech mapping (Bailly & Barbour, 2011) and avoid the irregular appearance of the words (Montero Perez et al., 2014b). In this view, PSC tries to mitigate the first and the second problems, which regard the overuse of bottom-up strategies and the over-reliance on reading the caption.

As a noteworthy feature of PSC, partialization and synchronization processes are done automatically so that the method can be applied to a large number of audio

and video contents. Besides, the selection of words in PSC is based on the learners' proficiency levels and the caption is tailored to meet the different learners' needs. This feature alleviates the third problem on addressing individual learner's requirements when selective words are shown in the captions.

1.5 Challenges and Approaches

1.5.1 Defining and Selecting Difficult Words

PSC strives to effectively remedy the shortcomings of the conventional full and keyword captioning methods. This tool is meant to facilitate the comprehension process while encouraging listening to the audio more than reading the caption. In this method keywords are not the selection criteria. Instead, the selection aims to include the difficult words or phrases in the caption. This idea raises the fundamental question of how can we define difficult words or phrases in the context of L2 listening? Finding a proper answer to this question is important to the creation of an effective tool that aims to assist L2 listeners.

To answer this question, we first referred to the body of research on L2 listening difficulties. These studies indicate that learners encounter a miscellaneous collection of factors that interfere with their listening process (Bloomfield et al., 2010). Among those, some features such as speech rate, word frequency, and word specificity are of prime concern for imposing more difficulties, hence being the main causes of L2 listening challenges (Griffiths, 1992; Révész & Brunfaut, 2013).

Based on the above argument, the detection of difficult words in the baseline system is inspired by the L2 studies and is based on speech rate, word frequency, and word specificity factors. These features were chosen as representative of major contributing factors to listening difficulties (Griffiths, 1992; Nissan et al., 1995; Schmitt & McCarthy, 1997; Révész & Brunfaut, 2013). Moreover, some of the other factors responsible for listening difficulty such as speaker accent, noise and length of the material could be easily circumvented and were not applicable to the content

of our study, hence not considered. Instead, the selected factors were feasible to be implemented and quantified automatically by the existing technologies.

1.5.2 Addressing Different Learners' Needs

For PSC to be effective, it is necessary to align the generated caption to the respective level of the learners. This is, in fact, one of the biggest shortcomings of keyword captioning, in which the selection of keywords is based on the content and does not consider various learners' requirements. In this view, one significant challenge to handle in generating effective PSC is to meet diverse learners' needs. While beginners may find more difficulty in processing the listening material due to limited vocabulary sizes, low tolerance to fast speech rate, etc., advanced learners may encounter less difficulty in comprehending the speech. This example highlights the importance of learner adjustment in PSC to differentiate the amount of shown words for different learners in effective ways. In doing so, PSC should be able to support learner adaptation feature i.e., to automatically adjust the level of the caption to suit the individual needs of different learners.

To address this challenge, this study investigates the use of evaluation metrics to assess the learners' proficiency levels and adjust the feature parameters of the PSC system, thereby provides the individual learners with words/phrases that lead to listening difficulties for them. In this regard, the vocabulary size of the learners and their tolerable rates of speech are determined based on the results of corresponding assessment tests and used to adjust the feature parameters in PSC. Finally, the level of difficulty and the amount of shown words in PSC is tailored to the requirement of different learners at different levels. While fewer numbers of words are shown to the advanced learners as compared to the beginners, the number of shown words is always kept to a minimum to ensure minimal but sufficient assistance to various learners.

1.5.3 Discovering Listening Difficulties in Speech in Different Videos

While PSC's baseline features can address the fundamental L2 listening problems, there are still other factors such as those related to the perceptual difficulty that are

not covered by the baseline version. As a result, baseline PSC system sometimes includes fairly easy to recognize words and occasionally exclude supposedly difficult words or phrases, which highlights the importance of exploring other possible features to enhance the system. One main source of difficulty for many L2 listeners is the perceptual difficulty of the speech, which includes instances such as phonological neighbors, identical pronunciations and word boundary locations (Cutler, 2005; Field, 2008; Broersma, 2012). For instance, for many language learners, finding the right boundaries between the words in the connected speech is very demanding, thus many L2 learners end up being confused with breached boundaries. Such difficulties severely hinder listening but are not addressed in the baseline PSC's selected words because they are not easy to detect without analyzing the nature of the speech. In fact, these factors are not only associated with the learners' perceptual difficulties but also linked to the speaker's clarity of utterances; hence these intrinsic speech difficulties need to be discovered using an elucidative source that analyzes the speech.

To address this challenge, this study proposes the use of ASR errors as a source to predict difficulties for L2 listeners. ASR systems process the speech signal to generate a transcript of the audio file. This process 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 addressing the perceptual difficulties and enhancing the baseline PSC system.

In this thesis, we focus on finding useful patterns or features in the ASR errors to detect problematic speech segments for L2 listeners. The discovered patterns are evaluated 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 realize an enhanced version, which aims to provide better assistance for the second language learners.

1.6 Organization of the Thesis

The thesis describes how the use of ASR technology and computational linguistics together with other components frame this system and enables the automatic generation of PSC. The thesis also reports on the evaluation of this new method compared with the full captioning method and investigates the room for its improvements. In the next step, the thesis aims to alleviate the shortcoming of the baseline PSC system. To this end, it approaches the problem of defining the listening difficulty for individual sentences/words by performing an analysis of the underlying features causing ASR errors and those that make L2 listening difficult for language learners. The thesis then provides the result of such analysis, presents the supportive experimental evaluation on the effective ASR errors and explains how these ASR clues can contribute to word selection in PSC. The enhancement of PSC is explained and evaluated by experiments and future directions of this study are introduced. Figure 1.2 demonstrates the organization of the thesis based on each chapter.

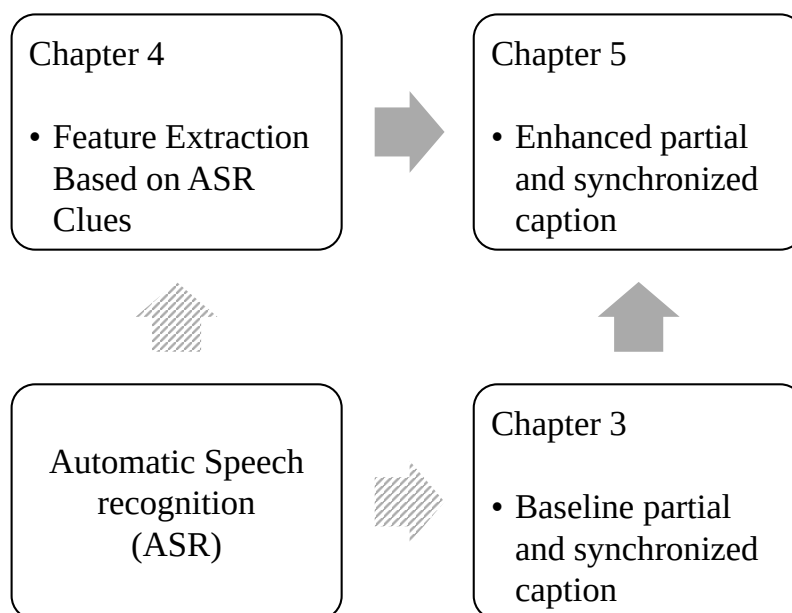


Figure 1.2: Overview of the thesis

Following this introduction, the next chapter reviews the CALL systems and the use of ASR systems in L2 teaching and learning. The chapter also introduces the

factors acting as barriers for L2 listening comprehension and presents an overview of captioning and different existing captioning methods. Finally, it reports, in detail, the limitations of the conventional captioning methods and explains the importance of investigating new methods of captioning for training the L2 listening skill.

Chapter 3 introduces the first main contribution of this thesis on developing PSC as a new technique of captioning and describes how the use of ASR technology and English language corpora together with other components makes this method an effective tool. It also explains the details on the selection criteria for PSC, which addresses the first challenge of the thesis regarding the detection of difficult words for the PSC system. Following this, the experimental procedure conducted to evaluate our method is presented. Finally, the results of the experiments are analyzed and discussed and the conclusion is provided.

Chapter 4 focuses on the second contribution of the thesis and investigates the similarities between the ASR errors and L2 listeners' problems in recognizing the speech, presuming that some of the ASR errors signify problematic speech segments for L2 learners. This approach addresses the second challenge of the thesis regarding the discovery of factors that lead to L2 listening difficulties in speech. This is done for the purpose of improving the choice of words in PSC by detecting other challenging speech segments. To conduct the analysis, those ASR errors that were not detected by PSC, as cases of learners' difficulties, were further analyzed. The root causes of some of these errors, which closely indicated the challenging nature of the audio led to the discovery of four categories including homophones, minimal pairs, negatives, and breached boundaries. The chapter reports the results of an experiment in which ASR-generated transcripts were compared with L2 learners transcription to check if these errors indicate problematic speech segments. This chapter is concluded by showing how these features lead to L2 difficulties and result in ASR errors.

Chapter 5 regards the third contribution of the thesis and explains how the use of ASR clues contributes to the enhancement of the baseline PSC system. It divides the enhancement to two parts: learning from ASR correct cases to remove easy instances from PSC and benefiting from ASR erroneous cases to address more difficult

words in PSC. The chapter describes how the discovered categories of errors are embedded into the baseline PSC to make the enhanced version and how this version is evaluated in an experiment with L2 learners. The chapter concludes, based on the experimental results. Findings indicate that the enhanced PSC addresses most of the L2 learners' difficulties and better assists them in comprehending challenging video segments compared to the baseline.

Finally, Chapter 6 discusses the overall findings, concludes this thesis, and points out the future directions for this research in order to exploit the full potential of the application of partial and synchronized captioning in language learning environments.

Chapter 2

Literature Review

This chapter presents a review of the computer-assisted language learning (CALL) systems and elaborates on the use of automatic speech recognition (ASR) systems in the domain of second language learning and teaching. In addition, the focus will be put on L2 listening skill and the different factors that affect L2 listeners' recognition and comprehension. As the focus of this thesis is on the use of captioning in developing L2 listening skill, this chapter also provides an introduction to the various existing methods of captioning and explains the limitation of each method.

2.1 Computer-Assisted Language Learning (CALL)

The advancement of 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 and learning environment, the interaction between teacher and learner, and the students' learning process ([Chapelle, 2001](#)). CALL systems provide the materials that meet the requirement 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, as well as various forms of computer-mediated communication with native speakers ([Amaral & Meurers, 2011](#)).

CALL has been introduced in the late 1990s. Then, in the following years, it has been enriched by the support of natural language processing (NLP) technologies, which led to the emergence of Intelligent CALL (ICALL). ICALL, as an extension of CALL, explores the application of artificial intelligence (AI) techniques for language learning to further enhance CALL systems ([Gamper & Knapp, 2002](#)). ICALL started its own research field more than a decade ago, when NLP technologies were advanced enough to be included in language learning systems, at least in experimental settings.

Linguistic analysis and NLP technology are introduced to enhance learner's awareness of language forms by providing individual feedback on learner's errors. The commencement of this new field goes back to the Intelligent Tutoring Systems (ITS), which made use of some NLP features to extend the traditional language learning systems ([Anderson et al., 1985](#)). This research field further explores the integration of ITS and NLP to create an instructional framework for foreign language learning by using some techniques such as grammar checking, error analysis and tutoring ([Swartz & Yazdani, 2012](#)). Many NLP technologies have been employed to illustrate linguistic structures, make language comprehensible, provide varied exercise material, and spot and correct errors ([Aldabe et al., 2006](#)). As instances of these technologies, we can refer to concordance programs that are a sort of NLP programming, text alignment programs that are used to align bilingual texts, speech recognition and synthesis technologies to create and/or check pronunciation of words, morphological analyzers to provide easy access to corpus and dictionary look-up, parsers to clarify linguistic structures, and finally machine translation (MT) as an application builder ([Heift & Schulze, 2003](#)).

Nowadays, ASR, MT and the combination of different technologies are largely applied ([Levy & Stockwell, 2013](#)). Advances in technologies further enabled the manifestation of mobile-assisted language learning (MALL) applications, which have expanded the learning and teaching spectrum by transcending the time and the place boundaries ([Kukulska-Hulme & Shield, 2008](#)). Through the use of these technologies, on top of the classroom activities, the students are encouraged to practice learner autonomy, proceed at their own pace, and to participate in real-time interactions

with the teacher, their peers and the native speakers (Levy, 1997; Chapelle, 2001; Schwienhorst, 2012). CALL and MALL as ubiquitous devices have facilitated the access to the authentic materials and enabled flipped classroom. In doing so, these technologies are anticipated to bridge the gap between formal and informal learning, promote learner's motivation, and foster language learning.

In recent years, several CALL systems were introduced to create a more comprehensive language-learning environment where negotiations between the learner and the system were enabled. This was done by the advances in ASR technology, which allowed CALL systems for training the pronunciation and communication skills, hence focusing on speaking skill development. It should be noted, however, that while diverse CALL systems have been implemented either focusing on a particular aspect of language learning or encompassing several aspects, little attention has been given to building CALL systems that focus entirely on listening skill development. Instead, more and more attention was paid to the use of advanced technologies such as ASR systems for improving other aspects of L2 learning such as speaking skill development.

2.2 ASR Systems in Second Language Learning

The increasing need for innovative tools that foster language learning procedure has led to a growing interest in CALL systems that utilize ASR technology. Some possible applications include the pronunciation evaluation in order to improve oral skills and caption generation in order to facilitate listening comprehension (Shimogori et al., 2010; Thomson & Derwing, 2014). These systems can improve oral proficiency, which is considered as a problematic skill considering the time investments and the costs.

CALL systems with the integration of ASR technology are often known as CAPT (Computer Assisted Pronunciation Training) systems, and has gained increasing attention for their ability, which includes understanding the learner's input, reacting and providing feedback on the learner's pronunciation quality, thereby realizing a more realistic learning process (Neri et al., 2003).

ASR is also used for generating automatic captions for the videos, hence provide the learners with the textual form of the speech to train L2 listening skill at no cost. However, the ASR generated captions often involve a certain amount of errors, which result in imperfect captions that are not appropriate for language learning purposes.

While learners can in certain ways benefit from learning with speech-enabled systems, researchers are still skeptical about the usability of ASR for certain L2 learning purposes. In this study, an overview of the ASR framework, the errors generated by these systems and their applications in L2 teaching and learning is provided for a better understanding of the merits and demerits of such systems. This can lead to the discovery of new possible applications of ASR technology in the domain of foreign language learning and teaching.

2.2.1 ASR Framework

The speech recognition is a task that strives to find a mapping from acoustic observation to a single or sequence of words (Adami, 2010; Jurafsky & Martin, 2014). The acoustic observation is represented by $X = (x_1, x_2, \dots, x_t)$ that are taken from a speech signal. The sequence of words $W = (w_1, w_2, \dots, w_t)$ is taken from a fixed and known set of possible words, ω .

A statistical framework for speech recognition selects the sequence of words that is most likely to be produced given the acoustic observation. When the acoustic observation X is observed, the probability that the words W were spoken is denoted by $p(W|X)$. Selecting the most probable sequence of words, $\tilde{W}$, is formulated as

$$\tilde{W} = \underset{W \in \omega}{\operatorname{argmax}} p(W|X) \quad (2.1)$$

Direct modeling of the posterior probability $p(W|X)$ is difficult, therefore, it is commonly reformulated by applying the Bayes' rule.

$$\tilde{W} = \underset{W \in \omega}{\operatorname{argmax}} \frac{p(X|W)p(W)}{p(X)} \quad (2.2)$$

where $p(W)$ is the probability with which the word sequence W is uttered (the language model), and $p(X|W)$ is the probability with which the speaker utters sequence W in the acoustic form X (the acoustic model), and $p(X)$ is the probability that acoustic observation X is observed. The last term is often ignored in the maximization operation. A decoder encapsulates the process of searching through all possible sequences of words W that maximize the given equation.

When there are several possible pronunciations π for the word sequence X , the probability $p(X|W)$ is calculated as

$$p(X|W) = \sum_{\pi} p(X|\pi)p(\pi|W) \quad (2.3)$$

in which $p(\pi|W)$ indicates the probability of W uttered as π and $p(X|\pi)$ is the probability of pronunciation π takes the form of acoustic observation X . By considering the most likely pronunciation, the speech recognition equation should be rewritten as

$$W' = \operatorname{argmax}_{W, \pi} p(X|\pi)p(\pi|W)p(W) \quad (2.4)$$

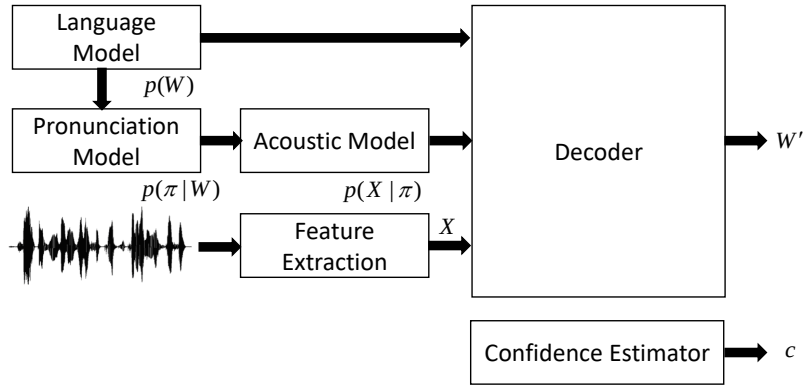


Figure 2.1: Typical Framework of ASR.

Figure 2.1 illustrates the main modules of an ASR system. Based on the given formulation, a typical ASR system is composed of the following components:

Feature Extraction: The speech features extracted from the audio signal as the acoustic observation.

Acoustic Model: This component calculates the likelihood of $p(X|W)$ using pattern recognition techniques. To this end, subword units such as phones, initial-finals, and syllables are used.

Language Model (LM): This component computes the probability of $p(W)$ by using a generative model that serves as a linguistic prior probability to constrain the word sequence. Common language models are grammar-based LM, statistical N-gram LM, and neural network-based LM. While grammar-based and statistical N-gram LMs can be used in the decoding process, the neural network-based LM can only be used for re-scoring of the initial decoding result.

Pronunciation Model (Lexicon): A pronunciation model defines the way a word is composed of sequences of subword units, e.g. phones. Some of the words have more than one possible way of pronunciation, and as a result, they have multiple entries in the lexicon. Equation (2.2) showed the way that multiple pronunciations can be considered for a word, and the final pronunciation is automatically determined by the Viterbi algorithm. The probability table is obtained by processing a corpus with pronunciation annotation.

Decoder: The decoder searches for the best matching word sequence to the acoustic evidence using Equation (2.4). It is the most complex component of speech recognition, and there are numerous variations in the algorithms and features it can use. The “Julius” decoder that is used in this study is a two-pass large vocabulary continuous speech recognition (LVCSR) decoder and performs almost real-time decoding on most current PCs in 60k word dictation task. The system utilizes word N-gram and context-dependent HMM and supports various search techniques such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, and Gaussian selection. This system is modularized carefully to be independent of model structures, and various HMM types are

supported such as shared-state triphones and tied-mixture models. Other famous decoders include HTK¹, Sphinx², RASR³, Juicer⁴ and Kaldi⁵.

Confidence Estimator: This module calculates a recognition confidence score W' that indicates the reliability of the recognition result c . The most popular method for generating the confidence measure scores (CMS) is an approximation of posterior probabilities over a lattice and its variants. However, in recent years, discriminative models such as conditional random fields (CRF) models and deep neural network (DNN) are used, which combine multiple sources of information such as acoustic, lexical and linguistic features with contextual information to estimate the CMS.

2.2.2 Factors Causing ASR Errors

The performance of speech recognition systems is usually evaluated in terms of accuracy and speed. Accuracy is usually measured with word error rate (WER), whereas speed is measured with the real time factor. Analyzing ASR errors is a difficult task because there are several types of errors, which can range from a simple mistake on the number agreement to the insertion of an irrelevant word for the overall understanding of the sequence of words. They can also affect neighboring words and create a whole area of erroneous words (Ghannay et al., 2015). The key issue of the ASR is the nature of the encodings used in this complex system: (i) in the level of integrating prior acoustic, phonetic and linguistic knowledge; (ii) in the level of algorithms for estimating the parameters of the model(s), (iii) in the level of recognition i.e., finding the most likely interpretation of the observations given the overall model, and (iv) in the level of integrated search over all of the constraints. These principles lead to a set of behaviors that may lead to various types of ASR errors (Moore & Cutler, 2001). These errors appear in the form of insertions, deletions, and substitutions and can be detected when the ASR transcript is aligned with the original transcript.

¹<http://htk.eng.cam.ac.uk/>

²<http://cmusphinx.sourceforge.net/>

³<https://www-i6.informatik.rwth-aachen.de/rwth-asr/>

⁴<https://github.com/idiap/juicer>

⁵<http://kaldi-asr.org/>

In general, the accuracy of speech recognition systems are largely influenced by the following factors:

- Vocabulary size and confusability,
- Speaker dependence vs. independence,
- Isolated, discontinuous, or continuous speech,
- Task and language constraints,
- Read vs. spontaneous speech,
- Adverse conditions.

In this view, many factors play a role in degrading the ASR system performance including the external factors like environmental noise and/or intra- and inter-speaker variability as well as the internal factors like acoustic models and acoustic similarities between word sequences. While the robustness of ASR systems against extrinsic variability (especially when arising from additive noise) has been studied extensively, the robustness against intrinsic variations of speech (i.e., the natural variability that is produced by the speaker) is not clearly understood ([Pietquin & Beaufort, 2005](#)). Factors that contribute to such intrinsic variability include foreign and regional accents, speaker physiology, speaking style, the rate of speech, the speaker's age and emotional state. These variations degrade the classification performance of automatic recognizers even in the optimal acoustic conditions ([Benzeghiba et al., 2007](#)).

Many studies have investigated the influence of variations in speaking rate on the performance of ASR systems, given that slow and fast speech rate cause different problems for these systems ([Mirghafori et al., 1995](#); [Siegler & Stern, 1995](#); [Stern et al., 1996](#); [Martinez et al., 1997](#)). In this regard, fast speech is known to increase deletion and substitution errors, whereas slow speech is reported to increase insertion errors ([Martinez et al., 1997](#); [Nanjo & Kawahara, 2004](#)). Utterances slower than 3 syllables/sec or faster than 6 syllables/sec have more word-recognition errors than their counterparts in the normal speaking range ([Greenberg & Chang, 2000](#)).

Speaking effort (i.e., loudly and softly spoken speech), speaking style (read or spontaneous), dialect and accent (i.e., phonological differences compared to a standard language that depends on the regional origin of the speaker), and children speech

are other influential factors to degrade the performance of the ASR system (Meyer et al., 2011).

The analysis based on articulatory features showed that for utterances with increased speaking effort and high speaking rate, the differentiation between voiced and unvoiced sounds was especially problematic in ASR (Meyer et al., 2011). There is approximately three times the number of articulatory-feature errors (per phone) when the word is misrecognized, regardless of the position of the segment within the word (Greenberg & Chang, 2000).

Additionally, ASR performance degrades when recognizing accented speech and non-native speech (Kubala et al., 1994; Lawson et al., 2003). Un-modeled factors in ASR systems such as out-of-vocabulary words (Chen et al., 2013) or unfamiliar phonemes uttered by non-native speakers (Flege et al., 2003) have been also considered as one of the major sources of the ASR errors.

In the spontaneous casual speech or under time pressure, reduction of pronunciations of certain phonemes or syllables often happens. It has been suggested that this “slurring” affects more strongly sections that convey less information. In contrast, speech portions where confusability (given phonetic, syntactic and semantic cues) is higher tend to be articulated more carefully, or even hyperarticulated. This, in turn, causes higher ASR error rate in the less informative category of words such as determiners or function words (Bell et al., 2003).

The composition of the words is another role-player in the performance of ASR systems. Acoustic perplexity rules were studied by Pietquin and Beaufort (Pietquin & Beaufort, 2005) to investigate what words are most confusable with each other when transcribed by ASR. Analyzing the syllable structures revealed that the highest error rates are associated with vowel-initial syllables. Syllables with complex (i.e., consonant cluster) onsets and codas tend to exhibit a relatively low word-error rate, as do polysyllabic words. This effect is particularly pronounced with respect to deletions, and vowel-initial syllables are particularly prone to such errors (Greenberg & Chang, 2000).

Prosodic stress also affects the performance of the ASR systems. There is a higher probability of a recognition error when a word is entirely unstressed. The relation between lexical stress and the word-error rate is particularly apparent for deletions and is manifested across all ASR systems (Goldwater et al., 2010).

Finally, the length of the word is known to affect the ASR performance with more errors generated on shorter words (Shinozaki & Furui, 2001; Goldwater et al., 2010).

2.2.3 ASR Applications in Second Language Learning

Speech recognition technologies used to train L2 speaking skill can be divided into two main groups. One includes those dealing with discrete speech input, and the other one is able to handle continuous speech input. More precisely, the discrete speech recognition system analyzes single patterns, which are known to the system and is mostly used to train pronunciation or fluency, where the user can choose from a predefined set of patterns. Continuous speech recognition, on the other hand, aims at analyzing free and fluently spoken input (Thomson & Derwing, 2014). However, ASR technology used for discrete speech recognition is still more reliable than continuous speech recognition, because the user's input is limited to several options or predefined patterns (Adami, 2010). Continuous speech recognition, on the other hand, allows processing freely produced sentences. Given the limitations of these systems to handle non-native speech, only a few CALL systems have integrated this type of ASRs. In fact, developing CALL systems based on ASR technology that can provide training and feedback for second language speaking is not trivial. There are two main problems that these kinds of the system should deal with: (i) handling non-native speech and (ii) dealing with erroneous speech.

Handling Non-native Speech: Non-native speech is atypical in many aspects, which poses serious problems to ASR systems (Lawson et al., 2003). Non-native speech may differ from a native speech in terms of pronunciation, morphology, syntax, and the lexicon notably due to the interference of the first language (Flege et al., 2003). Such difference may concern prosodic or segmental aspects of speech or both, which can blur phonemic differences and hence have serious consequences for intel-

ligibility (Van Doremalen et al., 2009). This big pronunciation difference makes it difficult for the system to provide a correct analysis. A possible solution would be to elicit restricted output from learners by having them read aloud an utterance from a constrained set of answers with limited freedom to formulate responses, as in systems like “*Subarashii*” (Ehsani et al., 2000) and the “*Let’s Go*” (Raux and Eskenazi, 2004). While this solution would alleviate the problem, it is notable that more freedom is needed for user responses in ASR-based CALL systems to let the learners produce target language in meaningful ways (Chapelle, 1998).

Dealing with Erroneous Speech: Learners’ spoken input to the CALL system would include mistakes. Thus, learner’s utterance needs to be recognized and to be diagnosed for mistakes, assessed and corrected (Tsubota et al., 2004). Moreover, it is important to provide the learners with meaningful and valid feedback in order to improve their speech. Assessment entails more than just using the statistical recognizers. Pronunciation assessment should be able to distinguish between possible pronunciations of words, whereas traditional ASR attempts to recognize commonalities between different pronunciations of words, which means distinguishing between the words, not pronunciations. Thus, the best tool for speech recognition is not necessarily the most proper one for pronunciation assessment. A proposed solution to this problem involves the attempt to improve decoding and comprising methods for the acoustic models, the language model, and the lexicon in order to compensate for the deviations in pronunciation, morphology, and syntax. The main task of these systems is to detect each phoneme, which is pronounced, and to assess how close the pronunciation was to that of a native speaker. Finally, it aims at correcting the learners’ mistakes by giving them the feedbacks or by providing them with the correct pronunciation (Neri et al., 2008). However, to achieve high speech recognition performance including accurate detection of erroneous utterances by non-native speakers is still an ongoing research.

Apart from speaking skill development, ASR systems with reasonably high performance can contribute to facilitating listening comprehension and understanding native speakers in real-time conversation by providing a real-time or automatic tran-

script for the given speech (Pan et al., 2010; Shimogori et al., 2010). Munteanu et al. (Munteanu et al., 2006) used ASR to generate transcripts of webcast lectures for examining native speakers' comprehension on the videos. They found out that ASR generated transcripts are useful when word error rate (WER) is lower than 20%. This finding was generalized to L2 learners in a study by Shimogori et al. (Shimogori et al., 2010), who suggest that captions with 80% accuracy improve the understanding of Japanese learners of English. However, the errors generated by these systems, particularly the delay caused by the technical problems, make the output inevitably inappropriate and relatively misleading for the purpose of training listening skill, specifically at the beginning stage.

To sum up, ASR technology has not yet fully exploited in the field of language learning and teaching. The acoustic models (speaker dependent, independent, or adapted), input quality (noise levels, microphone, sound card), and input style (discrete or continuous input) have an impact on speech recognition performance. Despite these limitations, ASR technology has formed an integral part of many language learning tools and CALL systems particularly for evaluating, training and improving L2 pronunciation and speaking skill (Neri et al., 2003; Witt, 2012; Thomson & Derwing, 2014). However, this technology has rarely been used for training and enhancing L2 listening skill. In fact, to design systems for training L2 listening is a demanding task, as many factors influence perception, recognition and comprehension in the listening process and different skills must be used hand-in-hand ranging from top-down to bottom-up strategies, the use of background and contextual schema combined with the understanding of the acoustic or phonological signals, words, etc. (Mayor, 2009). Thus, understanding the factors that affect L2 listening skill is a prerequisite step toward developing a system that can effectively lead to training this skills.

2.3 Factors affecting L2 Listening Skill

A number of factors in speech and language varying from acoustic level to lexical, syntactic and pragmatic level affect comprehension. While each of these features plays

a role in listening difficulty, some are largely referred to as the dominant obstacles of L2 listening, as reported in the following.

2.3.1 Listening Strategies

Many complex cognitive processes underlie the listening construct. Both top-down and bottom-up strategies, for instance, play active roles in listening comprehension. Top-down strategies refer to the use of background and contextual knowledge, exploited to construct meaning. In contrary, the bottom-up strategy is to derive meaning from processing small units (such as phones), decoding the sounds and building up. Both top-down and bottom-up strategies play a crucial role in effective comprehension, but L2 listeners need to learn the extent to which they should use each strategy and make a balance between the two processes (Rost, 2005). However, the majority of L2 learners overemphasizes the use of the bottom-up process, hence ignore the contextual clues and prior knowledge. Such learners often adopt hindering strategies such as word-by-word decoding, mental-translation and over-reliance on using bottom-up process (Osada, 2004). Learners who adopt such strategies tend to follow every word and are obsessed with grasping the meaning of each word in order to gain full comprehension. Given the limitations of working memory and the speed of speech flow, these learners face a lot of anxiety and thus fail to comprehend the audio (Hasan, 2000; Goh, 2000; Osada, 2001). Accordingly, it is sometimes necessary for the learner to tolerate vagueness and deal with the incompleteness of understanding (Vandergrift, 2011). To guide learners towards successful comprehension, listening instructions should encourage a strategy to link information, infer meaning and draw conclusions without heavily fixating on the perception of each word (Mayor, 2009).

2.3.2 Listening Materials

Apart from the listening strategies, there are different factors that may hinder or interfere with the listening comprehension process. Above all is the aural medium itself, which influences the listening comprehension in certain ways (Thompson, 1995). In this view, the content attributes such as the pragmatic information, the topic of

the material, and the length of the input are other influential factors on listening comprehension.

Pragmatic Information: Pragmatic information refers to the understanding of implied meanings, indirect message or culturally related information that leads to the complexity of the listening material. Insufficient pragmatic knowledge, hence, results in listening deficiencies (Bloomfield et al., 2010).

Length of Material: The length of the audio may influence the extent to which the learners can follow the speech, retain and process the information. The length of the speech can be calculated in different measures such as syllables per second, the duration of time, and the total number of words or sentences. Long speech may flood the learners with information which overwhelms their working memory and restrain their comprehension (Henning, 1990).

Topic of Material: Considering recall comprehension, greater working memory is related to better performance for familiar topics (Leeser, 2007). Besides, the role of background knowledge for L2 listeners emphasizes that passages about familiar topics are typically easier to comprehend than unfamiliar ones (Tyler, 2001; Sadighi & Zare, 2006).

2.3.3 Lexical Factors

There are a number of lexical factors that can impede listening for L2 learners. These factors which include word frequency, specificity, idiomaticity, length, part of speech, etc. are reported in many studies to cause L2 listening difficulties (Nissan et al., 1995; Schmitt & McCarthy, 1997; Bloomfield et al., 2010; Révész & Brunfaut, 2013).

For instance, when the speaker chooses the difficult words, which are beyond the vocabulary size of the listener, comprehension may be impeded as the listener finds it difficult to recognize such words, hence often fails to grasp the overall meaning of the speech (Goh, 2000; Bloomfield et al., 2010; Webb, 2010). However, to define a difficult word is a persistent problem and requires extensive investigations. While there are many factors that account for the difficulty of a word in terms of listening, the followings present some key features:

Frequency: Basically, the more frequent a word is, the more likely it is for the learner to know it or to have encountered it before. Furthermore, the occurrence of infrequent words in speech is correlated to its complexity and would lead to difficulty in comprehension (Nissan et al., 1995; Bloomfield et al., 2010).

Specificity: Words for specific purposes are often more difficult to learn. For instance, if a learner intends to advance to high-level academic study such as graduate school, there is a clear need to learn academic words (Schmitt & McCarthy, 1997).

Idiomacity: Comprehending idioms or idiomatic expressions can be difficult because even high-frequency words in the context of idioms may have a different meaning from what they commonly mean. Learners, however, often assume that the meaning of an idiom equals the sum of meanings of its components (Laufer, 1990). A positive correlation is found between the presence of an idiom in a passage, used for the test item, and the item difficulty (Kostin, 2004).

Polysemy: Polysemous words have more than one related senses (e.g., the word “class”). When words have multiple related senses, their meanings overlap (Murphy, 2004) and learners often mistakenly assume that the familiar sense is the only meaning (Laufer, 1990).

Word Length: The length of a word has a strong effect on its recognition. The shorter words are often more frequent, however, these words are usually less focused in the listening process (Field, 2003).

Part of Speech: When addressing the difficulty of a word, it is important to consider which part of speech it belongs to. For instance, nouns predominate over predicate terms in most of the languages, as nouns are often learned before verbs (Gentner, 1982).

Different Coding of L1: Native language of the learners can affect confusion of similar lexical forms. This assumption contributes to the understanding of why some foreign language words are more problematic for some learners than others (Ludwig, 1984; Laufer, 1990).

2.3.4 Acoustic, Speech and Perceptual Factors

Apart from lexical factors that affect L2 listening comprehension, many factors in speech and acoustic levels such as speech rate, pronunciation, accent, noise, distortion etc. can also influence comprehension. Furthermore, perceptual confusion is another crucial challenge for L2 listeners. The perceptual difficulty is indeed a fundamental factor for listening comprehension impair, which receives the most complaints from language learners. The following summarizes some of the significant factors in this category that account for listening difficulties:

Speech Rate: The fact that there is no option to control the speed of delivering the speech makes it difficult for the listener to comprehend the audio. Speech rate is defined as the number of words or syllables per time unit, which often involves pauses or silent intervals (Tsao & Weismer, 1997). Whether it is too fast or too slow, speech rate can deteriorate listening comprehension (Wingfield et al., 1985; Dunkel, 1988).

Pronunciation, Stress, and Intonation: Pronunciation, stress, and intonation are easily modified by the speaker when delivering the speech. Pronunciation can be ambiguous or unclear at some points due to speaker's pronunciation style, which is difficult to be recognized. Likewise, stress and intonation patterns may influence listening comprehension (Osada, 2004).

Accent: Speaker's accent also influences successful retrieval of the information by the listener (Floccia et al., 2009). For non-native listeners particularly, comprehension of unfamiliar accent is a demanding task and even the difference between American and British accents may hamper comprehension if the learners are more used to one rather than another.

Interlocutor's Variations: Speaker's gender and emotions are found to influence listening for L2 learners. Male speakers have a faster articulatory rate, which makes it difficult for L2 listeners to follow their speech as compared to female speakers (Quené, 2007). Moreover, speaker's vocal emotion can alter his/her speech rate, which in turn may influence listening comprehension. For example, anger and fear

are associated with higher speech rate, whereas sadness and disgust will lead to lower speech rate (Murray & Arnott, 1993).

Hesitation and Pauses: Hesitation and pauses are sometimes inevitable when a person is giving a speech especially when it is given spontaneously. Pauses, repetition and fillers are used by the speaker for various reasons such as repairing the false starts, adding extra explanations, and making more time to think of what to say (Buck, 2001). These disfluencies are considered as beneficial for some listeners but hindering for others. Basically, the effect is dependent on the proficiency level of the listeners. Beginners may fail to recognize these signals and cannot identify whether they are used for repeating a point or as afterthoughts, etc. (Underwood, 1989).

Distortion and Noise: Acoustic distortions significantly influence the listening comprehension of any listener, including the native speakers (Adank et al., 2009). Distortion and noise are known as profound barriers for listening as degraded acoustic input interferes with the recognition process (Aydelott & Bates, 2004).

Pronounceability: This factor is associated with the difference between the learner's first language and second language. Pronounceability has a strong effect on word difficulty as the acoustic similarity is considered more important than orthographic similarity (Ellis & Beaton, 1993).

Perceptual Confusion: While words may be defined as a fixed sequence of phones, the learner does not have direct access to this information. Rather, what the learner experiences in the input is complex statistical information over a corpus of utterances resulting from the concatenation of sub-word units (Saffran et al., 1996, p. 608). L2 listeners process the speech mostly in word-level, but they often do so through rough approximations and in many cases their matches do not correspond exactly to the sounds that they have heard (Field, 2008). Thus, learners' inability to recognize specific phonemes causes inaccurate guesses about the target words, which can inhibit them from processing the following words accurately. For instance, if the learner confuses the word "worth" with the word "worse" when listening to a sentence, it is likely that he/she goes a way off track. There is no doubt that the context would be helpful to disambiguate in such cases, but this requires the

effective use of top-down strategies to link the ideas together and correct the initial assumptions, which is not easy for the majority of language learners (Osada, 2004).

In this view, phonological neighbors such as minimal pairs and identical phonological forms such as homophones make recognition difficult for L2 learners as they end up with several potential candidates to choose from and need to use the speech context to decide on the correct interpretation (Weber & Cutler, 2004). Minimal pairs as words with a single phonemic difference (e.g., “face” and “faith”) are confusing for L2 listeners because this subtle difference makes the two words completely distinct in terms of meaning. The inaccurate recognition of minimal pairs might disrupt speech comprehension, as listeners have to resort to contextual information to determine the intended meaning of a word (Broersma, 2012, p. 1206). The same argument holds for homophones, where the pronunciations of the words are identical, but the meanings are totally different (e.g., “plain” and “plane”). Listeners need to go through a high-level of semantic analysis to distinguish these words as different meanings may be activated and active competition may happen (Weber & Cutler, 2004; Broersma, 2012). Given the dynamic nature of speech, making the right decision on multiple simultaneous lexical activations requires much skill and learners frequently go off the track as the first full word they hear may not be the intended word, but only a spuriously embedded form (Cutler, 2005). With more active candidates, more competition occurs, which in turn slows down the recognition process (Norris et al., 1995).

The perceptual confusion is not bounded to individual word, but can be extended to the misrecognition of boundaries between the words. The onsets and the endings of the words are not clear in the speech as compared to the text, which leads to difficulty in word segmentation for the majority of L2 listeners (Vandergift, 2007). Learners need to parse and segment the speech as they hear the continuous stream of sounds and they need to rely on their lexical knowledge, segmentation strategies, rhythmic cues, etc. (Weber & Broersma, 2012). Language learners try to scan the utterances to find the familiar matches to the known vocabulary (Cutler, 2005). Given that pauses occur in the natural speech only at about every 12 syllables (Field, 2003), continuous and consistent segmentation of the speech is a very challenging task for

L2 listeners compared to the readers who have access to markers of word boundaries. Thus learners often fail to locate the word boundaries in the process of matching the utterances to the words, which leads to breached boundaries (Field, 2008). When a listener misrecognizes the phrase “*made out*” as “*may doubt*”, this small mistake may hinder the effective processing of the next speech segments.

The role of the speaker is also important in producing a word in a way that it may be misrecognized with its phonological neighbors or leads to breached boundaries. Words can be deviated from their standard form in the connected speech due to the phenomena such as assimilation, resyllabification, reduction, etc. (Field, 2008). Consequently, errors in auditory discrimination and/or articulation of these sounds may result in misunderstanding and misinterpretations of the word, phrase or sentence (Nilsen & Nilsen, 2010, p. 15). However, detecting such kinds of perception or production confusions is very difficult unless there is a source for analyzing intrinsic speech difficulties.

2.4 Captioning Methods

While many factors affect the listening process and make listening comprehension difficult for language learners, a number of tools are developed to assist these learners in handling difficulties of comprehending the speech. In this view, captions are used to help L2 listeners improve certain aspects of the target language. The type of captioning, however, influences the effect of this assistive tool on language learning. Although the conventional full captioning method is still the mainstream of contemporary education, other methods such as keyword/paraphrase captioning have drawn some attention (Garza, 1991). Moreover, the advances of the ASR technology have enabled the generation of synchronized captions. Unlike the typical captions where chunks of words appear on the screen, in synchronized captions, the emergence of words on the screen is concurrent to the speaker’s utterance. This method fosters word recognition but promotes word-by-word decoding that is known as a hindering strategy.

2.4.1 Full Captions

Full captioning is defined as visual text delivered along with audio or video via multimedia where the language of verbatim text matches the spoken content (Markham & Peter, 2003; Leveridge & Yang, 2013). Without being affected by accent, pronunciation and audio deficiencies, full captions allow the listeners to parse the speech stream into meaningful chunks (Garza, 1991; Winke et al., 2010), which is an essential process for learning (Ellis, 2003). Full captioning also aids with the phonological visualization of audio to make listeners more certain of ambiguous input (Bird & Williams, 2002).

A considerable amount of literature has been published on beneficial effects of full captioning. Studies have investigated the effect of this method on word learning (Bird & Williams, 2002; Danan, 2004; Sydorenko, 2010; Montero Perez et al., 2013), reading development (Markham & Peter, 2003), word recognition (Bird & Williams, 2002) and listening comprehension (Danan, 2004; Taylor, 2005; Winke et al., 2010). A recent meta-analysis investigated the overall effect of full-captioned video on listening comprehension and vocabulary learning based on 18 studies. The findings revealed a large effect of full captioning on listening comprehension and vocabulary acquisition (Montero Perez et al., 2013).

Most of these findings are derived from the dual coding theory, which posits that by both verbal and visual inputs, learners can construct referential connections and thus learn more efficiently (Paivio, 1990). However, each of verbal and visual channels has limited capacity and can manage a limited amount of processing at a time (Mayer & Moreno, 2003). This assumption is central to the cognitive load theory and working memory theory (Baddeley, 1992; Sweller, 1994).

Although these theories have not been widely researched in the context of L2 (Diao et al., 2007; Sydorenko, 2010; Mayer et al., 2014), some studies indicated that when authentic videos are accompanied with captions, learners' differences in cognitive capacity may influence their attention and makes it difficult for them to focus on audio, video, and text (Taylor, 2005; Sydorenko, 2010). In this view, learners tend to

focus on the input which is easier to perceive and some learners prioritize reading over listening especially when they cannot divide their attention to the all sources of input equally (Lund, 1991; Pujolà, 2002; Chang, 2009). In such cases, many L2 learners may find the audio difficult to process quickly and use the captions as “the most understood, relevant and thus preferred stimuli” (Leveridge & Yang, 2013, p. 202). In line with this, in a study by Sydorenko (Sydorenko, 2010), who examined the effect of input modality and learner’s attention to the input, participants reported that they paid most attention to the captions, then to the video, and finally to the audio. While the combination of different inputs is beneficial for L2 learners in many ways (such as vocabulary learning, comprehension, form recognition, etc.), when the goal is to train the listening skill, full captions may not be a preferable tool (Diao et al., 2007) and hence the input should be enhanced to fit this purpose. Meanwhile, different types of captions (e.g., verbatim or keyword) can influence listening in different ways (Garza, 1991). Accordingly, some alternative methods such as keyword captioning and word-level or phoneme-level synchronized captioning have gained instructional value.

2.4.2 Synchronized Caption

Advancement of automatic speech recognition (ASR) technology has enabled automatic text-to-speech alignment, which led to the development of word-level or phoneme-level synchronized captioning (Braunschweiler et al., 2010). Here, synchronized captioning is to show the words or phonemes in the caption one by one, which is realized by mapping each word or phoneme to its corresponding speech segment. In this sense, the major difference between this method and the full-captioning method lies in the synchronization unit (See Figure 2.2).

In the full-captioning method, synchronization is done in sentence level so that chunks of words appear on the screen, stay there for seconds, and disappear, before another chunk will appear. On the other hand, word-level synchronized captioning is a sequential word-by-word captioning, in which the emergence of words on the screen is concurrent to the speaker’s utterance, i.e., instead of chunks of words appearing altogether, words appear one after another, from left to right, and in precise synchrony

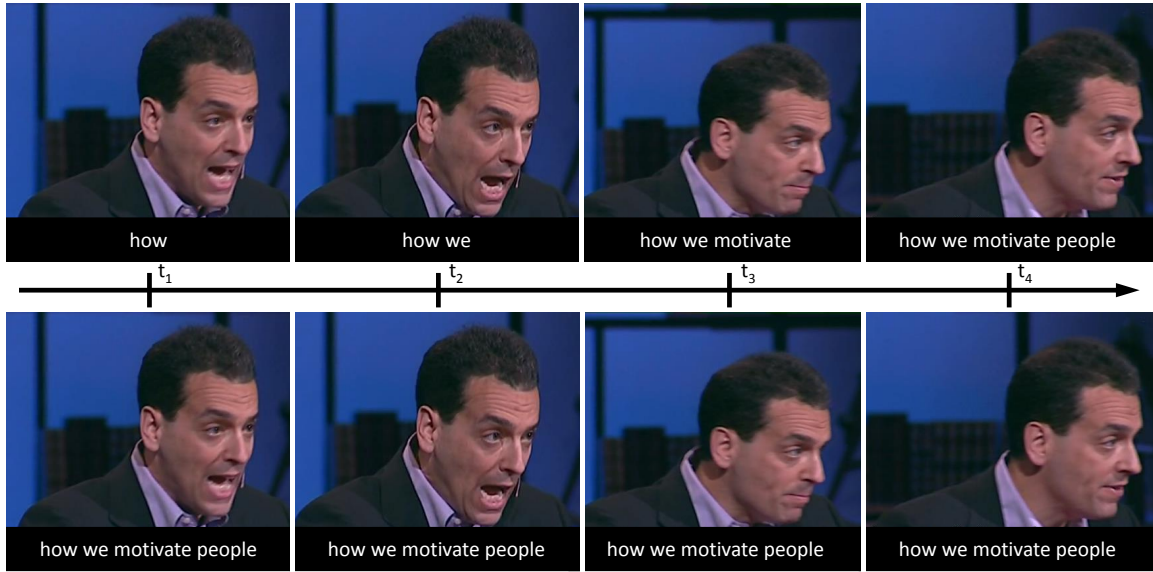


Figure 2.2: Screen shot for a TED talk with word-level synchronized caption (top) vs. conventional full caption (bottom) for consecutive time frames ($t_1 \sim t_4$). ©TED Talk by Daniel Pink: Puzzle of motivation.

as the speaker says each word. Synchronization in phoneme-level is the same except that the alignment unit is set to the phonemes instead of words.

Some examples of making speech-to-text correspondences can be found in audio books, albeit in a different context. In this context, text-to-audio alignment is used to facilitate reading and text tracking. The Talking Books Project (Medwell, 1998) is a preliminary example in which the text is read while each word is highlighted as spoken. These books were found beneficial for assisting children in text-to-speech mapping, word decoding and L1 reading development (Medwell, 1998). Interactive electronic books are enhanced versions used to assist word reading and story comprehension (Korat, 2010). Likewise, digital books, which use ASR technology to align each spoken word with the read text can potentially help L2 learners in reading development and vocabulary learning (Trancoso et al., 2007).

In addition to audio books, “karaoke-style” display, where the text is highlighted in colors as the audio moves by, has gained some instructional value. Bailly and Barbour (Bailly & Barbour, 2011) developed a system that exploits the alignment of text with audio at various levels (letters, phones, syllables, etc.) and experimentally investigated the use of such a karaoke-style reading system for learning sound-to-letter

mapping. This system uses a data-driven phonetizer trained on an aligned lexicon of 200,000 French entries to display a time-aligned text with a speech at phoneme level. This system allowed the learners to select any segment of the audio and listen to it as the phonemes were highlighted incrementally. A cursor was set to move smoothly in real time under the phonemes that were being vocalized. In this study, the subjects' scores on a word orthography test were compared using two different audio + text reading systems: synchronous and non-synchronous. The study suggested that the multi-modality of synchronous-reading implicitly facilitates text-to-speech mapping in French for native and non-native subjects.

As found in prior research ([Medwell, 1998](#); [Trancoso et al., 2007](#); [Korat, 2010](#); [Bailly & Barbour, 2011](#)), word-level and phoneme-level synchronization can be beneficial for developing reading skill, facilitating aural-written verification, promoting word recognition and assisting word-boundary detection. While the effectiveness of this method on L2 listening development is hardly investigated, most of the aforementioned beneficial aspects are found equally advantageous for effective listening ([Vandergrift, 2007](#)). This method can consequently be used for L2 listening development to exploit its pedagogical effectiveness.

Meanwhile, caution should be taken as the method has potential to promote word-by-word decoding strategy, which hinders effective listening ([Vandergrift, 2004](#)). Accordingly, some complementary method may be needed to offset this effect. This can be accomplished through highlighting only particular words or sentences in the caption, as in keyword captioning.

2.4.3 Keyword Caption

Guillory ([Guillory, 1998](#)) examined the use of keyword captioning for beginning learners of French and reported that students who received keyword captions performed as well as those who received full captions. In her study, keywords were selected manually based on their importance to the main idea of the videos. Guillory discussed that “learners no longer need to be subjected to a volume of text to read; they can, in fact, comprehend authentic video with considerably less pedagogical support” ([Guil-](#)

lory, 1998, p. 95). She concluded that keyword captioning decreases cognitive load, improves multichannel processing, and motivates learners to listen more but read less.

Montero Perez et al. (Montero Perez et al., 2014b) investigated the effectiveness of keyword captioning and found no significant difference between the scores of keyword captioning group and no-captioning group in the comprehension question. Yet, they also investigated the perceived usefulness of the keyword captioning and reported that this method was highly distracting. The researchers argued that the salient and irregular appearance of keywords on the screen could lead to such distraction. Note that the appearance of keywords is not synched to their utterances.

On vocabulary acquisition, however, a study showed that keywords were as helpful as full captioning (Montero Perez et al., 2014a). In their study, full captioning, keyword captioning and full captioning with highlighted keywords were compared against the no-captioning condition. The findings revealed that the captioning groups scored equally well on form recognition and clip association and significantly outperformed the no-captioning group.

2.4.4 Limitations of Captioning Methods

The majority of studies reported positive effects of full captioning, but some argued that it cannot be concluded whether the learners' performance on the subsequent comprehension tests is based on reading the text or listening to the audio (Pujolà, 2002; King, 2002). Winke et al. (Winke et al., 2013) investigated what learners usually do when they watch a captioned video using the eye-tracker technology. Their data revealed that when caption was on the screen, learners read the text on average 68% of the time. We do not clearly know how learners balance their attention to simultaneous sources of input – audio, video, and text. However, there is skepticism that the full-captioning method may promote reading over listening (Guillory, 1998; King, 2002).

Word-level synchronized captioning is another method to provide textual clues along with the audio. This method fosters text-to-speech mapping but is likely to

encourage reading the captions perpetually and promoting word-by-word decoding as the synchronization feature facilitates reading to a great extent.

Keyword captioning, in contrary, aims to solve this problem by providing limited keywords. However, it suffers from the abrupt appearance of keywords on the screen, which makes it distractive (Montero Perez et al., 2014b). Another major drawback of this method lies in the manual selection of keywords, which is fairly content-specific and does not consider the proficiency level of the learners, hence may not provide each learner with an adequate amount of support. In this regard, Guillory (Guillory, 1998) concluded that the limited number of keywords in her study might not have provided enough information for the beginners.

On one hand, the learner needs to be able to deal with a real-world situation where there is no access to any supportive tool, and on the other hand, we cannot expect a non-native listener to follow the authentic input without any support. Hence, the listening instruction should focus first and foremost on assisting the language learners to cope with aural input difficulties while maintaining a tendency to develop compensatory strategies for listening in real-time. Thus, further research should be conducted to investigate an effective method for assisting learners to gain adequate comprehension, without becoming too much dependent on captions.

According to Krashen's input hypothesis (Krashen, 1985), learning occurs if the learner receives comprehensible input $[i+1]$, which is slightly above his or her current knowledge $[i]$. Considering the challenging nature of authentic inputs, they usually contain information far beyond the knowledge of the learner $[i+n]$. Given this fact, comprehension of such materials without any assistance leads to frustration for many L2 learners. Therefore, captioning should aim to make the authentic input comprehensible ($[i+1]$ instead of $[i+n]$) for different learners.

Incentivized by this demand, we investigate a captioning method that amends the shortcoming of the current methods and allows the learners to practice for a stage where they are able to cope with authentic materials without any assistance. To this end, we looked for a solution that strives to foster L2 listening and decrease dependence on the caption simultaneously. The next chapter introduces the proposed

solution, explains the main idea and the advantages of this method and evaluates the effectiveness of this method as compared to the other existing methods.

Chapter 3

Baseline Partial and Synchronized Caption

To facilitate the comprehension of authentic materials, captions are widely used as assistive tools. Existing captioning methods, however, suffer from some limitations and may not necessarily improve learners' listening skill, rather may encourage reading the text instead of listening to the audio. Contrarily, in a real-world communication, learners must solely rely on their listening skill, as no assistive tools are available.

This chapter presents a novel method of captioning, partial and synchronized caption (PSC), as an alternative captioning tool to train L2 learners' listening skill, while decreasing dependence on reading the captions. As the name suggests, PSC is based on partialization and synchronization methods. The proposition of adopting these two methods is motivated from a number of viewpoints. First, when full caption text is used, many language learners just read the caption rather than listening to the audio because they are often better at reading than listening ([Pujolà, 2002](#)). Thus, they try to comprehend the content by merely reading the text. In contrast, through limiting the number of shown words in PSC (partialization), the learners are encouraged to listen to the audio and read for a limited number of words.

Second, when the captions are synchronized in word-level, learners can constantly match each shown word to what they have heard and hence can smoothly follow the audio without being distracted by the emergence of the words on the screen.

Moreover, in the case of hidden words, learners are still able to follow the word boundaries with the help of synchronization.

PSC benefits from synchronization and partilization and it aims to scaffold the learners by detecting difficult words or phrases and presenting them on the caption while removing easy words to encourage the learners to listen more. This explanation highlights the importance of defining effective criteria for selecting difficult words. This chapter builds on the previous chapter and explains that the selection criteria of the baseline PSC system are based on an extensive review of L2 studies to learn the lexical or speech related factors that are most problematic for L2 listeners.

To realize the PSC system, a speech recognition system is trained using TED talks and employed to generate a caption text aligned in precise timing with the speech signal of the respective words (synchronization). Next, partial captions are automatically generated by selecting words/phrases, which are likely to hinder learner's listening comprehension (partialization). This selection is based on three features: speech rate, word frequency, and word specificity as dominant factors that impede L2 listening (feature extraction). The caption is then adjusted to the level of the learner by considering learner's vocabulary size and tolerable rate of speech (learner adaptation). The final caption includes only a subset of words, in accordance to the learner's needs, presented one after another in sync with the speech, both to encourage listening and to aid speech-to-text mapping. The effectiveness of PSC is evaluated in comparison with no-caption and full caption conditions by 58 Japanese learners of English.

3.1 Concept of Partial and Synchronized Caption

Based on the discussion in the previous chapter, full captioning is criticized for: encouraging a word-by-word decoding strategy, promoting the use of bottom-up skill (Osada, 2004), allowing comprehension of audio by just reading the text without listening (Pujolà, 2002), and imposing a high-level of cognitive load by providing a large number of textual clues together with the audio (Sydorenko, 2010). To

address these limitations, the author proposes partial and synchronized captioning (hereinafter, PSC) as an alternative to the existing methods for training L2 listening comprehension. Figure 3.1 illustrates a screenshot of PSC as compared with full captioning.

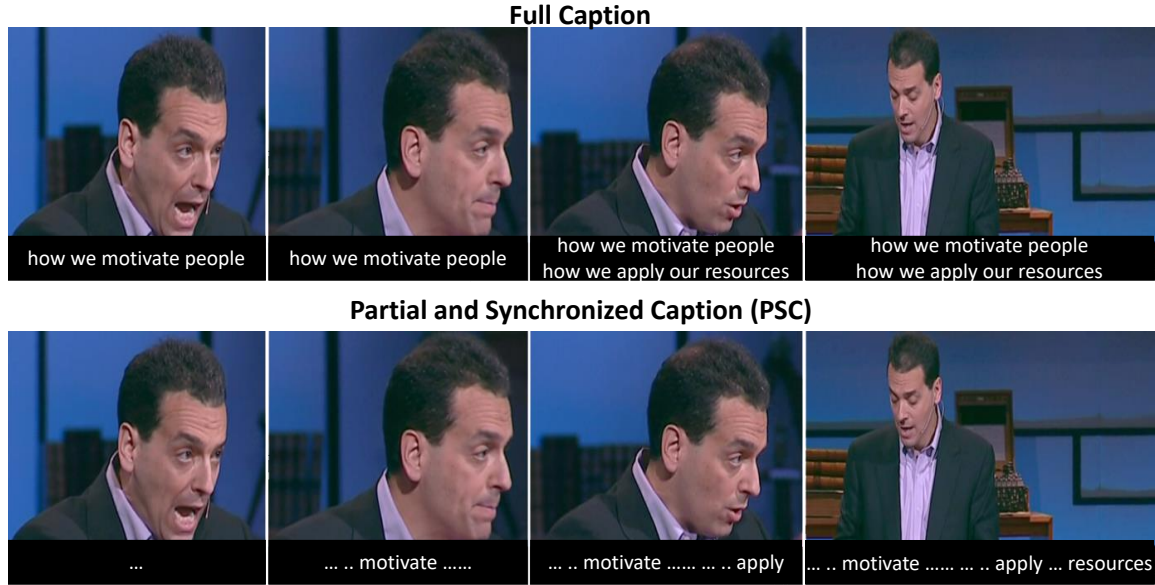


Figure 3.1: Screenshot of the full caption vs. the Partial and Synchronized Caption. © TED Talk by Daniel Pink: Puzzle of motivation.

PSC is 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. This system 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.

More precisely, in this method, the original transcript is automatically reduced to a partial caption, which includes only a selected set of words or phrases (partialization) while each word is forced to appear on the screen in sync with its utterance and in a one-by-one sequence (word-level synchronization). The word-level synchronization is realized by ASR technology, which precisely maps each word to its corresponding speech segment. The filtering process of words to be presented takes

into account not only the hindering factors of comprehension but also the assessed knowledge of the learner (learner adaptation).

Table 3.1: Comparison of different captioning methods

Different Advantages	Different Types of Captions				
	Full	Synchronized	Keyword	Partial	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			✓	✓	✓

Given that PSC benefits from both partialization and synchronization features, the integrated method involves the merits of each feature while their complementary nature counteracts the demerits of each other. As the Table 3.1 suggests, word-level synchronized captioning neatly presents the word boundaries and fosters word recognition, but may promote word-by-word decoding. To address this issue, partial captioning was proposed to limit the number of words in the caption. However, the partial captioning method alone suffers from the abrupt appearance of words on the screen, which is distracting for many language learners. Through the use of word-level synchronization together with partialization, the irregular appearance of the words in the partial captioning method can be circumvented. Another advantage of PSC lies in automation, which enables quick and effective generation of appropriate captions for different learners. Compared to keyword captioning, the selection criteria of PSC are both learner-specific and content-specific. This is important, especially because not every learner can benefit from a caption in which keywords are selected generally.

PSC aims to assist the learners to cope with aural input difficulties without constantly referring to the verbatim caption. It acts as an intermediary stage before the learner is totally independent of the captions.

Based on this explanation, the PSC system 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. But the main question is what criteria should be used for detecting difficult words and phrases i.e., what kind of features needs to be extracted to realize this goal?

3.2 Feature Extraction

In the partial captioning procedure, the system selects a subset of words from the transcript and presents them in the caption while masking the remaining words. The effectiveness of PSC highly depends on the choice of words to appear in the caption. Unlike keyword captioning, which considers important words to appear on the screen, in PSC the major obstacles of listening comprehension are considered as a prudent choice for word selection.

There are different factors that may interfere with the listening comprehension process. Both lexical and acoustic factors affect L2 listening comprehension. The lexical-level factors involve word frequency, specificity, idiomaticity, length, part of speech, etc. (Nissan et al., 1995; Schmitt & McCarthy, 1997; Bloomfield et al., 2010; Révész & Brunfaut, 2013), whereas the acoustic-level factors are known as accent, speech rate, pronunciation, noise, distortion and perceptual difficulties (Griffiths, 1992; Buck, 2001; Field, 2003; Osada, 2004; Cutler, 2005; Bloomfield et al., 2010).

In this study, words of “low frequency”, those delivered at “high speech rate”, and those recognized as “academic or specific terms” form the basis of word selection. These features were chosen above others for being repeatedly mentioned as major contributing factors to listening comprehension impair (Griffiths, 1992; Nissan et al., 1995; Schmitt & McCarthy, 1997; Révész & Brunfaut, 2013). Another reason for choosing these criteria is that factors such as speaker accent, noise, length and the topic of material can be circumvented by content selection, as it is also the case in this study. For instance, the videos of this study included TED talks delivered by American speakers and trimmed to short segments. Therefore, the videos are noise-free,

non-native accent is not involved, and length is controlled. However, word specificity is considered because the videos involve technical terms. Moreover, the factors adopted in this study are feasible to be implemented and quantified automatically by the existing technologies.

3.2.1 Speech Rate

Rost (Rost, 2005, p. 506) states that speech rate is “a major factor in the comprehensibility of speech for L2 listeners”. High speech rate can negatively affect both native speakers’ and L2 listeners’ comprehension (Buck, 2001). However, it is difficult to define a turning point where speech rate is beyond the learner’s tolerance. Nevertheless, language learners should be able to deal with normal speech rate. Although defining a reliable threshold for normal speech rate is yet another issue, studies have reported a range of 160 to 190 words per minute (Pimsleur et al., 1977) and 3.83 to 4.66 syllables per second (Tauroza & Allison, 1990).

Calculation of syllables is based on the structural syllabification of the text, which is realized by using Natural Language Toolkit (Bird et al., 2009)¹. The duration of each word is measured by using the time-stamps obtained from the alignment process of the ASR system. In PSC, the speech rate is quantized into several bins (from slow to fast) based on the standard rates of speech reported in Table 3.2, and set as the system’s default thresholds. Yet, the thresholds can be modified according to a learner’s preference.

Table 3.2: Standard Rates of Speech (Tauroza & Allison, 1990; Pimsleur et al., 1977)

Speed	Range (wpm)	Range (sps)
Fast	Above 220 wpm	Above 320 spm (5.33 sps)
Moderately Fast	190–220 wpm	280–320 spm (4.66–5.33 sps)
Average	160–190 wpm	230–280 spm (3.83–4.66 sps)
Moderately Slow	130–160 wpm	290–230 spm (3.16–3.83 sps)
Slow	Below 130 wpm	Below 190 spm (3.16 sps)

¹<http://nltk.org/>

3.2.2 Word Frequency

The occurrence of less frequent words is another problem of L2 listening, which may stimulate learners to pay too much attention to those words, therefore hindering the listening comprehension (Goh, 2000; Bloomfield et al., 2010; Webb, 2010).

The term word frequency can be defined as the relative number of times a given word is used in a language. In this view, the frequency of words in written or spoken corpora is related to word difficulty mainly because learners are less likely to be familiar with infrequent words (Nissan et al., 1995). Moreover, for both L1 and L2 listeners, processing low-frequency words takes more time (Buck, 2001; Bloomfield et al., 2010).

Generally, when watching captioned videos, good readers know how to scan for selected words (Guillory, 1998). Eye-tracking studies also suggest that high-frequency words are skipped more (Rayner, 1998) and fixated less (Inhoff & Rayner, 1986; Moran, 2012). Partial captions should then suffice many listeners by presenting infrequent words that interfere with comprehension.

The frequency of the word is calculated based on its occurrence in spoken or written corpora. Nation (Nation, 2006) categorizes English vocabulary into high-frequency (the most frequent 2000–3000 word families), mid-frequency (3000–9000 word families), and low-frequency (beyond the 9000 frequency band). Based on this, Nation and Webb (Nation & Webb, 2011) designed 25 word family lists, each including 1000 word families, plus four additional lists: (i) an ever-growing list of proper names, (ii) a list of marginal words including swear words and exclamations, (iii) a list of transparent compounds, and (iv) a list of abbreviations. The first two lists are hand-selected while the rest are based on the following two famous corpora:

- The British National Corpus (BNC)
- The Corpus of Contemporary American English (COCA)

In this study, the frequency of each word is calculated using the word family lists and cross-checked with COCA (Davies, 2008) to get an exact value. BNC includes 100+ million words of both written and spoken language from a wide range of sources,

designed to represent a wide cross-section of British English. On the other hand, COCA contains 520+ million words of text (20 million words each year since 1990) and is hence more comprehensive. COCA is the largest corpus of American English, which includes millions of words, equally divided among spoken, fiction, popular magazines, newspapers, and academic texts, and is updated regularly. It is more suitable for this study, which uses TED talks, delivered by American speakers.

3.2.3 Word Specificity

By word specificity, we refer to special words that are used in a particular technical domain. Examples include academic terms, jargon, and terminologies. According to Goh (Goh, 2000), limited vocabulary, especially for academic words, is often a cause of L2 listening comprehension impair. Révész and Brunfaut (Révész & Brunfaut, 2013) noted that a higher frequency of academic words is associated with greater listening difficulty. Webb (Webb, 2010) also showed that glossaries consisting of low-frequency word families and technical vocabularies have a great value in assisting comprehension. Since TED talks used in this study involve many academic or specific terms, word specificity is also considered as a feature for generating PSC. The system detects these academic terms by using Academic Word List (Coxhead, 2000). Besides, Academic Vocabulary List based on COCA (Gardner & Davies, 2013) is also consulted to achieve higher accuracy.

3.3 System Implementation

Based on the explanation in the previous section, it can be summarized that PSC is a system that uses TED talks (system input) for training L2 listening skill and enables the synchronization of the text to speech in word-level using ASR technology (synchronization). As a baseline, the detection of difficult words is realized based on three defined features: speech rate, word frequency, and word specificity (partialization). The level of difficulty and the amount of shown words in PSC is tailored to the requirement of different learners at different levels (learner adaptation). Accordingly,

the system, as shown in Figure 3.2, consists of three main modules, synchronization, partialization, and learner adaptation.

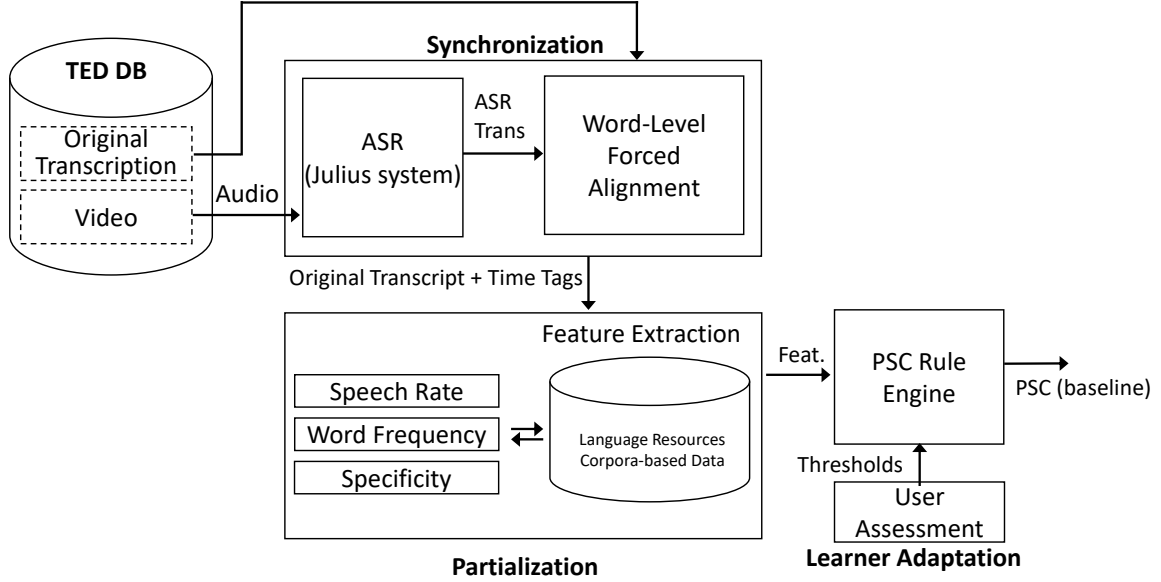


Figure 3.2: Schematic of Baseline PSC System. Baseline employs ASR system to synchronize words with their speech segments, and then partialize the text based on its features. The features then are tuned in learner adaptation module through receiving a threshold from the user, which is obtained by several tests that measure the vocabulary size and the tolerable speech rate of the learner to adjust the caption to the level of that learner.

3.3.1 System Input

TED² talks form the database of the system and used as the medium for PSC. TED is the abbreviation form for “Technology, Entertainment, and Design”. It is a non-profit organization that welcomes the world’s most fascinating people from every discipline to deliver a talk. TED is devoted to spread the ideas, usually in the form of short, powerful talks (18 minutes or less, roughly equivalent to 2,500 words), which cover a wide variety of topics. Each talk is well prepared and presented by a skillful speaker. TED prepares the video recordings of the talks available under the Creative Commons license. All talks with corresponding English captions are freely available on its website.

²<http://www.ted.com/>

TED talks are used for 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. These videos can meet different interests and immerse L2 listeners in listening to inspiring talks, while being exposed to the authentic material.

Automatic transcription of TED talks has been investigated in the IWSLT challenges. Huang et al. (Huang et al., 2013) reported word error rate of 19.7% by the GMM-based system and a 14.0% by the DNN-based system, which was trained using 167.8 hours of 760 TED talks only with caption texts instead of faithful transcripts. In their study, 81.1 hours of WSJ (Paul & Baker, 1992) and 62.9 hours of HUB4 English Broadcast news (Graff et al., 2002) provided by the Linguistic Data Consortium (LDC) were also used.

3.3.2 Synchronization

To make PSC, TED talks are transcribed by Julius v.4.3.1 (Lee & Kawahara, 2009). The system receives a video and its corresponding transcript as input, both taken from TED website. First, the audio is ripped from the input video and the embodied speech is transcribed automatically by the ASR system. Synchronization is realized by the word-level-alignment feature of ASR. For precise alignment, however, a dedicated acoustic-phonetic model is necessary. For example, standard speech recognizers that are trained with a corpus of read speech do not work well for spontaneous speech (e.g., lectures) even in the alignment task. Acoustic and language model used in this system were trained using 780 TED talks (180 hours) through the lightly-supervised learning method (Naptali & Kawahara,). The resultant ASR system outputs highly accurate transcripts with estimated timestamps for tokenized words (Figure 3.1).

Finally, the original transcript and the ASR output are aligned to make precise synchronization. This process is done through the force-alignment method to eliminate the ASR errors. The outcome of this process realizes word-level synchronization and specifies the onset of each word, which in turn enables the calculation of each word's duration.

3.3.3 Partialization

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 PSC caption based on three features: (i) speech rate as a dominant factor that 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 & 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 last factor also affects listening in a sense that many language learners are not familiar with these specific words (Webb, 2010; Révész & Brunfaut, 2013).

The feature extraction module processes the generated caption text and converts it into a feature vector. This module calculates the speech rate and frequency of each word in the transcript and detects the specific words.

The system first calculates the speech rate of the speaker, $sr(w_i)$, when delivering each individual word w_i , where $i \in \{1, \dots, N\}$. There are several 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 & Stern, 1995). SPS, on the other hand, has fairly uniform distribution over speech rate and is more robust against the variations in speech (Wang & 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 & Allison, 1990).

To estimate the frequency of each word, $fr(w_i)$, the system refers to two comprehensive corpora: British National Corpus - BNC, which includes 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 & 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 & Beglar, 2007), which are compatible with the word family list (Nation & 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 (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 & 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 & Webb, 2011), and the stop list.

The rule engine in the decision-making module decides on the inclusion or exclusion of a word in the final caption. In the first stage, the decision about a word is made based on the defined features, i.e., if the word has high speech rate (i.e., above the learner’s understandable rate of speech), low-frequency (i.e., beyond the learner’s vocabulary size), or it is categorized as an academic term. In the second stage, the system handles special instances such as abbreviations, proper names, num-

bers, interjections, transparent compounds, and repeated appearance of words. These general features act on each word, either as excitatory or inhibitory. For instance, abbreviations and proper names are always shown while interjections are discarded. Accordingly, it is possible to decide about the special instances by categorizing them into *keep* or *hide* categories:

$$keep(w_i) = \mathbb{1}(w_i \in \{ppn \cup abb \cup dcp\}) \quad (3.1)$$

and

$$hide(w_i) = \mathbb{1}(w_i \in \{itj \cup stp \cup ecp\}) \quad (3.2)$$

where indicator function $\mathbb{1}(\cdot)$ 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.

$$show(w_i) = \mathbb{1}\left(\mathbb{1}(fr(w_i) - \theta_{fr}) + \mathbb{1}(sr(w_i) - \theta_{sr}) + sp(w_i) + keep(w_i)\right) \times (1 - hide(w_i)) \quad (3.3)$$

In this view, the system carefully selects the appropriate words for the learners to foster L2 listening skill training.

3.3.4 Learner Adaptation

The final step is to tailor the caption to adjust for different language learners at different levels. At this stage, the system conducts several tests to estimate the learners' current level of proficiency. These include a vocabulary size test (Nation & 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 vocabulary size test is used to measure the knowledge of particular frequency levels of words (for example, the first 1000 and

second 1000 words). It covers 20,000 word families, consisting of 140 multiple-choice questions, which is used with only non-native speakers. The speech rate sample test tries to check the learner's listening speed by evaluating the learner's recognition after listening to a speech at various speeds (the different speeds are defined based on standard rates of speech). 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.

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.

3.3.5 Caption Generation

The next step is to generate the caption. To this end, the formatting and display module generates the final PSC using the user display parameters. These parameters regard the sequence of the words that should be readable and understandable for the learners. We handle words after numbers and words after "*apostrophe s*" in this version. If a word is decided to appear in the caption, it will be copied intact in the output caption, otherwise, a character mask (here we use "dots") replaces every letter of the word. This will emulate the speech flow and presents the location of each and every word in the caption in sync with the respective utterances. For example, "*express*" will be replaced by "....." and "*don't*" will be replaced by "....".

In addition, the readability of the captions is another issue, which affects the learner's focus on the listening comprehension although it is mostly ignored in the literature. For instance, inappropriate font size regarding the aspect ratio of the video, cluttered and miss-positioned caption area, surprising caption pop-up, short/long caption display duration, bad justification of the caption, and empty lines as the factors that lower the readability of captions are handled in our system. The output file is generated using the time tag of words, the number of words decided for one

caption line, the learner preferences, and the readability considerations based on the pre-defined settings.

Finally, to display the generated caption compatible with the latest media players, the captions should be converted into SAMI format. This structured markup language is created to standardize the playback of media in sync with the caption.

3.4 Experimental Evaluation of Baseline PSC

Given the novelty of PSC method, the following research questions investigate its potential effectiveness:

1. Do captioned videos (using PSC and full caption) lead to better comprehension compared to non-captioned videos?
2. Can PSC be substituted for the conventional full-text captioning method?
3. Do proficiency differences affect the usefulness of PSC?
4. Does PSC help the learner comprehend the subsequent segment of the video later without any captions?

3.4.1 Participants

The participants of this experiment were 58 Japanese students in two classes (28 in class A and 30 in class B) who enrolled in CALL courses at the university. The subjects were 19-22 years old engineering students. Most of them started studying English from the age of 10-13. The participants' scores on CASEC® (Computer Assessment System for English Communication³) test ranged from 560 to 850, indicating that they had different proficiency levels. This is important for the design of the study to investigate the effectiveness of PSC for different proficiency levels. Participants were divided into three groups based on their CASEC scores: beginners (560–599), pre-intermediates (600–759) and intermediates (760–850). CASEC is a standard test, which evaluates the learners on their knowledge of vocabulary and listening ability with approximately 0.96 reliability (Nogami & Hayashi, 2009).

³<http://casec.evidus.com/>

3.4.2 Material

The clips were selected from the TED website which provides authentic videos plus almost accurate transcripts under the Creative Commons license. The selection was carefully done to include videos of American speakers only in order to avoid the influence of other accents.

Moreover, caution was taken to exclude videos that contained many difficult technical terms. Besides, the average speech rate and word frequency of the videos were calculated to exclude the videos with very high speech rate or those with too many infrequent words. As Figure 3.3 shows, selected videos shared approximately comparable speech rate and word frequency. Nine videos were used in the experiment. All videos were trimmed to approximately 5-minute meaningful segments. Appendix I includes a list of all videos used in experiments throughout this study.

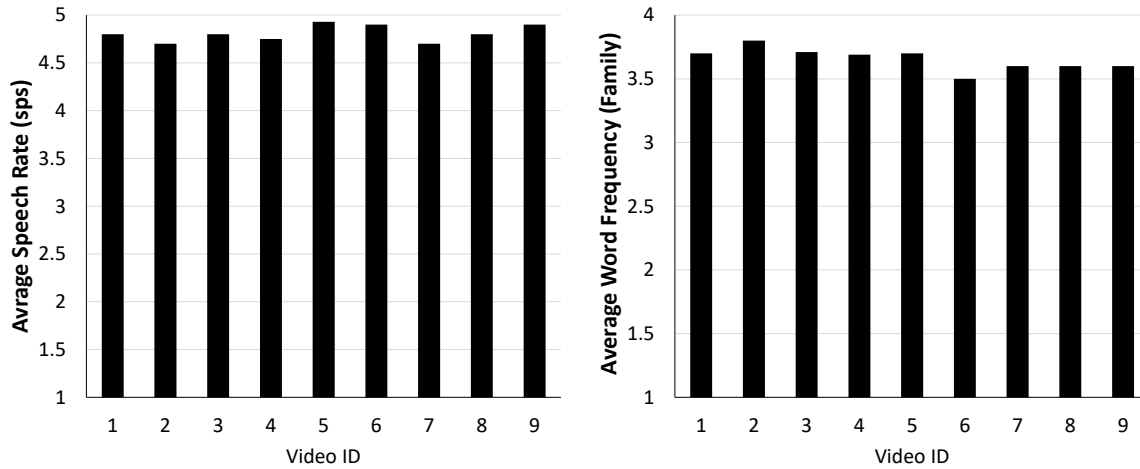


Figure 3.3: Statistics on the speech rate and word frequency of the video clips

Each video was prepared with three settings of captions: no caption (NC), full caption (FC) and PSC. PSCs were generated for different proficiency groups based on the vocabulary size and the tolerable rate of speech. The percentage of shown words in the final captions, however, did not exceed 30% of the transcript across all proficiency levels, for any of the videos.

Figure 3.4 shows statistics of PSC generated for pre-intermediates. As the figure illustrates, the percentage of shown words is different based on each acting feature

(i.e., speech rate, frequency, and specificity). In the experiment, only the PSCs generated based on all three features were used.

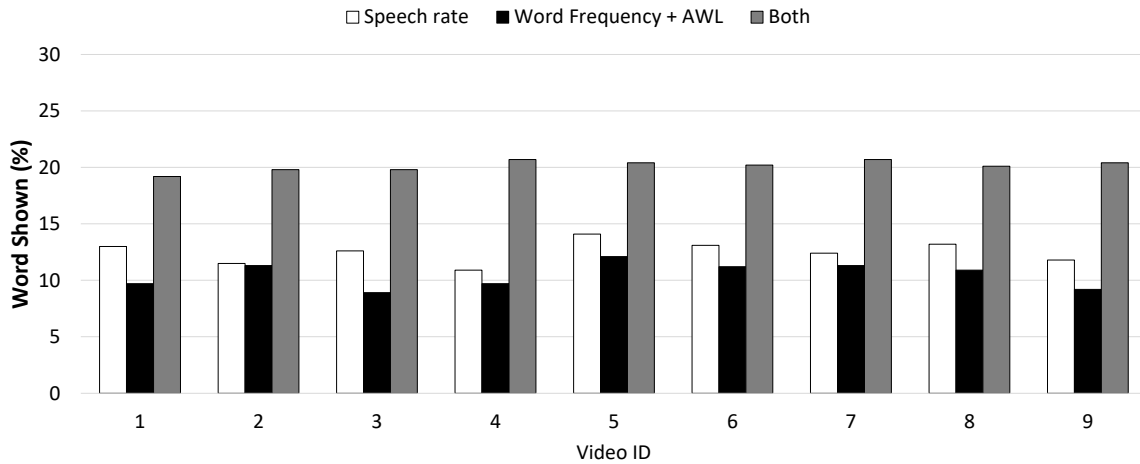


Figure 3.4: Percentage of words shown in PSC for pre-intermediate group

3.4.3 Data Collection Instruments

Vocabulary Size Test (frequency thresholding): A vocabulary size test designed by Nation and Beglar (Nation & Beglar, 2007) with the approximate Rasch reliability of 0.96 was used to evaluate the participants' vocabulary reservoir. The results of this test were used to determine the frequency threshold for our caption generator. This test consists of 140 multiple-choice questions, with 10 items from each 1000 word family. Since the caption generator uses the same 25 word family lists as its references, the result of the test is appropriate to be set as a threshold. The test was taken online.

Speech Rate Test (speech rate thresholding): The subjects were given several short conversations taken from a TOEIC[®]⁴ practice test with the approximate reliability of 0.90, reported by ETS⁵. The speed of the conversations was modified into three different levels: slow, moderate and fast. The participants were asked to answer some questions about the related conversation and report whether the audio was too slow, too fast or appropriate for them. Data on this test was only used to

⁴<http://www.toeic.or.jp/english.html>

⁵<https://www.ets.org/>

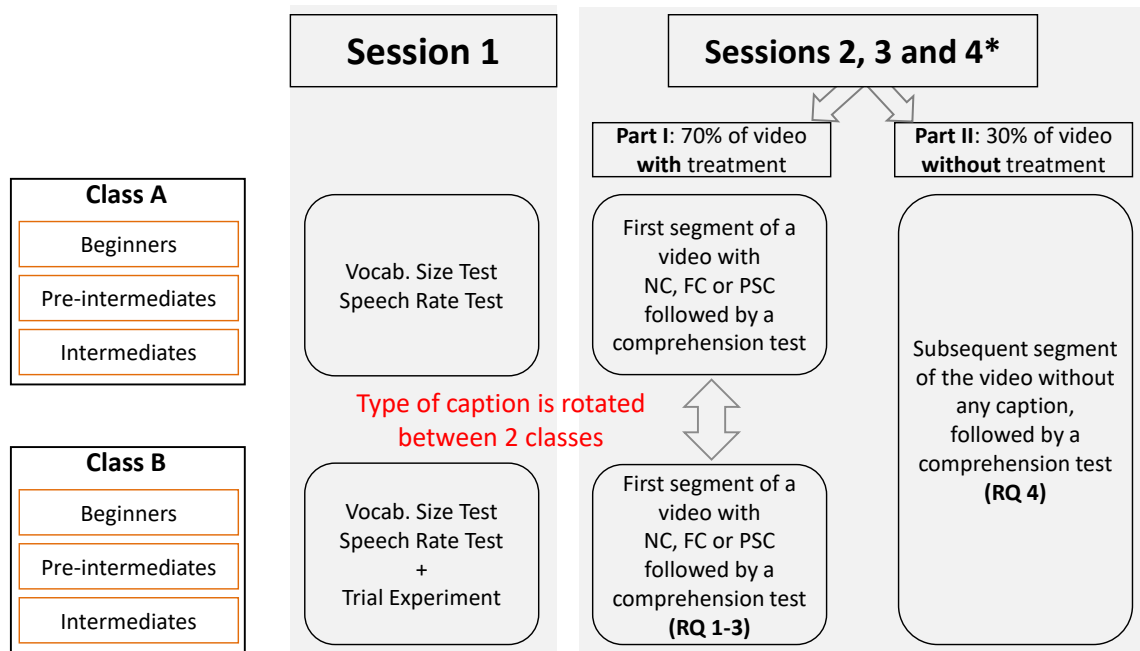
define suitable speech rate thresholds for the system, mapped to the standard rate of speech, in order to adjust PSC for each group. Hence, the scores were not counted in the final result analysis.

Comprehension Tests: Once the participants watched a video with a specific caption type, they were asked to take a comprehension test in the form of multiple-choice and cloze test. The multiple-choice questions focused both on the main idea and the specific information of the video. For cloze-test questions, the participants listened (once) to a short audio (20–30 seconds) extracted from the video they have watched and filled in the blanks in a corresponding paragraph (verbatim text of the audio). The missing words were selected from those appeared in the caption in order to constrain the choice of words and differentiate between the conditions. Appendix II includes a sample of the questions.

Questionnaire: Since the proposed PSC was used for the first time in a real language-learning environment, we conducted a 5-point Likert-scale questionnaire to solicit some feedbacks on the PSC system and to discover the opportunities for its future enhancement. The questionnaire consisted of 12 questions: two questions focusing on the participants' attitudes toward using PSC, four questions comparing their usage of PSC and FC, two questions focusing on the synchronization feature, two questions investigating the partialization aspect and the selected words in PSC, and the last two questions on the readability of the captions. The questions were reviewed by several experts to assure face or content validity. Appendix III includes the questionnaire and the participants' feedbacks.

3.4.4 Procedure

The experiment was conducted once a week in CALL classes and lasted for 4 sessions. Each experimental session took approximately 60 minutes. Figure 3.5 illustrates the experimental procedure. The first session was dedicated to the vocabulary size test and the speech rate test in order to generate appropriate PSC for each video (adjusted for each proficiency group). In this session, the students in class B had a trial experiment with PSC.



**In session 4 the participants also filled in a questionnaire.*

Figure 3.5: Experimental procedure

The experiment had two parts:

Part I evaluated the effectiveness of each captioning method (NC, FC, PSC) on learners' comprehension and provided data for research questions 1-3 by allowing a comparison of these three methods and evaluating PSC for different proficiency levels.

Part II was designed to evaluate the immediate effect of each captioning method on preparing the participants to listen without any caption. This part relates to research question 4 to investigate if PSC can help the learner comprehend the video later without any captions.

To realize this design, each video was divided into two segments: 70% from the beginning of the clip and the subsequent 30% (from the last cropped frame to the end of the clip). The longer part of the clip (70%) was used for Part I of the experiment and hence was captioned with NC, FC, and PSC, whereas the shorter part of the clip (30%) was preserved for Part II of the experiment and was used without any caption. First, the subjects did Part I and watched a video (70% long) with NC, FC or PSC, followed by a comprehension test. Next, the subjects did Part II and watched the

rest of the same video (subsequent 30%) without any caption, followed by another comprehension test.

In the final session, the subjects filled out the questionnaire about the effectiveness of PSC. The same procedure was adopted in both classes to maintain similar conditions. As shown in Figure 3.6, the types of captions were rotated between the two classes to alleviate the effect of content variability. The order of watching the videos is also rotated among the participants.

Class Proficiency	2 nd Session			3 rd Session			4 th Session		
	Video1	Video2	Video3	Video4	Video5	Video6	Video7	Video8	Video9
A beginners	FC	NC	PSC	NC	PSC	FC	PSC	FC	NC
B beginners	PSC	FC	NC	FC	NC	PSC	NC	PSC	FC
A pre-intermediates	PSC	FC	NC	FC	NC	PSC	NC	PSC	FC
B pre-intermediates	NC	PSC	FC	PSC	FC	NC	FC	NC	PSC
A intermediates	NC	PSC	FC	PSC	FC	NC	FC	NC	PSC
B intermediates	FC	NC	PSC	NC	PSC	FC	PSC	FC	NC

Figure 3.6: Experimental design for Part I (70% of videos with treatment). The remaining 30% of videos are used for Part II without any captions.

3.4.5 Results

Analysis of Variance (ANOVA) test was used to examine the effect of conditions (types of caption) on comprehension. The significance level is set to 0.05. The effect size (η_p^2) is reported and interpreted based on Cohen's rules of thumb: small ($\eta_p^2 > .01$), medium ($\eta_p^2 > .06$) and large ($\eta_p^2 > .14$). Fisher's LSD posthoc test was used to compare the effect of captioning methods. Finally, a paired-sample t-test was used to check the differences between PSC and FC conditions across three proficiency groups.

Research Question 1

The first research question investigated whether the use of caption (FC or PSC) can lead to better comprehension compared to no-caption condition (NC). This question deals with Part I of the experiment in which the participants watched 70% of videos with different captioning methods (NC, FC, PSC). Table 3.3 summarizes the mean scores on this part.

Results presented in Table 3.4 reveal that scores under FC and PSC conditions were statistically higher than the NC condition. Thus, the condition (caption type) significantly affected the comprehension scores [$F(1, 57) = 59.5, p < .001, \eta_p^2 = .51$].

Table 3.3: Descriptive statistics of comprehension scores – Part I

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

Table 3.4: Repeated-measure ANOVA on the effect of 3 types of caption - Part I

Source	Caption	df	F	p-value	η_p^2	Obs. power
Within Subject	NC, PSC, FC	1	59.54	< .001	.51	1.00
	Error	57				

Fisher's LSD post-hoc analysis (Table 3.5) revealed that the participants' scores under the PSC condition ($M = 52.89, SD = 19.39$) and the FC condition ($M = 54.25, SD = 17.33$) were significantly higher than the NC condition ($M = 35.69, SD = 14.68$). This finding provides a positive answer to the first research question.

Research Question 2

The second research question asks if PSC can be substituted for FC method. The question relates to Part I of the experiment (watching 70% of videos with different captions). As Table 3.5 presents, the difference between the scores of the PSC and

Table 3.5: LSD posthoc comparisons on scores of different conditions - Part I

Caption		Mean Difference	Std. Error	p-value	95% Confidence	
					LB	UB
PSC	NC	17.14	2.50	< .001	12.13	22.14
	FC	-1.38	2.76	.619	-6.91	4.15
FC	NC	18.52	2.38	< .001	13.75	23.28
	PSC	1.38	2.76	.619	-4.15	6.91

FC conditions was not statistically significant [$F(1, 57) = .25, p = .62, \eta_p^2 = .004$]. The findings suggest that PSC, while presenting less than 30% of the text, leads to the statistically equivalent level of comprehension as FC, which presents 100% of the text in the caption.

Research Question 3

The third research question relates to Part I of the experiment and concerned the effectiveness of PSC for different proficiency groups. Results of the paired-sample t-test in Table 3.6 revealed that within each proficiency group, the average scores of the members under the FC condition and the PSC condition are not significantly different: beginners ($t(18) = .22; p = .83$), pre-intermediates ($t(18) = -1.09; p = .23$) and intermediates ($t(19) = .23; p = .98$).

Table 3.6: Paired-sample t-test on scores of PSC vs. FC conditions across proficiencies - Part I

ProficiencyCaption		Mean	SD	SD	95% Confidence				
					LB	UB	t	df	p-value
				Mean	(2-tailed)				
Beg.	PSC-FC	.94	18.51	4.25	-7.98	9.86	.22	18	.83
Pre-Inter.	PSC-FC	-5.20	20.79	4.77	-15.22	4.82	-1.09	18	.23
Inter.	PSC-FC	.12	23.32	5.22	-10.80	11.03	.023	19	.98

While three levels of PSC were used (with different amount of text for each proficiency group), the finding indicates that PSC successfully adjusted its content to the learners' proficiency levels so that they gained a similar level of comprehension as watching videos with full captions.

Research Question 4

This question investigates the effectiveness of PSC in preparing the learner for the real-life context and regards Part II of the experiment, i.e., watching the subsequent 30% of videos without any captions after the previous segment (70%) was seen with a caption. As shown in Table 3.7, the best scores were gained when the learners first watched videos with PSC ($M = 56.59, SD = 17.34$) compared to NC ($M = 40.12, SD = 17.39$) and FC ($M = 42.65, SD = 13.37$) conditions.

Table 3.7: Descriptive statistics of comprehension scores – Part II

Caption	Proficiency Level	N	Mean	SD
NC	Beginner	19	32.95	16.04
	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

Table 3.8 shows that this difference is statistically significant based on a Fisher's LSD test. This means that the learners' comprehension of an uncaptioned video was significantly higher when they had watched the previous segment with PSC than with full captions or no captions [$F(2, 118) = 20.5, p < .05, \eta_p^2 = .26$].

Table 3.8: LSD posthoc comparisons on scores of different conditions – Part II

Caption		Mean Difference	Std. Error	p-value	95% Confidence	
					LB	UB
PSC	NC	16.47	2.95	< .001	10.56	22.38
	FC	13.94	2.63	< .001	8.66	19.21

Although this is a short-term enhancement partly because of adaptation to the video, this finding is still promising. We also compared the scores of participants when they watched the videos with no-caption in Part I ($M = 35.69, SD = 14.68$) and the remainder of the same videos without a caption in Part II ($M = 40.12, SD = 17.39$). Under these two similar conditions, there was no statistically significant difference between the scores [$t(57) = -1.296; p = 0.20$]. Consequently, the results of Part II of our experiment are affected by the treatment (caption) used in Part I.

3.5 Discussions

3.5.1 Overall Effect of Different Captioning Methods

The first research question aimed to compare the effect of captioning conditions (FC or PSC) with no-captioning condition (NC). The quantitative results on this question corroborate the findings of previous studies and suggest that the presence of captions significantly aids listening comprehension (Markham & Peter, 2003; Danan, 2004; Taylor, 2005; Winke et al., 2010; Montero Perez et al., 2013). This is confirmed regardless of whether PSC or FC was used and is inline with the dual coding theory of Paivio (Paivio, 1990). When reading caption forms part of watching a video, learners can benefit from multiple input modalities. In this regard, Bird and Williams (Bird & Williams, 2002) emphasized that the use of text and sound results in better recognition memory. However, even in the FC condition, the participants' scores on the tests are below 60%, which indicates the difficulty in understanding the video content. TED talks are apparently difficult to comprehend for most of the non-native speakers and are more appropriate to be used for advanced learners whereas the participants in this study were beginner to intermediate level.

3.5.2 Effectiveness of PSC Compared to FC

In the context of research question 2 (comparing the effectiveness of PSC with FC) we found that the test scores under the FC and PSC conditions were not statistically different. This may indicate that the two methods can be used interchangeably.

However, PSC by presenting fewer numbers of words (less than 30%) encourages more listening than reading, compared to FC, which has been criticized for promoting reading over listening (Pujolà, 2002; King, 2002). But how can PSC be as effective as FC for comprehension, while it only shows a small amount of the text? The following two reasons are considered.

Cognitive load reduction: While PSC improves comprehension following the assumption of the dual coding theory, the effectiveness of this method may also be explained by the cognitive load phenomenon and limitation of working memory that occurs in the case of excessive information (Baddeley, 1992; Sweller, 1994). With smaller amounts of text in the visual channel, learners are less likely to encounter overload to multi-channel processing and more likely to achieve fuller comprehension of the information coming through the auditory channel (Guillory, 1998, p. 97). Thus, for effective learning, multimedia instruction should minimize any unnecessary cognitive load (Mayer & Moreno, 2003). In this view, PSC aims to foster listening by providing minimal assistance and hence acting as a trade-off between the dual coding theory and the cognitive overload. Moreover, its word-level synchronization facilitates text-tracking and reduces the amount of scanning for text-to-speech mapping.

Appropriate selection criteria: An explanation to strike a balance between these theories (dual coding and cognitive load) may lie in the appropriate selection criteria adopted by PSC – with appropriate selection of caption text, learners can gain better comprehension without excessive cognitive load. As such, the results of this study are mainly based on the three features (speech rate, word frequency, and specificity), which realized PSC and formed its selection criteria. Rationale of these selection criteria can be found in (Griffiths, 1992; Zhao, 1997; Buck, 2001) on speech rate and in (Nissan et al., 1995; Webb, 2010; Révész & Brunfaut, 2013) on word frequency and specificity.

3.5.3 Effectiveness of PSC across Proficiency Levels

The results revealed that the scores gained under the FC and PSC conditions were not statistically different within each proficiency group. In other words, the subjects

in each proficiency group could gain a similar level of comprehension under the PSC condition (with less than 30% of the text shown) and the FC condition (where 100% of the text is shown). PSC tries to provide enough assistance by adopting the amount of text in the caption to the learners' needs and proficiency levels (three levels of PSC, adjusted for each proficiency group, were used in our experiment. See Figure 3.4 as an example for pre-intermediates).

Studies that explored the use of captions for learners with different proficiency levels have reported mixed results. The results in this study are in line with Markham's (Markham, 1989) and Guillory's (Guillory, 1998) findings on the effectiveness of captions (full and keyword respectively) for low-proficiency learners and therefore contradicts Taylor (Taylor, 2005) who reported that first-year students scored better in NC condition than FC condition. A possible explanation may lie in the noteworthy feature of PSC that evaluates the learner's level $[i]$, adjusts the amount of textual information to match that level and hence provides comprehensible input by changing $[i+n]$ to $[i+1]$ for different learners. However, due to the limited number of participants in each group, the results must be interpreted with caution and cannot be extrapolated.

3.5.4 Effectiveness of PSC to Prepare Learners for Real-life Situations

The results of the final research question confirmed that after watching videos with PSC, the participants had a better performance in comprehending a video without any assistance.

Reliance on captions is an individual matter that cannot be universally applied. Consequently, before considering the addition or removal of captions, instructors should have evidence to the degree at which the individual learners rely on captioning support for comprehension (Leveridge & Yang, 2013, p. 211). As explained earlier, PSC provides a different amount of assistance to learners at different levels. More importantly, it allows the learner to adjust the amount of textual clues by changing the thresholds. This method acts as a source of scaffold and allows the learners to prepare for the NC condition at their own pace. It is anticipated that following

the same strategy in the long term, the learners can entirely rely on their listening skills for comprehension. Given the nature of listening skills, however, a long-term experiment is required to confirm this effect.

One limitation of this study regards Part II of the experiment on the immediate post-effect of captioning methods. The result on this part may reflect a temporary enhancement and should be confirmed by more experiment. Moreover, our participants were beginners to intermediate Japanese learners of English. Thus the results cannot be generalized to other participants with different L1 or different proficiency levels. Administrating a longitudinal study and involving more participants can form more solid analysis.

3.6 Conclusion

This chapter introduced a smart type of captions that allows the use of limited textual clues and promotes listening to the audio in order to comprehend the material. The proposed method, PSC, is based upon three factors that contribute to listening difficulty: speech rate, word frequency, and specificity. Using these features the system generates a caption that tries to deal with the limitation of the previous methods. With the ASR technology, the system synchronizes the text-to-speech, which emulates the speech flow, facilitates text-to-speech mapping and avoids the salient appearance of the words on the screen. The system assesses the tolerable speech rate and the vocabulary size of the learner to adjust the caption to the proficiency level of the learner.

Evaluated in two CALL classes, the results of the experiments showed that learners scored better when using PSC compared to the no-caption condition. PSC resulted in comparable comprehension as the full-caption condition. Furthermore, the learners gained significantly higher scores on a new segment of the video without any caption when they had watched the video with PSC first. The finding highlights the positive effect of PSC on preparing learners for listening in simulated real-life situations, where they do not have any means of assistance such as captions or speed



controllers. The results also indicate that the method can assist learners to obtain adequate comprehension of videos by presenting less than 30% of the transcript. This method is expected to be effective particularly for Japanese students who heavily rely on caption text in order to comprehend the content of the video.

It must be noted, however, that we need to enhance this system to encompass other features that affect L2 listening comprehension. This is crucial for increasing the accuracy of word selection in PSC and hence providing better assistance to the learners. A wide range of features related to speech and lexical aspects can be considered in the following chapters.

Chapter 4

ASR Errors to Predict L2 Listening Difficulties

This chapter proposes a new paradigm for detecting difficulties in speech for L2 listeners. This is the first study, which uses an external element (ASR system) as a model for predicting L2 learners' listening difficulties. ASR errors are compared with L2 listeners' transcription mistakes. 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, the comparison between ASR errors and L2 learners' recognition errors, the term we coined as ASR-L2SR, has not been closely examined. This research is motivated by the challenges associated with the detection of perceptual ambiguities in the speech for L2 listeners, which requires a revelatory source to shed light on the intrinsic speech difficulties in different listening materials. Accordingly, the objectives of this chapter are to perform such comparison and determine whether ASR errors can highlight challenging speech segments that signal recognition difficulties for L2 learners, hence provide insights for PSC enhancement.

4.1 ASR versus Human Speech Recognition

While human listeners have little difficulties in dealing with recognition of spoken language in acoustically challenging situations, ASR often lacks the same robustness

that is achieved by the auditory system (Meyer et al., 2011). This observation has motivated research that 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 & Cutler, 2001; Scharenborg et al., 2003; Meyer et al., 2006; Scharenborg, 2007; Vasilescu et al., 2012).

The subjects of these studies are either a native speaker of the target language or non-native speakers with no knowledge of the target language (e.g., Japanese with no knowledge of French tested with French audio, which includes words with the maximum phonetic similarity between the two languages). These studies have investigated the robustness of ASR systems against extrinsic variability (especially when arising from additive noise) along with robustness against intrinsic variations of speech (i.e., the natural variability that is produced by the speaker).

Most of these studies have emphasized the importance of conducting fair HSR-ASR comparisons by restricting the influence of background information, using logatoms/pseudowords (Meyer et al., 2006). This is especially important to consider when evaluating the system against native speakers, to maintain a comparable situation. Otherwise, native speakers' recognition will in most cases surpass the ASR performance. The use of pseudo words is not required in the experiment with non-native speakers, who are not familiar with the target words. However, to make a credible ASR-HSR comparison, words should be chosen carefully to include instances that have the shortest phonetic distance between the human's native language and the target language.

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). Dialect and accent were considered as other factors that significantly affect HSR and ASR recognition scores. Furthermore, speaking rate, effort, and style as well as the choice of speaker, contribute considerably to the variance of recognition scores in HSR and ASR (Goldwater et al., 2010).

Intrinsic variations also have a significant effect on resynthesized speech. In this condition, however, the choice of speaker seems to have a more dominant effect than speaking rate, effort and style. In the case of ASR, the contribution of changes in rate, effort and style are more important than in HSR with resynthesized speech, which is consistent with the high sensitivity of ASR against such kinds of variations (Meyer et al., 2011).

While these factors are remarkably more influential on ASR systems (Benzeghiba et al., 2007), restricting the contextual information can also affect human recognition (Kitaoka et al., 2014). Another main difference between HSR and ASR is the strategy that human listeners employ to detect speech components in a signal. For example, Miller and Licklider (Miller & Licklider, 1950) performed an experiment with interrupted (gated) speech and found that word recognition scores are only slightly degraded when the interruptions occur at modulation frequencies between 10 and 100 Hz. The authors assumed that a high intelligibility could be obtained as long as listeners get a glimpse at each phoneme of the presented word.

Through these studies, researchers attempt to compare ASR errors with human (native or non-native) misrecognition in order to unfold solutions for improving the ASR systems (Shen et al., 2008). However, ASR errors were not compared with second language learners' misrecognition. Inspired by ASR-HSR comparisons, this study strives to detect the similarities or differences between ASR and L2SR in order to identify L2 listeners' difficulties.

4.2 Reviews on ASR versus L2 Speech Recognition

In ASR-HSR studies, ASR errors are accounted 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., 2012). 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. 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.

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 error in lexical level. 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 (Shinozaki & Furui, 2001; Toutanova et al., 2003). 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., 2012).

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 make L2 listening difficult for the learners. Many factors account for L2 listening difficulties, some of which have already been explained and covered by PSC features. We use such features to conduct ASR-L2SR analysis:

Lexical Factors: In the lexical level, the frequency of the words 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). Similarly, findings on the ASR error analysis emphasize the importance of this factor in the performance of the system (Shinozaki & Furui, 2001). Another factor is the word length, which has also been found to be a useful predictor of higher error rates in ASR systems (Shi-

nozaki & 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 dominate prepositions (Field, 2008).

Acoustic, Speech and Perceptual Factors: Among these factors, speech rate, whether too fast or too slow, is the main source of difficulty for many L2 learners (Griffiths, 1992). This factor is also influential for ASR systems, which are largely affected by variations in the speech rate (Fosler-Lussier & Morgan, 1999; Shinozaki & 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 differences of speakers, 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). Furthermore, the gender of the speaker may have some effect on the ASR performance, as it is reported that male speakers are more problematic for ASR systems than females (Adda-Decker & Lamel, 2005). A similar result was reported for L2 listening recognition; males have generally faster articulatory rates than females, which makes their speech difficult to recognize for L2 learners (Quené, 2007). Finally, the ambiguities in speech such as the occurrence of homophones, assimilation and ambiguous word boundaries are found to severely impede L2 listening comprehension for language learners (Broersma, 2012) as well as the ASR performance (Forsberg, 2003).

Table 4.1 summarizes a comparative analysis we performed through an extensive investigation of background studies to compare the factors that affect the performance of ASR and those that influence L2 listening.

Table 4.1: ASR-L2SR Comparison

	ASR Difficulties	L2 Listening Difficulties
Lexical Factors	Infrequent words are more likely to be misrecognized. (Shinozaki & Furui, 2001)	The occurrence of infrequent words in speech is correlated to complexity (Bloomfield et al., 2010).
	Word length has also been found to be a useful predictor of higher error rates. (Shinozaki & Furui, 2001)	The length of a word has a strong effect on its recognition. (Laufer, 1990)
	Open class words (N. and V.) cause less errors compared to the closed class (Prep., articles). (Goldwater et al., 2010)	Recognition of content words is easier than function words and nouns predominate over predicates/verbs. (Gentner, 1982; Nitta et al., 2010a)
Acoustic, Speech and Perceptual Factors	Fast or very slow speech rate raises the ASR errors (Fosler-Lussier & Morgan, 1999; Shinozaki & Furui, 2001).	Whether it is too fast or too slow, speech rate can deteriorate L2 listening. (Griffiths, 1992)
	Co-articulation, pronunciation, speaking style, disfluencies, accent, age, physiology, and emotions of the speaker lead to the ASR difficulties. (Benzeghiba et al., 2007)	Pronunciation can be unclear due to assimilation, reduction, etc. Stress, intonation patterns, and accent affect L1 and L2 listening. (Osada, 2004; Bloomfield et al., 2010)
	Male speakers are more problematic for ASR systems than female speakers (Adda-Decker & Lamel, 2005).	Male speakers have faster articulatory rates than females. (Quené, 2007)
	Ambiguity in the speech such as the occurrence of homophones or ambiguous word boundaries are the factors that lead to recognition difficulties for the ASR systems. (Forsberg, 2003)	Phonological neighbors and words with identical pronunciation make L2 recognition hard (Broersma, 2012). Assimilation, reduction, etc. leads to breached boundaries and attenuate L2 listening (Field, 2003).

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 order to further improve the performance of the PSC 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 L2 listening difficulties is complex. This study addresses this issue by investigating the factors causing 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. In this view, the use of ASR errors as an indicative of listening difficulties can provide important insights for discovering useful features for PSC improvement.

4.3 ASR Error Analysis

Recent advances in ASR systems such as the use of deep neural networks has led to a significant increase in the accuracy of these systems and realized comparable performance to the native speakers with very low error rate ($\sim 10\%$). As a result, the ASR errors generated by such systems are very limited and do not contain enough instances to enable effective ASR-L2SR comparison. The version of the ASR system we used in this study provided us with a reasonable amount of errors while maintaining an acceptable performance ($\sim 20\%$).

It is also possible to generate a list of the N most likely hypotheses by the ASR system (Jurafsky & Martin, 2014). These hypotheses are usually sorted based on high-level knowledge sources. The hypotheses in the N -best list can provide more and diverse type of errors and allow for exploring additional ASR erroneous cases. In practice, we found that our ASR system's 1-best output could still provide us with a reasonable amount of errors to perform the ASR-L2SR analysis with the purpose of discovering useful features to enhance the baseline PSC system.

4.3.1 ASR Error Statistics

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. As presented in the Table 4.2, 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 belong to the substitution category:

Table 4.2: ASR Error Analysis on 70 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%)

- **Substitution Errors:** Instances where ASR transcript and ground truth are different in one or more words (17.53% of all words).
- **Deletion Errors:** Instances that ASR failed to transcribe, but are present in the ground truth (1.81%).
- **Insertion Errors:** Instances that ASR transcriptions is not present in the ground truth. (2.00%).

4.3.2 ASR Error Trends

To begin the analysis, ASR errors are examined to discover the underlying trends. There are many factors accounted for the emergence of ASR errors. Shinozaki and Furui (Shinozaki & Furui, 2001), using different corpora, reported that word recognition error tends to be higher if the word has a small number of phonemes, spoken fast or too slow, and observed less frequently in the language-model training corpus.

Based on their findings, features such as speech rate, word frequency, and word length are good predictors of ASR errors. Since the first two features are also used in the baseline PSC system for detecting difficult words, these features seem to be prudent choices for investigating ASR errors in this study. Accordingly, our analysis is performed based on PSC’s baseline features (speech rate and word frequency) together with the word length.

The speech rate of the ASR errors was calculated in SPS and its trend was explored in four bins: slow (~ 3.83 SPS), moderate ($3.83 \sim 5.33$ SPS), fast ($5.33 \sim 8$ SPS) and too fast ($8 \sim$ SPS) based on the standard rates of speech (Tauroza & Allison, 1990).

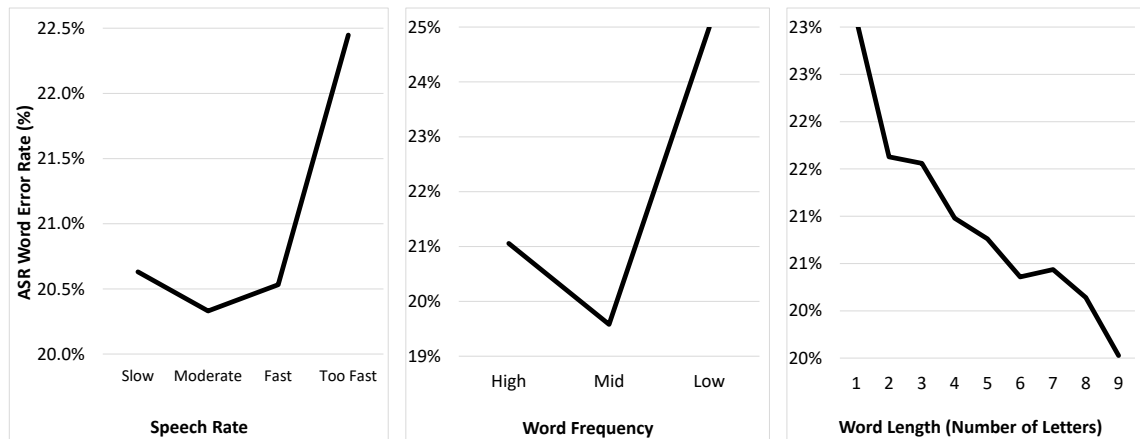


Figure 4.1: Trend Analysis on ASR Errors

Figure 4.1(left) illustrates how the ASR error rate increases when the speech rate rises. The trend is in line with those reported in L2 studies (Nitta et al., 2010b). With increasing speech rate, L2 learners are more prone to make listening mistakes (Rost, 2005). 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. (Nitta et al., 2010b) reported that at 4 SPS, L2 learners missed or mistook 4.2% of the words, of which 2.7% was function words and 1.5% was 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. Furthermore, L2 studies have shown that misrecognition increases among L2 listeners when listening to audio with too slow speech rate (Griffiths, 1992), which is also the case in our ASR errors.

Similar trend analysis is performed on the ASR errors considering the word frequency feature. The frequency of words in ASR errors is calculated by referring to Nation's family lists (Nation & Webb, 2011) along with BNC and COCA. The frequency is partitioned into 3 bins - low frequency (~ 3000 word families), mid-frequency ($3000 \sim 6000$ word families), and high-frequency (above 6000 word families) according to (Schmitt & Schmitt, 2014). Figure 4.1(mid) shows 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 a short length and pronunciation variations.

The errors were also investigated regarding word length feature. Although the length of a word is not used in the baseline PSC selection criteria, this feature has been frequently investigated in the studies focusing on ASR error analysis (Goldwater et al., 2010) as a good predictor of ASR errors. In addition, the feature is also considered in studies focusing on L2SR difficulties (Laufer, 1990). Figure 4.1(right) shows that ASR error rate decreases as the word length increases. The longer the words are, the better the ASR can recognize them. Longer words have a longer duration, which makes it easier for ASR system to identify them. This finding is similar to the results of ASR error analysis (Fosler-Lussier & Morgan, 1999). These results can be explained by the findings of L2 studies reporting that learners pay more attention to longer words in speech and strive for recognizing them accurately (Field, 2008). Moreover, longer words are often articulated more carefully or even hyper-articulated (Bell et al., 2003), which in turn make it easier for ASR systems to recognize them, and also attract learners' attention when listening to the speaker (Field, 2003).

Findings of the analysis revealed that similar trends are discovered on ASR errors and L2 listeners' misrecognition. Moreover, the trends we extracted are in line with those reported in previous studies on ASR error analysis using other ASR systems.

4.4 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 the baseline 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. In this system, these errors are specifically compared with PSC choices to find the overlaps and seek further enhancement. To

investigate any plausible similarities, the baseline PSC was generated for all 70 TED videos, controlling for high speech rate, low frequency, and specific or academic words. The selected words by the baseline PSC were then compared against ASR errors to find the degree of overlap.

Table 4.3 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, yet these should be removed. This finding highlights the importance of investigating these categories to discover the underlying features. These mismatches are automatically extracted and further analyzed to discover the challenging speech segments that are not yet handled by the PSC system and to detect easy cases, which are presented in the baseline PSC, indicating the system flaw.

Table 4.3: ASR Cases versus Baseline PSC Comparison (70 TED Talks)

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

4.4.1 Analysis on ASR Error and PSC Shown Cases

First of all, the ASR errors were analyzed by taking PSC's shown cases into account, i.e., frequency, speech rate, and specificity features. Figure 4.2 depicts the distribution of the mutual cases between the ASR errors and PSC's selected words.

As the figure suggests, speech rate is the primary factor that selects the words for PSC and is also the major factor that leads to the emergence of the ASR errors (58%). This is in line with our results of the ASR error trend analysis, which indicated that with the elevation of speech rate, word error rate (WER) also increases. Thus the

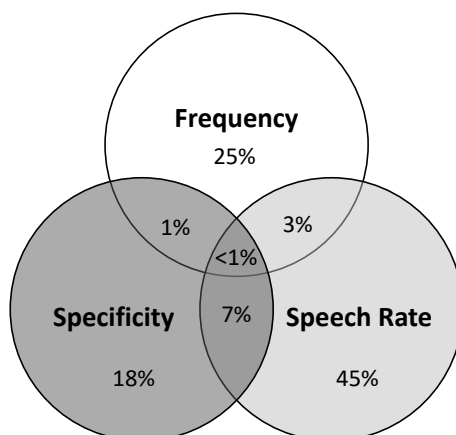


Figure 4.2: Feature analysis in ASR Error and PSC shown cases

factor addresses difficulty both in bringing difficult cases into PSC and for causing ASR errors.

The frequency factor shows 20% of overlap between the PSC shown words and the ASR errors. This finding indicates that the inclusion of infrequent words in speech will degrade ASR performance and addresses recognition difficulty for language learners, hence, is used to detect difficult words for PSC. Finally, specific words are by default set to be always shown in the PSC system, but only a small number of these words cause ASR errors (6%). The fact that our ASR system is trained on TED corpus explains for the correct detection of such cases, which are mostly included in the dictionary of the system.

4.4.2 Analysis on ASR Correct and PSC Shown Cases

While it is assumed that ASR errors can indicate problematic speech segments for L2 listeners, ASR correct cases can specify easy items, which may not be necessarily included in PSC. A thorough analysis on ASR correct and PSC shown cases will identify the reasons for PSC's decision to include these words. To conduct this analysis, the ASR errors were analyzed by taking PSC's features into account, i.e., frequency, speech rate and specificity features. Similar to PSC, the speech rate of the ASR errors were calculated in syllables per second, the frequency was estimated based on the corpus of contemporary American English - COCA ([Gardner & Davies](#),

2013), and the specific words were detected by referring to the Academic Word List (Coxhead, 2000) and academic corpus of COCA.

As Figure 4.3 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 for L2 learners (e.g., “every”, “who”, etc.). In fact, these words were easy to recognize using the contextual information despite the fast speech rate.

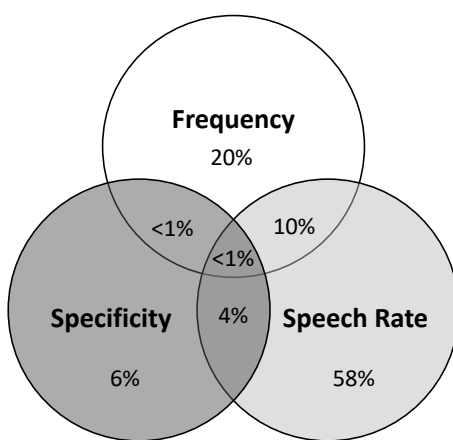


Figure 4.3: Feature analysis in ASR correct and PSC shown cases

It can be suggested that ASR correct cases can provide insightful clues on refining the speech rate threshold. In this view, by defining a secondary threshold for the speech rate feature on ASR correct cases we can apply stricter margins to show these instances and improve the word choices in PSC.

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, and few instances of the words shown in the PSC based on the frequency feature seem to be unnecessary. For instance, words such as “*dystopia*”, “*piggybacking*”, “*pandemic*”, “*larceny*”, “*abyss*” could be correctly transcribed by the ASR, but are infrequent to many L2 listeners and hence likely to be unknown. Findings of this analysis imply that the frequency feature is very effective and does not need any alteration.

The third feature is word specificity, which brings the academic words into 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. While these cases are simply categorized as specific words and shown in the baseline PSC system to foster listening, many of them are often so common that can hardly be labeled as specificity. To address this issue, a similar measurement can be taken as discussed for the speech rate feature, i.e., a secondary threshold can be defined for the specific words. This threshold, which takes into account the label of the ASR output, would be activated in case ASR correctly recognized a specific word to inhibit the word from appearing in the caption.

4.4.3 Analysis on ASR Error and PSC Hidden Cases

The next comparison deals with analysis on ASR erroneous and 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:

Homophones: words with the 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 to make a distinction (Field, 2003; Weber & Cutler, 2004).

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 L2 learners according to L2 studies (Weber & Cutler, 2004).

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 often misrecognize these cases and misunderstand the meaning (Field, 2003).

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.

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 easy to perceive if the contextual information is taken into account. While ASR systems generate plenty of such errors, these cases do not severely hinder comprehension and can be easily disambiguated.

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 an important case of misrecognition for L2 learners.

Noun Inflections: 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.

Interjections: words or expressions used to signify the speaker’s strong feeling, spontaneous emotion or reaction. They include 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 when it comes to listening.

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*”, “*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 strictly hinder L2 listening comprehension.

Stop List: cases, which are usually the most common words in a language and include short function words such as prepositions (e.g., “*at*”, “*on*”, “*up*”). This category also includes “*to be*” verbs, “*WH*” questions, etc.

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 apparently not so important for comprehension. 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

of agreement. The results of this annotation (percentage of usefulness) together with the occurrence ratio of each category are shown in Table 4.4.

Table 4.4: Distribution of patterns and their usefulness in ASR error and 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%

The annotation results (Table 4.4) 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 and PSC hidden category. Minimal pairs have the largest ratio of usefulness with 86%, followed by homophones (82%), negatives (72%), and breached boundaries (64%). Interestingly, these cases were identified to be particularly challenging for language learners according to L2 studies (Field, 2003; Weber & Cutler, 2004).

On the other hand, cases such as verb inflections, interjections and determiners lack the convincing amount of useful cases to be embedded into the PSC and their inclusion would contaminate the caption with many trivial words. Table 4.4 shows that in spite of involving more than 31% of useful words, the ratio of useful words to all words in each of these categories is relatively low. As a result, we regard them as impotent factors that are not useful to be incorporated into PSC.

4.5 Experimental Evaluation of Additional Features

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

4.5.1 Participants

The participants of this experiment 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 were considered as intermediates.

The participants attended this experiment as a part-time job, thus payment was considered as an incentive to encourage them to carefully do the tasks during the experiment. Prior to the experiment, all participants were informed about the procedure using a sample test. Almost all the students were familiar with the TED videos and had experiences of watching some TED talks before participating in the experiment. All participants listened with headphones and could use a pencil and paper in case they needed to take notes during the experiment.

4.5.2 Material

We selected 20 TED talks, opting for the 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 problems 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).

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 in the performance of L2 listeners on 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.

4.5.3 Procedure

The participants were asked to listen to these pieces of continuous 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. This procedure is demonstrated in Figure 4.4.

Through this experiment we aimed to answer several research questions regarding easy and difficult speech segments:

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?

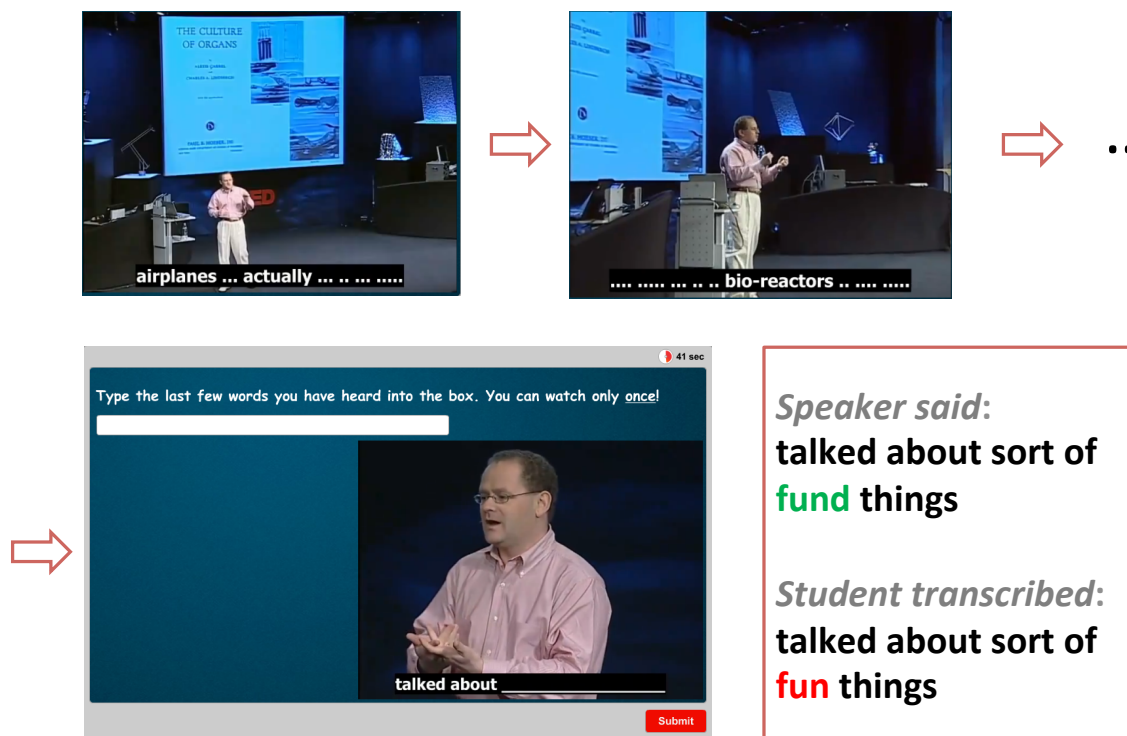


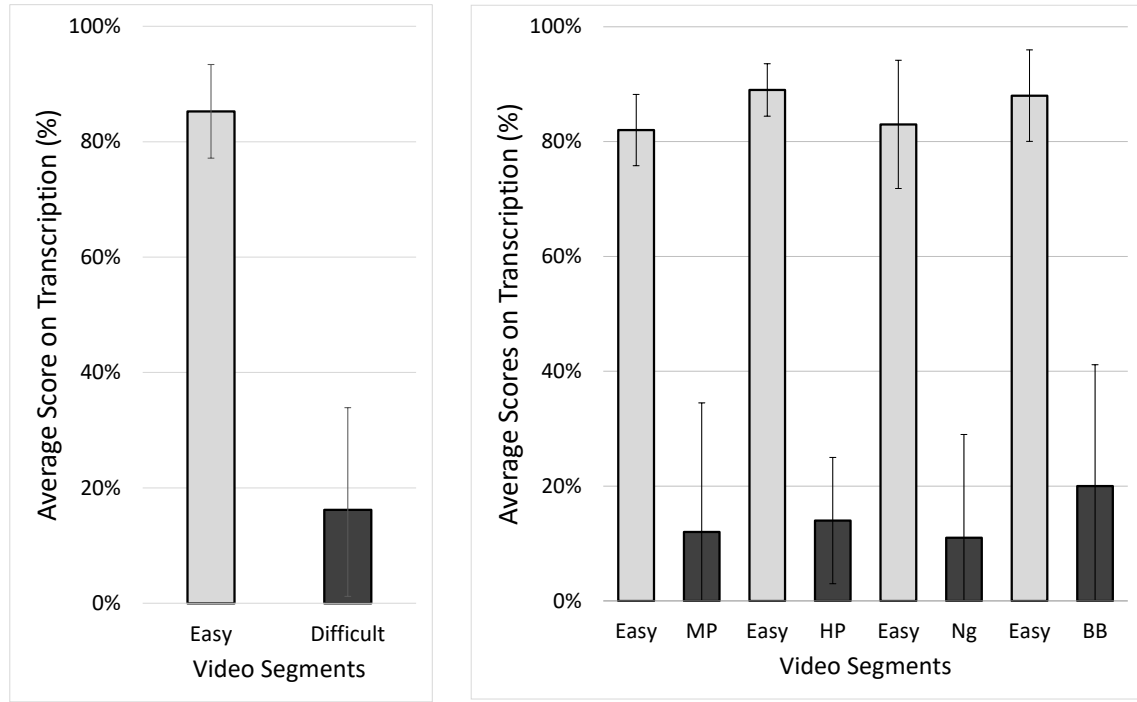
Figure 4.4: Experimental procedure of a transcription task with irregular pauses

4.5.4 Results

Figure 4.5(a) shows the statistics of participants' scores on transcribing (i) easy segments of the videos i.e., words correctly transcribed by ASR and (ii) difficult segments of the videos i.e., the words including ASR errors. As the figure shows, learners' scores on transcribing the easy segments ($M = 0.85$; $SD = 0.08$) are significantly higher than their scores on difficult segments ($M = 0.16$; $SD = 0.18$).

Figure 4.5(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 in 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 the extracted ASR errors in detecting problematic speech segments for L2 listeners.



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

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

Figure 4.5: Transcription scores on segments of ASR errors vs. ASR correct

4.6 Conclusion

This chapter investigated the use of ASR errors in detecting challenging speech segments of TED talks and improving the word selection criteria in PSC. The viability of using the ASR system as a model that can epitomize L2 listeners' problems in the perception of TED talks was explored. A root-cause analysis was conducted on the ASR errors to better understand the underlying features that make recognition difficult for such systems and they were compared with L2 listening influential factors. Such research has many pedagogical implications as it can provide the teachers with

useful data on the difficulties of authentic audio/visual material for L2 listeners and assist learners in overcoming their difficulties while listening.

Listening difficulties can be attributed to various factors such as fast speech rate, infrequent words, co-articulation, hesitation, and accent, etc. To collect such data, we focused on the errors made by an ASR system when generating transcripts for TED videos. Thus, ASR generated errors were stored and compared against the correct human-annotated transcript of the audio and the error trends based on PSC criteria were investigated. The results of this investigation provided us with some evidence that ASR errors, similar to the language learners, follow the same trends on speech rate, frequency, and length features.

Next, the ASR errors were further examined and the underlying factors that induced such errors were investigated through a root-cause analysis. While the root-causes of some of these errors could not be identified clearly, others closely indicated the challenging nature of the respective speech segments and were classified into different categories. Through the annotation of such cases, it was found that several categories in the ASR errors suggest the difficulties for L2 listeners and can be useful to be incorporated into the PSC system. These categories included homophones, minimal pairs, negative forms, and breached boundaries.

The discovered patterns were tested in the language learning environment to ensure that they cause difficulties for L2 listeners as they impede ASR performance. An experiment with L2 listeners confirmed the feasibility of using these ASR errors to predict L2 speech recognition difficulties. This finding provides means for future advances of the PSC system by exploiting ASR clues to optimize the choice of words. The next chapter elaborates on the enhancements of the baseline PSC system based on the findings derived from ASR error analysis.

Chapter 5

Enhanced Partial and Synchronized Caption

The baseline PSC's three features (i.e., speech rate, word frequency, and specificity) explain many of the L2 listening problems and account for the main causes of listening difficulties ([Griffiths, 1992](#); [Révész & Brunfaut, 2013](#)). However, not all listening challenges could be explained by these features. As a result, the selected words sometimes include easy words and occasionally exclude difficult words or phrases, which highlights the importance of exploring other features for word selection in this system. Nevertheless, the relationship between different factors and their significance in listening difficulties is complex. This calls for investigating another approach that can shed light on the difficulties of the speech specifically for L2 learners.

For instance, many L2 listeners have difficulty with lexical segmentation and they frequently fail in locating the right boundaries between the words in the connected speech ([Field, 2008](#)). Such difficulties have large effect on L2 listening impair but are complicated to detect without analyzing the nature of the speech, hence are missing in the baseline PSC's selected words.

To decipher listening challenges, in the previous chapter, the use of ASR errors was investigated 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 challenges of ASR systems is similar to L2

listeners when it comes to the transcription task. Thus, ASR errors in transcribing speech may derive from the same sources that lead to L2 misrecognition. As such, these errors can provide useful clues for the enhancement of PSC.

This chapter explains how useful errors are incorporated into the baseline PSC system to provide better assistance. It also describes how the amount of shown words in the baseline and enhanced system were maintained by removing easy cases from the baseline system, while the choice of words in the enhanced version was improved by embedding more of difficult cases. To attest the improvement, 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.

5.1 Using ASR Clues to Enhance Baseline PSC

Findings from the experiment in Chapter 4 showed the usefulness of four categories of ASR errors and indicated that these cases can be embedded into the PSC system to scaffold the learners on difficult speech segments. Meanwhile, findings revealed that some instances of shown words in PSC, which are correctly transcribed by the ASR system, are basically too easy and can be removed from the baseline PSC by defining a secondary threshold. Accordingly, we enhanced the baseline PSC system to provide better assistance for L2 listeners. The main idea is to view an ASR system as a model for the 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 the 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. In doing so, the label of the ASR output, correct or error, is assigned to each word and used as a clue to enhance the choice of words in PSC. Meanwhile, the aligned 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 in the baseline system. 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. Finally, all of the extracted features are integrated to decide about the inclusion of the word into the enhanced PSC system.

5.1.1 Improving Baseline PSC with ASR Correct Cases

Based on the analysis in Chapter 4, it was found that speech rate is the main reason to bring easy words into the baseline PSC, and many of the words brought by the speech rate factor could be correctly transcribed by the ASR system, indicating that these words were not too difficult. Examples include the words such as “one”, “every”, “open”, “look”, etc., which have high frequency, but have high speech rate because of the short length and can be simply excluded from PSC without causing a barrier for L2 listeners. Thus, the speech rate threshold is refined on ASR correct cases to prevent the inclusion of easy cases in PSC.

While a default threshold is set for PSC based on the user’s tolerance and literature standards θ_{sr} , a secondary threshold is 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, $\theta_{sr}^{ASRcor(w_i)=1} = \theta_{sr} + \Delta_{sr}$. Accordingly, $ASRcor(w_i)$ is a binary flag indicating the correctness of ASR output for word w_i according to the forced-alignment unit ($ASRcor(w_i) = 0$ signals the ASR error status), and Δ_{sr} is an added margin for ASR correct cases.

In addition, while specific words are always shown in PSC, many of them are not infrequent. For example, words such as “positive”, “science” and “research” are categorized as academic terms. However, these words are very frequent and the majority of L2 listeners should have no problem with them. Likewise, highly frequent proper nouns (e.g., “China” and “Obama”) could be simply omitted from or repeated less in PSC, given that our ASR system could also correctly transcribe these words.

These findings suggested that a secondary threshold should also be considered for the frequency of specific words when deciding on their inclusion in PSC. Therefore we introduce 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 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 (i) PSC's speech rate threshold should be tuned based on ASR clues, (ii) the word frequency feature should be prioritized, and (iii) 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 some space for encompassing more useful cases. Considering these findings, equation (3.3) will be changed to:

$$show(w_i) = \mathbb{1}\left(\mathbb{1}(fr(w_i) - \theta_{fr}) + \mathbb{1}(sr(w_i) - \theta_{sr}^{ASRcor(w_i)}) + \mathbb{1}(fr(w_i) - \theta_{sp}^{ASRcor(w_i)}) \times sp(w_i) + keep(w_i)\right) \times (1 - hide(w_i)) \quad (5.1)$$

5.1.2 Augmenting Baseline PSC with ASR Erroneous Cases

To make use of ASR errors, erroneous segments of the 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 ASR erroneous phrase along with its corresponding phrase in the transcript is then scanned for possible matches of homophones, minimal pairs, and negative cases. In addition, 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 the 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 the 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 the 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 phonetic transcripts 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 (e.g., “work” and “works”) in the present tense and past tense for regular verbs (e.g., “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) the 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, 2005). 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., “*right you are*” as “*rye chew are*”). The assimilation rule is realized using Gimson’s English assimilation standards (Cruttenden, 2014), which are very systematic and follow restricted patterns.

- **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 the 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 & 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 the 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.

5.2 Feature Extraction from ASR Errors

To extract features from ASR errors, 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 based on the findings of the previous chapter: homophones $hp(w_i, \hat{w}_i)$, minimal pairs $mp(w_i, \hat{w}_i)$, negatives $ng(w_i, \hat{w}_i)$, and the 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} or successor w_{i+1} is detected in $\hat{w}_i$.

In the next step, 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 sent to PSC Rule Engine to determine whether it should be shown or not. Based on the ASR correctness flag $ASRcor(w_i)$ for word w_i , this unit selects the appropriate procedure and thresholds to make a decision that is summarized in equation (5.2). If ASR transcribes the word correctly, the word frequency, speech rate, and the frequency of specific words are compared with the strict thresholds. On the other hand, if the ASR transcript contains an error, the baseline thresholds (the ones obtained from user assessments) are used. The show-decision is then filtered out if the hide-list chooses to hide the word. 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. Words added to PSC by the new features should not be suppressed by the feature. Equation (5.2) extends equation (5.1) with the new features:

$$show(w_i) = \mathbb{1} \left[\left(\mathbb{1}(fr(w_i) - \theta_{fr}) + \mathbb{1}(sr(w_i) - \theta_{sr}^{ASRcor(w_i)}) + \right. \right. \\ \left. \mathbb{1}(fr(w_i) - \theta_{sp}^{ASRcor(w_i)}) \times sp(w_i) + keep(w_i) \right) \times (1 - hide(w_i)) + \\ \left. hp(w_i, \hat{w}_i) + mp(w_i, \hat{w}_i) + ng(w_i, \hat{w}_i) + bb(w_{i-1:i+1}, \hat{w}_i) \right] \quad (5.2)$$

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 in this case.

5.3 Enhanced PSC System Realization

5.3.1 Extended System Overview

Figure 5.1 (red) depicts the extension on the baseline PSC system (black). Via a word-level forced-alignment procedure, the original transcript is synchronized with the speaker’s utterance. Speech rate, word frequency, and specificity are extracted for each word in the Feature Extraction module.

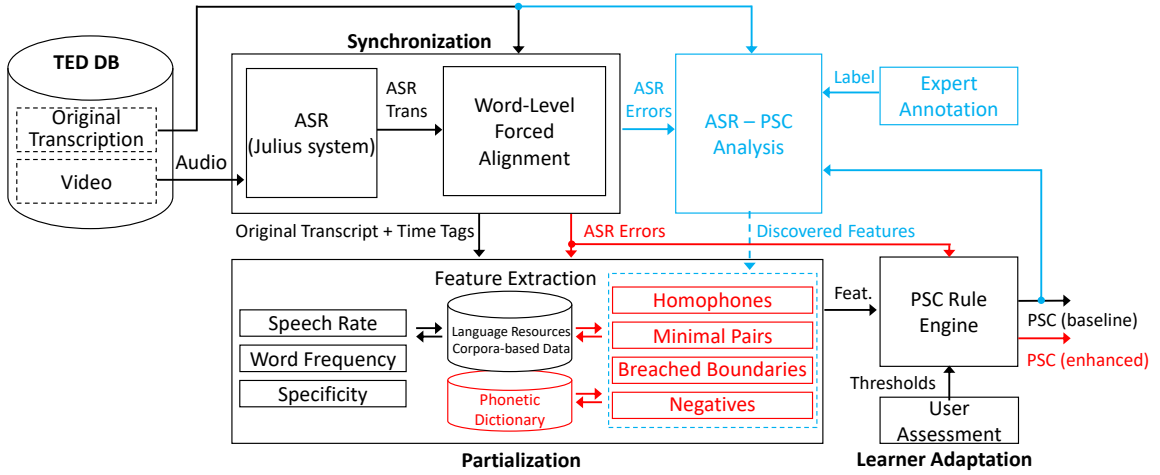


Figure 5.1: Enhanced PSC Process Flow: Baseline (black) employs ASR system to synchronize words with their speech segments, and then partialize the text based on its features. Via a root-cause analysis, ASR-PSC analysis (blue) examines several features to be incorporated into PSC’s feature extraction. Once the features are identified, they are added to the feature pool of the system. These features (red) enable the system to detect potentially difficult speech segments to be included in the enhanced PSC.

To improve the baseline PSC, the system is extended to extract the four categories of features derived from ASR errors (homophones, minimal pairs, negatives, and breached boundaries). Moreover, ASR correct cases that were contrarily detected to be difficult by PSC’s features were utilized as a signal for relatively easier segments of speech, hence, can be removed from PSC. 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 the ASR transcript, which in turn, is used to extract the new set of features. Moreover, the decision in the rule engine considers the thresholds suggested by the ASR error status along with the user proficiency. On

top of this, the amount of ASR-error driven features (e.g., breached boundaries) could be adjusted by exploring other candidates from the list of the N-best hypotheses, in addition to the 1-best output that is the default of the system. The enhanced PSC is expected to outperform the baseline PSC in providing essential clues for recognition of the listening tasks for the L2 learners.

5.3.2 Statistics of Baseline PSC versus Enhanced PSC

Table 5.1 indicates the statistical comparison between the baseline PSC and the enhanced PSC with regard to ASR correct and erroneous cases.

Table 5.1: Baseline PSC versus Enhanced PSC (70 TED Talks)

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

As the table presents, the enhanced PSC version includes 41% of ASR errors, compared with the baseline PSC, which includes 22% of ASR errors, while 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 the 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. Figure 5.2 shows a screenshot of baseline vs. enhanced PSC.

5.4 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



Figure 5.2: Screenshot of baseline (left) and enhanced (right) PSC. The original sentence was “if I tried to make a new ear”. The phrase “make a new ear” caused an ASR error. Many language learners had difficulty in transcribing this segment. Their transcriptions included cases such as “make a new year”, “making you here”, “make it in new air”, “make in you hear”, etc. ©TED Talk by Alan Russell: The potential of regenerative medicine.

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 the listening material (Buck, 2001).

5.4.1 Participants

In this experiment 36 Japanese and 2 Chinese undergraduate students, mostly from engineering fields, participated. The participants’ TOEFL® ITP¹ scores ranged from 450 to 560. ITP stands for “Institutional Testing Program” and uses 100 percent academic content to evaluate the English-language proficiency of non-native English speakers. The test evaluates skills in three areas:

1. Listening Comprehension: measures the ability to understand spoken English as it is used in colleges and universities,
2. Structure and Written Expression: measures recognition of selected structural and grammatical points in standard written English,

¹https://www.ets.org/toefl_itp

3. Reading Comprehension: measures the ability to read and understand the academic reading material in English.

The participants' scores can categorize them as pre-intermediate level. All participants in this experiment were enrolled in a CALL class, where the experiment was held. The participants were given instruction on how to perform the test both in English and in the Japanese language.

5.4.2 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. In this view, we circumvent a situation where learners prioritize a version over another because of the larger quantity of shown words. While both the baseline PSC and the enhanced version includes the same number of words in the target sentence, the shown words were different. More specifically, the shown words in the enhanced version included an instance of ASR errors and are assumed to be a better means for disambiguating the difficulties of speech.

5.4.3 Procedure

The experiment consisted of two parts:

Part I: 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 set of target words they had to transcribe. The participants were then 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 compared to the other version. It should be noted that learners were uninformed about which choice is the baseline PSC or the enhanced PSC.

Part II: To evaluate the enhanced PSC over the baseline PSC with a more quantitative approach, we also designed a paraphrasing test. Accordingly, in Part II of the experiment, the learners were divided into two groups:

1. those who received the baseline PSC along with the videos and
2. 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 experiment provide us with quantitative data on evaluation of the baseline and the enhanced PSC based on the learners' scores.

5.4.4 Results

Figure 5.3 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.

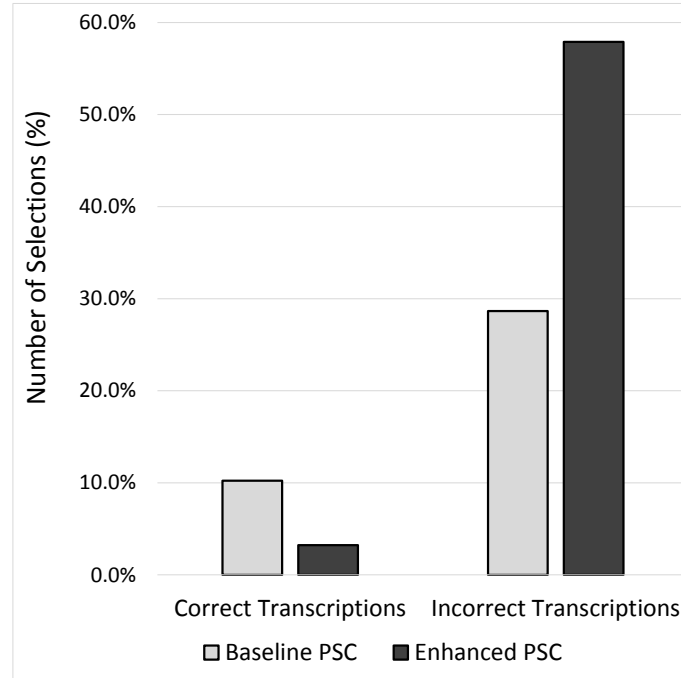


Figure 5.3: Experimental evaluation of Baseline PSC and Enhanced PSC – Part I: Participants’ preferences on choosing between baseline PSC and enhanced PSC after transcription.

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 5.4 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 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% incorrect responses. Findings of this experiment, which is based on the 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.

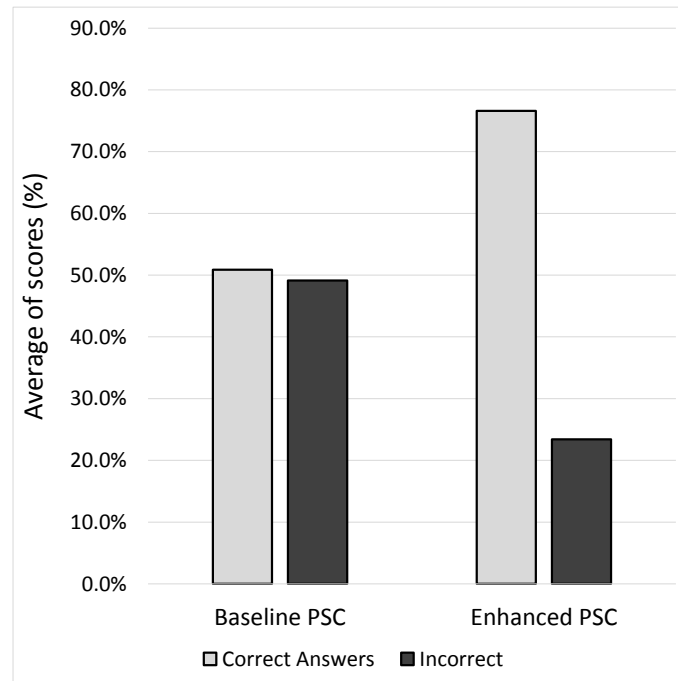


Figure 5.4: Experimental evaluation of Baseline PSC and Enhanced PSC – Part II: Paraphrasing scores of the participants in baseline PSC group versus the enhanced PSC group.

5.5 Conclusion

This chapter leveraged the ASR errors in detecting problematic speech segments and improving the word selection criteria in PSC. Following a thorough analysis in Chapter 4, it was found that several categories in the ASR errors signal the difficulties for L2 listeners. These categories included homophones, minimal pairs, negatives, and breached boundaries. In this chapter, we explained how the baseline PSC system was extended to detect these cases and generate the enhanced PSC, drawing upon the ASR clues. This was realized by analyzing the ASR output, which provided some insights on how to refine PSC's selection. In this view, ASR output shed light on both easy and difficult words and phrases in the input, which could be directly used to enhance PSC to better address L2 listeners' needs. The chapter explained the

procedure on omitting easy cases and embedding difficult ones to realize the enhanced version of the PSC system.

Experiments with the L2 listeners using the enhanced PSC, which included the ASR errors, revealed that the enhanced version could effectively assist the L2 listeners in recognizing the speech compared to the baseline system and is preferred to the baseline when learners' views were elicited on identifying which captions (baseline or enhanced) could assist them better and provided more useful words or phrases that they could not recognize easily through the course of listening.

Chapter 6

Conclusions

6.1 Contributions and Summary

This thesis introduced a novel captioning method, Partial and Synchronized Caption (PSC), as a tool for developing L2 listening skill and proposed a new approach, exploiting the ASR cues, to detect difficult speech segments for L2 listeners and improve the baseline PSC system. Accordingly, this research has three main contributions:

1. Proposing PSC as a new tool of captioning to train L2 listening skill,
2. Investigating ASR errors as a source to predict L2 listeners difficulties,
3. Using ASR clues to enhance the baseline PSC system.

This new motivation comes from the limitation of the previous captioning methods, which provide the full text thus allowing comprehension of the material merely by reading the text, promoting word-by-word decoding strategy and fostering dependence on reading the text rather than training the L2 listening skills. Nevertheless, the proposed method, PSC, deals with these limitations and promotes listening to the speech by presenting a selected subset of words, where each word is synched to its corresponding speech signal. It provides both teachers and learners with an effective tool to systematically improve L2 listening skill, by allowing the gradual reduction of textual clues in the caption through the course of training, while receiving a caption that matches the requirements of individual learners and scaffolds them only when it is necessary.

As the baseline form, PSC detects difficult words based on three factors that lead to listening difficulty: speech rate, word frequency, and specificity. Through calculating these features, the system generates a caption that allows the use of limited textual clues to foster L2 listening skill. Specifically, PSC presents difficult words on the screen and hides easy words to promote more listening and less reading. With the use of ASR technology, the system realizes word-level synchronization between the text and the speech, which emulates the speech flow, fosters text-to-speech mapping and alleviates the salient emergence of the words. Through learner assessment, specifically the tolerable speech rate and the vocabulary size, PSC strives to address learners' demands and adjusts the captions to their proficiency levels.

Experimental evaluations of the baseline system showed that learners scored better when using PSC compared to the no-caption condition. Furthermore, with a considerably fewer numbers of shown words (less than 30%), PSC could provoke a statistically equivalent level of comprehension as the full caption condition. Results also indicated that learners gained significantly higher scores when exposed to a new segment of the video without any caption right after they had watched the previous part of that video with PSC. The finding confirms the effectiveness of PSC for preparing learners to handle simulated real-life situations, where assistive tools are not available. PSC can assist learners to obtain adequate comprehension of videos by presenting less than 30% of the transcript, hence, it is expected to be effective particularly for those learners who read the captions extensively for comprehending the video.

Given that difficult words are not bounded to the baseline PSC's defined criteria, the system yet anticipated for improvement to cover other difficult cases. To address this issue, this study investigated 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 the hypothesis on the usefulness of ASR errors in predicting difficulties of the audio, 70 TED talks were transcribed by an ASR system and the ASR errors were analyzed to discover the underlying factors. Annotation of these errors identified four

categories among many possible features that deemed to be useful for enhancing the baseline PSC: homophones, minimal pairs, negatives, and breached boundaries. To confirm the usefulness of these features, which derived from ASR erroneous cases, an experiment was conducted with L2 listeners. Results showed that these four categories of ASR errors were problematic for L2 listeners, whereas the learners hardly faced difficulties in transcribing easy (control) cases.

Following this analysis, the baseline PSC was enhanced by leveraging ASR errors i.e., by embedding ASR erroneous cases as problematic speech segments and eliminating some of ASR correct cases as easy to recognize speech segments. To this end, secondary thresholds were defined for the speech rate and word specificity features in the baseline PSC based on ASR correct cases. Furthermore, useful ASR erroneous cases were incorporated into the baseline to form the enhanced version. The enhanced PSC was then compared against the baseline version in another experiment. The results of this experiment revealed that L2 listeners noticeably preferred the enhanced version 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 the 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 clues. In this view, not all ASR errors are good predictors of learners' difficulty in listening, but some of them, such as breached boundaries, were indeed worth investigating. Moreover, the enhanced PSC system, which includes ASR clues together with other factors accounting for L2 listening difficulties, is a learner-adaptive tool to train L2 listening by detecting challenging speech segments and difficult words/phrases to provide optimal assistance.

6.2 Future Work

The author concludes this paper with some future directions. We can encompass other features that affect L2 listening comprehension. This is crucial for increasing the accuracy of word selection in PSC and hence providing better assistance to the learners. A wide range of features such as accent, co-articulation, idiomaticity, part of speech, and even the pragmatic implications of words can be considered. In addition, there are other possible directions such as automation based on NLP, L2 learner modeling using ASR, learner adaptation using machine learning, and CALL system development, which form the future work of this thesis.

6.2.1 Data-driven PSC

The rule-engine of the proposed PSC system contains domain knowledge of linguistics and is fine-tuned by domain experts. Machine learning techniques, however, provide the means to extract such rules from data. For instance, decision trees are able to build a rule-engine with numerical and categorical features provided in this study. To this end, each word of the transcript would be assigned a binary show/not-show label. Through such data-driven approach, a decision tree (or similar methods) could learn the underlying rules of the PSC and generalize it to unseen videos.

Another alternative is to train a classifier to classify words as shown/not-shown in the PSC framework. Such system requires normalized real-valued features to work best, but the major problem is that it requires a large amount of annotated data. The system then uses the redesigned features to predict the label of unforeseen data.

6.2.2 ASR as a Model of Language Learner

The findings in this study 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. 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' demands. This process can be done by degrading the acoustic model or the language model through reducing the training data. In this regard, we can degrade the ASR models to the target learner's level. To this end, we can also 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 limiting the training data or omitting low-frequency words from the dictionary. The ASR error analysis unit is then provided with the transcript of these attenuated ASR outputs to find new candidates for PSC.

6.2.3 Learner Adaptation in PSC

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. The system should constantly analyze user's improvement over time in order to adapt the captions to individual learners. To this end, a history of the learner's performance should be embedded into the system. This can also help the teachers monitor the students' progress. In the meantime, the system should consider other measurements for user level assessment especially as new features are introduced and aggregated to the system for word selection. Moreover, through constant observation of the learner's profile information, use of the caption, performance on the evaluation tests, the history of exposure to specific terminologies, feedbacks, etc., the captioning system should adapt itself to the level of the learner.

6.2.4 PSC-Integrated CALL System Development

Finally, it is ambitious to use PSC as a core of a CALL system for training L2 listening skill. Ideally, learners can log into the system and create a profile, which constantly presents their progress, suggests the appropriate videos based on their progress, interest and background, provides the video with adjusted caption to them, evaluates their performance through suggesting the user to take several tests, provides the user with concrete feedback and recommends the user to activate/deactivate

several features to better train his/her listening skill. This system can also encompass a gamification feature, which can engage the learner into a game where the learner can compete with others using accumulating points, which are gained through lowering down the number of shown words and realizing a higher level of independence on caption, etc.

Moreover, the current version, which supports English, can be extended to other languages to be used by other L2 learners. In this sense, PSC may be used as a universal training tool for L2 listening development for learners aiming to learn different languages.

Appendix I

List of TED Talks

1. Helen Fisher (2008): Why we love, Why We cheat
2. Al Gore (2008): New thinking on the climate crisis
3. Elizabeth Gilbert (2009): Your elusive creative genius
4. Daniel Pink (2009): The puzzle of motivation
5. Becky Blanton (2009): The year I was homeless
6. Dan Buettner (2009): How to live to be 100+
7. Jane McGonigal (2014): Gaming can make a better world
8. Amy Cuddy (2012): Your body language shapes who you are
9. Matt Cutts (2011): Try something new for 30 days
10. Pamela Meyer (2011): How to spot a liar
11. Alan Rusell (2006): The potential of regenerative medicine
12. Alex Steffen (2005): The route to a sustainable future
13. James Howard (2004): The ghastly tragedy of suburbs
14. John Doerr (2007): Salvation (and profit) in greentech
15. Majora Carter (2006): Greening the ghetto
16. Robert Thurman (2006): We can be Buddhas
17. Robert Wright (2006): Progress is not a zero-sum game
18. Carolyn Porco (2007): This is Saturn
19. Dan Dennett (2006): Lets teach religion all religion in schools
20. David Keith (2007): A critical look at geoengineering against climate change



21. Dean Kamen (2002): The emotion behind invention
22. Dean Ornish (2006): Healing through diet
23. Erin McKean (2007): The joy of lexicography
24. Hod Lipson (2007): Building self-aware robots
25. Jonathan Harris (2004): The Webs secret stories
26. Kevin Kelly (2005): How technology evolves
27. Seth Godin (2003): How to get your ideas to spread
28. Steven Pinker (2005): What our language habits reveal

Demo videos are available in: <http://sap.ist.i.kyoto-u.ac.jp/psc/#DEMO>.

Appendix II

Comprehension Questions Sample

The following is a sample of comprehension questions used in our experiment. Please note that the students could see questions one by one. They could only see the next question after they had answered the previous one without being able to go back. The students were allowed to take notes if they wanted.

- **Who created the candle problem first?** *(B)*
 - A. Sam Glucksberg
 - B. Karl Dunker
 - C. Daniel Pink
- **What is the candle problem about?** *(C)*
 - A. Melt the candle and adhere it to the wall without using the matches.
 - B. Melt the candle and attach it to the wall without using any thumbtacks.
 - C. Adhere the lit candle to the wall so that it won't drip wax onto the table.
- **What was the reward offered in the Glucksberg's first experiment?** *(A)*
 - A. If you are the fastest to solve the problem, you receive 20\$.
 - B. If you are in the top 25% of the fastest times, you receive 10\$.
 - C. If you are in the top 10% of the fastest times, you receive 5\$.
- **What was the result of Glucksberg's first experiment?** *(B)*
 - A. The group "with" reward solved the problem more quickly.
 - B. The group "without" reward solved the problem more quickly.
 - C. None of the groups could find the solution in a given time
- **According to the speaker, which one is true:** *(A)*
 - A. Business is focusing on wrong motivators and incentives
 - B. Business should offer more rewards to improve thinking

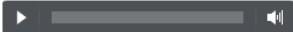
- C. Business should include more punishment for better results
- **The incentivized group beat the other one in second experiment, because (C)**
 - A. they were more creative and intelligent
 - B. they were given more time and reward
 - C. they were given a clear goal plus a reward
 - **The solution to the candle problem can be found by (B)**
 - A. sticking to functional fixedness
 - B. looking for solution in periphery
 - C. concentrating on the incentives
 - **What does speaker say about white-color workers in 21st century? (B)**
 - A. They are still doing rule-based tasks
 - B. They are less involved in routine tasks
 - C. They are doing more left-brain activities

The following is a sample of the cloze tests. The students could fill the blanks while listening to the audio. (Answers: *Solution – Fixedness – Platform – Experiment*)

Listen to the audio and while listening fill in the blanks.

この文章には書き落とされた言葉があります。その言葉の変わりに () があります。
オーディオを聴きながら, () に書き落とされた言葉を書きなさい。

YOU CAN LISTEN ONLY ONE TIME!!!!!!



And eventually, after five or 10 minutes, most people figure out the which you can see here. The key is to overcome what's called functional . You look at that box and you see it only as a receptacle for the tacks. But it can also have this other function, as a for the candle. The candle problem. Now I want to tell you about an using the candle problem.

Next

Appendix III

Questionnaire on Baseline PSC

A 5-point Likert-scale questionnaire with the scale of 1 (strongly disagree) to 5 (strongly agree) was used to get the learner feedback on PSC. Table III.1 presents the results.

Table III.1: 5-point Likert-Scale Survey Results

Statements	Mean	SD
S1: I think PSC is a good idea.	3.92	0.86
S2: I think PSC helps me understand.	3.84	1.00
S3: I think PSC helps me follow the audio without being distracted.	3.64	0.89
S4: I think PSC is better than FC.	2.84	1.00
S5: I think PSC is enough to understand.	3.18	1.16
S6: I think PSC helped me use my listening skill more.	3.20	1.06
S7: I think PSC is better than FC as I can't read all text.	2.93	1.01
S8: I think Synchronized Caption is very helpful.	3.75	0.84
S9: I think showing “...” instead of hidden words is a good idea.	3.57	1.02
S10: I could find most of words I did not know in PSC.	3.40	0.93
S11: I could find most of the words with high speech rate in PSC.	3.30	0.95
S12: I think the captions of videos were easy to read.	3.15	1.28

The first two statements explored the attitudes toward our method, and received almost positive responses and support (respectively 72% and 69% of the participants agreed with those statements - scored 4 or higher in Likert-scale). Items S4 through S7 elicit views on the effectiveness of PSC compared to FC method. Data on these

items ranged from 2.84 to 3.20 (averaged 3.03), indicating that the learners are still not sure if PSC can be substituted for FC.

Statements S3 and S8 focused on the word-level synchronization aspect of PSC. Responses to S3 do not reflect that PSC is distractive (only 2% of the participants found it very distractive). The results of S8 show nearly positive feedbacks on the synchronization feature of PSC (67% of the participants selected 4 or higher in the Likert-scale). By items S10 and S11, we investigated views on the partialization aspect of PSC and the selected words. Approximately 68% of the subjects (40 out of 58) chose Neutral (score=3) or Agree (score=4) and 16% chose strongly agree (score=5) in response to these items. To check PSC's readability, we designed items S9 and S12. Findings suggest some improvement should be considered in presenting the captions.

Appendix IV

Publications by the Author

Journal Papers (peer-reviewed):

1. Mirzaei, M. S., Meshgi, K., & Kawahara, T. (Under Review) Exploiting Automatic Speech Recognition Errors to Enhance Partial and Synchronized Caption for Facilitating Second Language Listening. *Computer Speech & Language Journal*, Elsevier.
2. Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2017). Partial and Synchronized Captioning: A New Tool to Assist Learners in Developing Second Language Listening Skill. *ReCALL Journal*, Cambridge University Press.

Conference Papers (peer-reviewed):

1. Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2016). Automatic Speech Recognition Errors as a Predictor of L2 Listening Difficulties. In D. Brunato, F. Dell’Orletta, G. Venturi, T. Franois, & P. Blache (Ed.), *Proceedings of COLING 2016 – Workshop on Computational Linguistics for Linguistic Complexity (CL4LC’16)*, Osaka, Japan, (pp. 192–201).
2. Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2016). Leveraging Automatic Speech Recognition Errors to Detect Challenging Speech Segments in TED Talks. In S. Papadima-Sophocleous, L. Bradley & S. Thouëсны (Ed.), *CALL Communities and Culture - Proceedings of the 2016 EUROCALL Conference, Limmasol, Cyprus* (pp. 313–318). Dublin: Research-publishing.net.
3. Mirzaei, M. S. (2016). Using Automatic Speech Recognition Technology to Assist L2 Listeners. *Proceeding of the 56 LET National Conference, Tokyo, Japan*.
4. Mirzaei, M. S., Meshgi, K., Akita, Y., & Kawahara, T. (2015). Errors in Automatic Speech Recognition versus Difficulties in Second Language Listening.

- In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Ed.) *Critical CALL - Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (p. 410). Dublin: Research-publishing.net.
5. Mirzaei, M. S., & Kawahara, T. (2015). ASR Technology to Empower Partial and Synchronized Caption for L2 Listening Development. In S. Steidl, A. Batliner, & O. Jokisch (Ed.) *INTERSPEECH 2015 – Workshop on Speech and Language Technology in Education (SLaTE'15), Leipzig, Germany* (pp. 65–70).
 6. Mirzaei, M. S. (2015). Automatic Speech Recognition Errors Enrich Partial and Synchronized Caption to Develop Listening Skill. *Proceeding of the 55 LET National Conference, Osaka, Japan*.
 7. Mirzaei, M. S., Akita, Y., & Kawahara, T. (2014). Partial and Synchronized Captioning: A New Tool for Second Language Listening Development. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Ed.), *CALL Design: Principles and Practice – Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 230–236). Dublin: Research-publishing.net.
 8. Mirzaei, M. S., Akita, Y., & Kawahara, T. (2014). Partial and Synchronized Caption Generation to Develop Second Language Listening Skill. In *ICCE 2014 – Workshop of Natural Language Processing Techniques for Educational Applications (NLP-TEA'14), Nara, Japan* (pp. 13–23).
 9. Mirzaei, M. S., & Kawahara, T. (2014). Listen More, Read Less: Partial and Synchronized Captions to Train L2 Listening using TED Talks. *Proceeding of the 54 LET National Conference, Fukuoka, Japan*, (pp. 106–107).

Technical Reports:

1. Mirzaei, M.S., & Kawahara, T. (2014). Partial and Synchronized Caption Generation to Enhance the Listening Comprehension Skills of Second Language Learners. *Proceeding of the 101 Joint Speech & Language Processing (SLP), Tokyo, Japan, 2014(15)*, (pp. 1–8).

Bibliography

- Adami, A. G. (2010). Automatic speech recognition: From the beginning to the portuguese language. In *9th International Conference on Computacional Processing of the Portuguese Language*.
- Adank, P., Evans, B., Stuart-Smith, J., & Scott, S. (2009). Familiarity with a regional accent facilitates comprehension of that accent in noise. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), 520–529.
- Adda-Decker, M. & Lamel, L. (2005). Do speech recognizers prefer female speakers? In *INTERSPEECH* (pp. 2205–2208).
- Aldabe, I., Arrieta, B., De Ilarraza, A. D., Maritxalar, M., Niebla, I., Oronoz, M., & Uria, L. (2006). The use of nlp tools for basque in a multiple user call environment and its feedback. In *The 13th Conference on Natural Language Processing (TALN 2006). April 10-13, 2006. Leuven (Belgium)* (pp. 815–824).
- Amaral, L. A. & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(01), 4–24.
- Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science(Washington)*, 228(4698), 456–462.
- Aydelott, J. & Bates, E. (2004). Effects of acoustic distortion and semantic context on lexical access. *Language and Cognitive Processes*, 19(1), 29–56.
- Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556.
- Bailly, G. & Barbour, W.-S. (2011). Synchronous reading: learning french orthography by audiovisual training. In *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* (pp. 1153–1156).
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2), 1001–1024.
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvett, D., Fissore, L., Laface, P., Mertins, A., Ris, C., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10), 763–786.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. ” O’Reilly Media, Inc.”.

- Bird, S. A. & Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics*, 23(04), 509–533.
- 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, DTIC Document.
- Braunschweiler, N., Gales, M. J., & Buchholz, S. (2010). Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *INTERSPEECH* (pp. 2222–2225).
- Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and cognitive processes*, 27(7-8), 1205–1224.
- Buck, G. (1988). Testing listening comprehension in japanese university entrance examinations. *JALT journal*, 10(1), 15–42.
- Buck, G. (2001). *Assessing listening*. Cambridge University Press.
- Chang, A. C.-S. (2009). Gains to l2 listeners from reading while listening vs. listening only in comprehending short stories. *System*, 37(4), 652–663.
- Chapelle, C. (1998). Multimedia call: Lessons to be learned from research on instructed sla. *Language learning & technology*, 2(1), 22–34.
- Chapelle, C. (2001). *Computer applications in second language acquisition*. Cambridge University Press.
- 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* (pp. 7418–7422).: IEEE.
- Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, 34(2), 213–238.
- Cruttenden, A. (2014). *Gimson's pronunciation of English*. Routledge.
- Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation.
- Cutler, A. (2005). The lexical statistics of word recognition problems caused by l2 phonetic confusion. In *Interspeech* (pp. 413–416).
- Cutler, A. & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31(2), 218–236.
- Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal des traducteurs/Translators' Journal*, 49(1), 67–77.
- Davies, M. (2008). The corpus of contemporary american english: 520 million words, 1990–2015. <http://corpus.byu.edu>. Accessed: 2013-07-04.
- Diao, Y., Chandler, P., & Sweller, J. (2007). The effect of written text on comprehension of spoken english as a foreign language. *The American journal of psychology*, (pp. 237–261).

- Dunkel, P. (1988). The content of l1 and l2 students' lecture notes and its relation to test performance. *TESOL Quarterly*, 22(2), 259–281.
- Ehsani, F., Bernstein, J., & Najmi, A. (2000). An interactive dialog system for learning japanese. *Speech Communication*, 30(2), 167–177.
- Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. *The handbook of second language acquisition*, 14, 63.
- Ellis, N. C. & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language learning*, 43(4), 559–617.
- 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(8), 2301–2312.
- Field, J. (1998). Skills and strategies: Towards a new methodology for listening. *ELT journal*, 52(2), 110–118.
- Field, J. (2003). Promoting perception: Lexical segmentation in l2 listening. *ELT journal*, 57(4), 325–334.
- Field, J. (2008). Bricks or mortar: which parts of the input does a second language listener rely on? *TESOL quarterly*, 42(3), 411–432.
- Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech communication*, 40(4), 467–491.
- Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in english: Can listeners adapt? *Journal of Psycholinguistic Research*, 38(4), 379–412.
- Forsberg, M. (2003). Why is speech recognition difficult. *Chalmers University of Technology*.
- Fosler-Lussier, E. & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2), 137–158.
- Gamper, J. & Knapp, J. (2002). A review of intelligent call systems. *Computer Assisted Language Learning*, 15(4), 329–342.
- Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, (pp. amt015).
- Garza, T. J. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals*, 24(3), 239–258.
- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. technical report no. 257.
- Ghannay, S., Camelin, N., & Esteve, Y. (2015). Which asr errors are hard to detect. In *Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing (ERRARE 2015) Workshop, Sinaia, Romania* (pp. 11–13).

- Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language teaching*, 40(02), 97–118.
- Goh, C. C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1), 55–75.
- 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(3), 181–200.
- Graff, D., Fiscus, J., & Garofolo, J. (2002). 1997 hub4 english evaluation speech and transcripts. *Linguistic Data Consortium, Philadelphia*, 133.
- Greenberg, S. & Chang, S. (2000). Linguistic dissection of switchboard-corpus automatic speech recognition systems. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.
- Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL quarterly*, 26(2), 385–390.
- Guillory, H. G. (1998). The effects of keyword captions to authentic french video on learner comprehension. *Calico Journal*, (pp. 89–108).
- Hasan, A. S. (2000). Learners' perceptions of listening comprehension problems. *Language Culture and Curriculum*, 13(2), 137–153.
- Heift, T. & Schulze, M. (2003). Error diagnosis and error correction in call. *CALICO Journal*, 20(3), 433–436.
- Henning, G. (1990). a study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on toefl listening comprehension item performance. *ETS Research Report Series*, 1990(2).
- Huang, C.-L., Dixon, P. R., Matsuda, S., Wu, Y., Lu, X., Saiko, M., & Hori, C. (2013). The nict asr system for iwslt 2013. In *Proc. Int. Workshop Spoken Language Translation*.
- Inhoff, A. W. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6), 431–439.
- Jurafsky, D. & Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson.
- King, J. (2002). Using dvd feature films in the efl classroom. *Computer Assisted Language Learning*, 15(5), 509–523.
- Kitaoka, N., Enami, D., & Nakagawa, S. (2014). Effect of acoustic and linguistic contexts on human and machine speech recognition. *Computer Speech & Language*, 28(3), 769–787.
- Korat, O. (2010). Reading electronic books as a support for vocabulary, story comprehension and word reading in kindergarten and first grade. *Computers & Education*, 55(1), 24–31.
- Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of toefl dialogue items. *ETS Research Report Series*, 2004(1).

- Krashen, S. (1981). Second language acquisition. *Second Language Learning*, (pp. 19–39).
- Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Addison-Wesley Longman Ltd.
- Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., & Zavaliagkos, E. (1994). Comparative experiments on large vocabulary speech recognition. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 1 (pp. I–561).: IEEE.
- Kukulska-Hulme, A. & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, 20(03), 271–289.
- Laufer, B. (1990). Words you know: How they affect the words you learn. *Further insights into contrastive linguistics*, (pp. 573–593).
- Lawson, A. D., Harris, D. M., & Grieco, J. J. (2003). Effect of foreign accent on speech recognition in the nato n-4 corpus. In *INTERSPEECH*.
- Lee, A. & Kawahara, T. (2009). Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference* (pp. 131–137).
- Leeser, M. J. (2007). Learner-based factors in l2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, 57(2), 229–270.
- Leveridge, A. N. & Yang, J. C. (2013). Testing learner reliance on caption supports in second language listening comprehension multimedia environments. *ReCALL*, 25(02), 199–214.
- Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.
- Levy, M. & Stockwell, G. (2013). *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.
- Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-put-er*. Citeseer.
- Lightbown, P. M. & Spada, N. (2006). *How languages are learned*. Oxford University Press.
- Ludwig, J. (1984). Vocabulary acquisition as a function of word characteristics. *Canadian Modern Language Review*, 40(4), 552–62.
- Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The modern language journal*, 75(2), 196–204.
- Markham, P. & Peter, L. (2003). The influence of english language and spanish language captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems*, 31(3), 331–341.

- Markham, P. L. (1989). The effects of captioned television videotapes on the listening comprehension of beginning, intermediate, and advanced esl students. *Educational Technology*, 29(10), 38–41.
- Martinez, F., Tapias, D., Alvarez, J., & Leon, P. (1997). Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Fifth European Conference on Speech Communication and Technology*.
- Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology*, 28(5), 653–660.
- Mayer, R. E. & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1), 43–52.
- Mayor, M. (2009). Call-enhanced l2 listening skills: Aiming for automatization in a multimedia environment. *Indian Journal of Applied Linguistics*, 35(1), 1–9.
- Medwell, J. (1998). The talking books project: some further insights into the use of talking books to develop reading. *Reading*, 32(1), 3–8.
- Mendelsohn, D. J. (1984). There are strategies for listening.
- Meyer, B., Wesker, T., Brand, T., Mertins, A., & Kollmeier, B. (2006). A human-machine comparison in speech recognition based on a logatome corpus. In *Speech Recognition and Intrinsic Variation Workshop*.
- 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(1), 388–403.
- Miller, G. A. & Licklider, J. C. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 22(2), 167–173.
- Mirghafori, N., Fosler, E., & Morgan, N. (1995). Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes. In *EUROSPEECH*, volume 95 (pp. 491–494).
- Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014a). Effects of captioning on video comprehension and incidental vocabulary. *Language, Learning & Technology*, 18(1), 118–141.
- Montero Perez, M., Peters, E., & Desmet, P. (2014b). Is less more? effectiveness and perceived usefulness of keyword and full captioned video for l2 listening comprehension. *ReCALL*, 26(01), 21–43.
- Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for l2 listening and vocabulary learning: A meta-analysis. *System*, 41(3), 720–739.
- 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)* (pp. 145–150).: Max Planck Institute for Psycholinguistics.
- Moran, S. (2012). The effect of linguistic variation on subtitle reception. *Eye tracking in audiovisual translation*, (pp. 183–222).

- Munteanu, C., Baecker, R., Penn, G., Toms, E., & James, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 493–502).: ACM.
- Murphy, G. (2004). *The big book of concepts*. MIT press.
- Murray, I. R. & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2), 1097–1108.
- Nanjo, H. & Kawahara, T. (2004). Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Transactions on speech and Audio Processing*, 12(4), 391–400.
- Naptali, W. & Kawahara, T. Automatic speech recognition for ted talks.
- Nation, I. S. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82.
- Nation, I. S. & Beglar, D. (2007). A vocabulary size test. *The language teacher*, 31(7), 9–13.
- Nation, I. S. & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.
- Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPHS* (pp. 1157–1160).
- Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in l2 dutch. *ReCALL*, 20(02), 225–243.
- Nilsen, D. L. & Nilsen, A. P. (2010). *Pronunciation contrasts in English*. Waveland Press.
- Nissan, S., DeVincenzi, F., & Tang, K. L. (1995). An analysis of factors affecting the difficulty of dialogue items in toefl listening comprehension. *ETS Research Report Series*, 1995(2).
- Nitta, H., Okazaki, H., & Klinger, W. (2010a). An analysis of articulation rates in movies. *ATEM Journal*, 15, 41–56.
- Nitta, H., Okazaki, H., Klinger, W., et al. (2010b). Missed word rates at increasing listening speeds of high-level japanese speakers of engilsh. *Senshu University Research Society*, 87, 171–198.
- Nogami, Y. & Hayashi, N. (2009). A japanese adaptive test of english as a foreign language: developmental and operational aspects. In *Elements of adaptive testing* (pp. 191–211). Springer.
- Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1209.

- Osada, N. (2001). What strategy do less proficient learners employ in listening comprehension?: A reappraisal of bottom-up and top-down processing. *Journal of Pan-Pacific Association of Applied Linguistics*, 5(1), 73–90.
- Osada, N. (2004). Listening comprehension research: A brief review of the past thirty years. *Dialogue*, 3(1), 53–66.
- Oxford, R. L. (1993). Research update on teaching l2 listening. *System*, 21(2), 205–211.
- Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.
- Pan, Y., Jiang, D., Yao, L., Picheny, M., & Qin, Y. (2010). Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1725–1734).: ACM.
- Paul, D. B. & Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language* (pp. 357–362).: Association for Computational Linguistics.
- Pietquin, O. & Beaufort, R. (2005). Comparing asr modeling methods for spoken dialogue simulation and optimal strategy learning. In *INTERSPEECH* (pp. 861–864).: Citeseer.
- Pimsleur, P., Hancock, C., & Furey, P. (1977). Speech rate and listening comprehension. *Viewpoints on English as a second language*, (pp. 27–34).
- Pujolà, J.-T. (2002). Calling for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, 14(02), 235–262.
- Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3), 353–362.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), 372.
- Révész, A. & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35(01), 31–65.
- Rost, M. (2005). L2 listening. *Handbook of research in second language teaching and learning*, (pp. 503–527).
- Rost, M. (2013). *Teaching and researching: Listening*. Routledge.
- Sadighi, F. & Zare, S. (2006). Is listening comprehension influenced by the background knowledge of the learners? a case study of iranian efl learners. *The Linguistics Journal*, 1(3), 110–126.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4), 606–621.
- Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5), 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 (1). *The Journal of the Acoustical Society of America*, 114(6), 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(04), 484–503.
- Schwienhorst, K. (2012). *Learner autonomy and CALL environments*. Routledge.
- Shen, W., Olive, J., & Jones, D. (2008). Two protocols comparing human and machine phonetic discrimination performance in conversational speech. In *Proceedings of Inter-speech*.
- Shimogori, N., Ikeda, T., & Tsuboi, S. (2010). Automatically generated captions: will they help non-native speakers communicate in english? In *Proceedings of the 3rd international conference on Intercultural collaboration* (pp. 79–86).: ACM.
- Shinozaki, T. & Furui, S. (2001). Error analysis using decision trees in spontaneous presentation speech recognition. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on* (pp. 198–201).: IEEE.
- Siegler, M. A. & Stern, R. M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1 (pp. 612–615).: IEEE.
- Stern, R. M., Acero, A., Liu, F.-H., & Ohshima, Y. (1996). Signal processing for robust speech recognition. In *Automatic Speech and Speaker Recognition* (pp. 357–384). Springer.
- Swartz, M. L. & Yazdani, M. (2012). *Intelligent tutoring systems for foreign language learning: The bridge to international communication*, volume 80. Springer Science & Business Media.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4), 295–312.
- Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, 14(2), 50–73.
- Tauroza, S. & Allison, D. (1990). Speech rates in british english. *Applied linguistics*, 11(1), 90–105.
- Taylor, G. (2005). Perceived processing strategies of students watching captioned video. *Foreign Language Annals*, 38(3), 422–427.
- Thompson, I. (1995). Assessment of second/foreign language listening comprehension. *A guide for the teaching of second language listening*, (pp. 31–58).
- Thomson, R. I. & Derwing, T. M. (2014). The effectiveness of l2 pronunciation instruction: A narrative review. *Applied Linguistics*, (pp. amu076).

- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173–180).: Association for Computational Linguistics.
- Trancoso, I., Serralheiro, A., Viana, C., Caseiro, D., & Mascarenhas, I. (2007). Digital talking books in multiple languages and varieties. In *3rd Language & Technology Conference, Poznan, Poland*.
- Tsao, Y.-C. & Weismer, G. (1997). Interspeaker variation in habitual speaking rateevidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research*, 40(4), 858–866.
- Tsubota, Y., Kawahara, T., & Dantsuji, M. (2004). Practical use of english pronunciation system for japanese students in the call classroom. In *Proc. ICSLP*, volume 15 (pp. 1689–1692).
- Tyler, M. D. (2001). Resource consumption as a function of topic knowledge in nonnative and native comprehension. *Language Learning*, 51(2), 257–280.
- Underwood, M. (1989). *Teaching listening*. Addison-Wesley Longman Ltd.
- Van Doremalen, J., Cucchiarini, C., & Strik, H. (2009). Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1), 973954.
- Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, 24, 3–25.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language teaching*, 40(03), 191–210.
- Vandergrift, L. (2011). Second language listening. *Handbook of research in second language teaching and learning*, 2, 455.
- Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT journal*, 42(4), 272–281.
- Vanderplank, R. (2010). Déjà vu? a decade of research on language laboratories, television and video in language learning. *Language teaching*, 43(01), 1–37.
- Vasilescu, I., Adda-Decker, M., & Lamel, L. (2012). Cross-lingual studies of asr errors: paradigms for perceptual evaluations. In *LREC* (pp. 3511–3518).
- Wang, D. & Narayanan, S. (2005). An unsupervised quantitative measure for word prominence in spontaneous speech. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1 (pp. I–377).: IEEE.
- Webb, S. (2010). Using glossaries to increase the lexical coverage of television programs. *Reading in a foreign language*, 22(1), 201.
- Weber, A. & Broersma, M. (2012). Spoken word recognition in second language acquisition. *The encyclopedia of applied linguistics*.

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

Wingfield, A., Poon, L. W., Lombardi, L., & Lowe, D. (1985). Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *Journal of gerontology*, 40(5), 579–585.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, 14(1), 65–86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, 97(1), 254–275.

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

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics*, 18(1), 49–68.