January 2015

THE RELATIONSHIP BETWEEN ACOUSTIC FEATURES OF SECOND LANGUAGE SPEECH AND LISTENER EVALUATION OF SPEECH QUALITY

Mengxi Lin
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Mengxi Lin

Entitled

THE RELATIONSHIP BETWEEN ACOUSTIC FEATURES OF SECOND LANGUAGE SPEECH AND LISTENER EVALUATION OF SPEECH QUALITY

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Alexander Francis
Chair

April Ginther

Margie Berns

Mary Niepokuj

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Alexander Francis

Approved by: Elaine Francis 6/12/2015

Head of the Departmental Graduate Program Date

THE RELATIONSHIP BETWEEN ACOUSTIC FEATURES OF SECOND LANGUAGE SPEECH AND LISTENER EVALUATION OF SPEECH QUALITY

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Mengxi Lin

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

To my parents, Jinjv Li and Shisong Lin, and my husband, Wei Zhang

ACKNOWLEDGMENTS

Undertaking this Ph.D. is truly a life-changing experience for me. I am deeply indebted to the many individuals who have supported my work and continually encouraged me through the writing of this dissertation. Without their time, encouragement, and patience, I would not have been able to see it through.

First and foremost, I would like to express my special thanks to my academic advisor, Dr. Alexander L. Francis, for being a tremendous mentor throughout the years. Alex is definitely one of the smartest persons I have ever met, and I have learnt so much from his enormous knowledge, sharp insight, and unique perspectives. I greatly appreciate his constant encouragement, thoughtful feedback, and selfless time and care, all of which allowed me to grow as a research scientist.

I am also greatly appreciative to Dr. April Ginther for serving as my committee member, for sharing the OEPT speech samples that made this study possible, and for providing invaluable advice and feedback to my research. April has been generously extending her research expertise throughout this work, and I also owe her so much for her personal support.

My thanks also go to Dr. Margie Berns and Dr. Mary Niepokuj, for serving on my committee and for their great support and assistance at all levels throughout the writing of this dissertation. I would like to thank Margie for helping me development my background in sociolinguistics, and to thank Mary for the insightful comments and suggestions on my research project.

I gratefully acknowledge the Purdue Research Foundation for granting me the funding for my Ph.D. dissertation research. My appreciation also goes out to the Linguistics Program of Purdue University, for all the academic, financial, and administrative assistance throughout my Ph.D. study. Moreover, I would like to thank all the students who took their time to participate in my experiments.

My gratitude is extended to my friends in the Purdue Linguistics Program, Yuanyuan Wang, Charles Lam, Chun Zheng, and Chuck Bradley, who have been always so helpful in numerous ways. The past five years would have been different without the stimulating discussions and all the fun we had together.

Finally, a heartfelt thank to my loving family. I am thankful to my dearest parents, Jinjv Li and Shisong Lin, for their unfailing support and all the sacrifice they have made for me. I am also fortunate to have met, dated, and married my husband, Wei Zhang, during the doctoral program. He has been always at my side, good times and bad, and is the greatest supporter of my academic pursuit. I am greatly indebted to him for everything.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

EIL	English as an International Language
ELF	English as a Lingua Franca
ESL	English as a Second Language
f0	Fundamental frequency
F1	First formant
F2	Second formant
Hz	Hertz
LFC	Lingua Franca Core
L1	First language
L2	Second language
NASA-TLX	National Aeronautics and Space Administration Task Load Index
OEPT	Oral English Proficiency Test
POA	Place Of Articulation
TOEFL	Test of English as a Foreign Language
VOT	Voice Onset Time
WMC	Working Memory Capacity

ABSTRACT

Lin, Mengxi PhD, Purdue University, December 2015. The Relationship between Acoustic Features of Second Language Speech and Listener Evaluation of Speech Quality. Major Professor: Alexander L. Francis.

Second language (L2) speech is typically less fluent than native speech, and differs from it phonetically. While the speech of some L2 English speakers seems to be easily understood by native listeners despite the presence of a foreign accent, other L2 speech seems to be more demanding, such that listeners must expend considerable effort in order to understand it. One reason for this increased difficulty may simply be the speaker's pronunciation accuracy or phonetic intelligibility. If a L2 speakers pronunciations of English sounds differ sufficiently from the sounds that native listeners expect, these differences may force native listeners to work much harder to understand the divergent speech patterns. However, L2 speakers also tend to differ from native speakers in terms of fluency the degree to which a speaker is able to produce appropriately structured phrases without unnecessary pauses, self-corrections or restarts. Previous studies have shown that measures of fluency are strongly predictive of listeners' subjective ratings of the acceptability of L2 speech: Less fluent speech is consistently considered less acceptable (Ginther, Dimova, & Yang, 2010). However, since less fluent speakers tend also to have less accurate pronunciations, it is unclear whether or how these factors might interact to influence the amount of effort listeners exert to understand L2 speech, nor is it clear how listening effort might relate to perceived quality or acceptability of speech. In this dissertation, two experiments were designed to investigate these questions.

The first experiment was designed to explore the acoustic features that have the greatest impact on listeners' evaluations of L2 speech quality. The speech of twenty

L2 speakers of English varying in proficiency (high and intermediate) and native language (Chinese and Korean) was evaluated by native listeners of American English. Subjective measures (listening effort, acceptability and intelligibility) were compared to the objective measure of word intelligibility, and to acoustic measures of fluency and pronunciation. Results showed that listening effort, acceptability and subjective intelligibility were highly related to one another and to word intelligibility, and were most strongly predicted by a set of fluency measures, including speech time ratio, speech rate, mean syllables per run, silent pause number, and silent pause time. Segmental and suprasegmental acoustic-phonetic properties did not predict subjective speech quality. These results suggested that fluency may effectively differentiate proficiency levels among relatively advanced L2 learners.

The second experiment was designed to further address the question of whether increasing fluency may reduce listening effort and improve the perceived intelligibility and acceptability of L2 speech when phonetic pronunciation remains constant. To this end, the fluency of the intermediate-proficiency L2 English speech samples used in the first experiment was increased by removing all non-juncture silent and filled pauses. The original and manipulated speech samples, as well as the high-proficiency L2 English speech samples, were evaluated by native American English listeners in terms of listening effort, intelligibility, and acceptability. Each listener's working memory capacity was also measured. Results show that the manipulated speech received significantly higher ratings on all three measures compared to the original intermediate-proficiency speech, and was rated as similarly intelligible and acceptable as the high-proficiency speech samples. It was also demonstrated that listeners of relatively higher working memory capacity expended significantly less effort for processing all speech types and perceived them to be more intelligible than did listeners with lower working memory capacity. These results suggest substantial cognitive benefit of improved fluency on listeners' perception of L2 speech.

Overall, this study suggests that level of L2 fluency plays an important role in predicting listeners subjective ratings, possibly due to the manner in which fluency

modulates listening effort through working memory capacity. These findings further enhance our understanding of the relationship between L2 speech fluency and intelligibility, and will have a direct impact on L2 instruction and assessment.

1. INTRODUCTION

1.1 Preliminaries: Listening to Speech

Speech is one of the most common human activities, and is likely the greatest triumph of the evolution of human kind. It is the vocalized form of human communication that involves a speaker who uses his or her articulatory organs to produce speech units that consist of consonants and vowels, as well as a listener who receives the speech signals, process the acoustic information, and understanding the meaning.

Listening to speech seems to be such an easy task that most of the time we do not even notice any difficulty involved, nor do we not recall how we have learnt it. Yet in other circumstances it can be so difficult that we struggle to understand speech (for example, when listening to speech in noise, or listening to heavily accented speech). Whether our listening experience may be effortless or effortful, listening to speech is essentially a cognitive process that characterizes enormous complexity. As Cutler (2012) elaborates:

> When we are listening, we are carrying out a formidable range of mental tasks, all at once, with astonishing speech and accuracy. Listening involves evaluating the probabilities arising from the structure of native vocabulary, considering in parallel multiple hypotheses about the individual words making up the utterances we hear, tracking information of many different kinds to locate the boundaries between these words, and paying attention to subtle variation in the way words are pronounced, and assessing not only information specifying the sounds of speech–vowel and consonants–but also, and at the same time, the prosodic information, such as stress and accent, that pans sequences of sounds. (p.2)

Despite such complexity, one may wonder why listening to speech remains perceptually easy in most circumstances. One possible answer may be that it depends so much on listeners' previous linguistic experience. For example, listening to a speaker of one's native language (L1) seems undeniably effortless and automatic. This may be partially attributed to the fact that when listening to L1 speech, listeners can recognize speech sounds with high speed and accuracy because most of the sounds are relatively good exemplars of the phonetic categories that they represent. At the same time, listeners are also highly flexible in processing either idiosyncratic variations between talkers (Allen & Miller, 2004; Theodore & Miller, 2010) or systematic variation between dialects (Clopper & Bradlow, 2008; Cutler, Smits, & Cooper, 2005; Evans & Iverson, 2004, 2007) due to long-term experience communicating in the native tongue. Listeners can also efficiently exploit other familiar patterns of the native language, such as prosodic variation, phonotactice constraints, etc., to establish the processing mechanism that assist them to rapidly map speech information to the internalized linguistic knowledge stored in long-term memory, which results in the automaticity of speech processing (Akker & Cutler, 2003; Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012).

On the other hand, such customary sensitivity to the phonetic details of the native language also enables listeners to rapidly detect differences arising from an unfamiliar accent (Ernestus & Mark, 2004; Magen, 1998), and these deviations from listeners' expectation may render the speech more difficult to recognize. Listeners may not be able to re-calibrate their perceptual criteria for phonetic categorization as effectively as their perceptual flexibility for handing L1 variability, simply because they have not yet encountered a wide range of possible forms of speech sounds in their native language carrying various foreign accents (Cutler, 2012). As a result, listeners may need more time to recognize sounds produced by non-native speakers and disambiguate word, not to mention that prosodic variation may further slow down speech processing. As a result, listening to a second language (L2) speaker seems more effortful and requires the commitment of greater cognitive resources for

more controlled processing (Engle & Oransky, 1999). The difference between listening to L1 speech and L2 speech is further amplified in challenging listening conditions. For example, the presence of noise often has been found to affect L2 speech recognition more severely than L1 speech recognition (Bent & Bradlow, 2003; Lecumberri & Cooke, 2006; Lecumberri, Cooke, & Cutler, 2010; Munro, 1998).

When listening to speech is effortful, listeners may judge it as less acceptable and may even find it less intelligible, which may have consequences. For example, in a classroom taught by an L2 English speaker with a heavy accent, students who are native speakers of English may find the course difficult because they are unfamiliar with the instructor's accent. The possibly unfamiliar patterns of phonetic realization of speech sounds may force students to strive to decipher the meaning of the instructor's English utterances before attempting to understand the concepts, which may be difficult to learn in the first place. This scenario, which is not rare in colleges and universities across the US, raises a few questions that are worth serious consideration. First of all, what are the factors that cause this situation? In particular, what is it in the L2 speech that interrupts native English students' listening experience and affects their understanding of the instructor's meaning? Moreover urgently, how to most effectively train the L2-speaking instructors to improve their intelligibility and ease students' listening difficulty? As studies have shown that training can assist listeners to adapt to foreign accent (Bradlow & Bent, 2008; Jongman, Wade, & Sereno, 2003), are there effective ways to help undergraduate English-speaking students establish listening strategies to better adjust to different English accents?

Although these questions appear to be pedagogically oriented, they are essentially related to basic questions addressed by theories of speech production and perception, especially second language speech processing. Research on second language acquisition has been growing rapidly over the past few decades, and a number of models have been proposed in an attempt to account for the production and perception of phonetic segments in L2 speech, such as the Speech Learning Model (Flege, 1995), the Perceptual Assimilation Model (Best & Tyler, 2007), and the Native Language

Magnetic Model (Kuhl & Iverson, 1995). Some of the ultimate questions that these models aim to solve are: How is L2 speech produced, perceived, and processed? Why is listening to L2 speech so different, and most of the times, so much more difficult, than listening to native speech? While a great amount of empirical research has been carried out to address these questions in different ways, most has been focusing on the fine-grained details of segmental production in relation to listeners' perception. More studies are needed to examine the production and perception of other characteristics of L2 speech and how they are related to listening experience and speech intelligibility and acceptability. Such investigations may open a window through which the intricacies of L2 speech production and perception are further disentangled.

1.2 Overview of the study

The growing role of English as a language of communication in today's world means that native speakers of English increasingly find themselves communicating with people who speak English as a second language. While many L1 English listeners can understand the speech of many L2 English speakers, some L2 speech may require more effort to process and may be less readily accepted because it is perceived as being more difficult to understand. Such perception may derive from a variety of factors that may or may not be independent, including both pronunciation and fluency.

Pronunciation is perhaps the most salient aspect in L2 speech that distinguishes it from native speech, and it is often observed that even highly proficient L2 speakers do not achieve native-like pronunciation (Major, 1987, 2001; Scovel, 1988). Thus the degree to which L2 speakers' pronunciation approximates the linguistic forms expected by native listeners will apparently affect listeners' perception of speech quality. If an L2 speaker's pronunciations of English differ sufficiently from those that native listeners expect, these differences may cause native listeners to misunderstand the speech, or at least to have to work harder to understand the divergent speech patterns, making that speech less acceptable than more native-like speech.

Another contributing factor may be fluency – the degree to which a speaker is able to produce appropriately structured phrases without unexpected pauses, self-corrections or restarts. Less fluent speech may also require listeners to work harder to understand the intended message, and previous research has shown that less fluent speech is consistently considered less proficient (Ginther et al., 2010). However, since less fluent speakers tend also to have less native-like pronunciation, it is unclear whether fluency directly affects evaluations of L2 speech quality. That is, is less fluent L2 speech perceived as being less proficient (and thus less acceptable) simply because it is also produced with more divergent phonetic features, or do specific properties of less fluent speech affect evaluations of speech quality independently of phonetic properties, for example by directly increasing listening effort?

Investigating these relationships is important since they may have consequences on listeners' evaluations of L2 speech. In particular, it is important to develop a better understanding of the role of listening effort in how listeners respond to L2 speech, because such speech is often encountered in situations where full attention is already needed for multiple tasks. For example, in an Algebra class taught by an L2 speaker, native English students' attention would be split between listening and learning, such that reducing listening effort may have positive effect on learning.

The overarching questions of the present study are twofold: 1) To identify the acoustic variables related to L2 fluency and pronunciation that have the greatest impact on listeners' subjective evaluation of intelligibility and acceptability as well as of the effort required to listen to L2 speech; and 2) to investigate how improvement in L2 fluency (independently of pronunciation) may contribute to these listener evaluations. Subjective measures of L2 English speech quality (listening effort, acceptability and intelligibility) were compared to the objective measure of word intelligibility by native English listeners, and to acoustic measures of fluency and of phonetic properties related to pronunciation. While previous studies on L2 speech have typically investigated either fluency or phonetic features of pronunciation independently, this study includes acoustic measures relating to fluency and pronunciation simultane-

ously. The ultimate goal is to contribute to understanding the multi-dimensional properties that affect listeners' evaluations of L2 speech quality, in order to provide a basis for developing effective strategies for L2 instruction and assessment.

1.3 Dissertation outline

In addition to the introduction (Chapter 1), this dissertation consists of five chapters and is organized as follows. Chapter 2 presents a theoretical review of the speech quality constructs used in this study, including acceptability, intelligibility, fluency, and listening effort. The purpose of the literature review is to address the theoretical bases of these constructs as well as to introduce various approaches of measurement. It will also review how these methods are applied in empirical studies to explore the multitude dimensions of speech intelligibility, fluency, listening effort, and acceptability. This chapter also presents the research questions investigated in this study.

Chapter 3 reports the methods and results of Experiment I. It provides details on the design of the experiment, the demographic information of the participants, the characteristics of the L2 speech samples, and the procedures of the two experimental tasks. It also introduces the list of acoustic measures that were used to analyze the fluency and phonetic features of the speech samples. Statistical techniques are then presented and the results of descriptive and inferential statistics are discussed in relation to the research questions.

Chapter 4 presents the methods and results of Experiment II. It provides details on the fluency manipulation of the speech samples, the design and procedure of experiments, and the demographic information of the participants. Data analysis of this experiment focused on comparison of subjective evaluations between the manipulated speech samples and the original speech samples. Further analysis was carried out by dis-aggregating the dataset based on participants' working memory capacity index. Statistical results are discussed in relation to the research questions and in light of existing literature.

Chapter 5 summarizes the study by a discussion on theoretical implications and pedagogical applications. It also reflects on the limitations of this study and presents thoughts for future research directions.

2. LITERATURE REVIEW

The phenomenon of second language speech has been extensively studied for over half a century, and generally speaking, consensus has been reached that different factors are involved in determining how one speaks an L2 and how L2 speech is perceived by listeners. Research on second language speech still lacks consensus on the constructs underlying L2 speech performance, as well as on the most effective ways to assess the quality of L2 speech, either from a theoretical or from a pedagogical perspective. Many studies on L2 speech quality evaluation are distributed across different fields, such as phonetics and psycholinguistics, sociolinguistics, instruction and pedagogy, testing and assessment, or combinations of these. Using different methods, these studies altogether have shed insight on L2 speech from a variety of perspectives.

The focus of this chapter is on the theoretical and empirical work on the many approaches to L2 speech quality evaluation. The chapter is divided into five sections. The first section presents an overview of studies on speech intelligibility, not only focusing on L2 intelligibility but also introducing frameworks and methodologies from related fields of speech sciences, such as speech pathology and information processing. The second section focuses on the concept of fluency in L2 speech production and perception, in particularly the quantification and modeling of L2 fluency. The third and fourth sections offer a brief review of listening effort and speech acceptability respectively, which are two constructs commonly used in speech sciences. The two sections also discuss on how listening effort and speech acceptability may be related to speech intelligibility, and how they may benefit L2 speech research. The fifth section presents the overarching research design and questions of the present study.

2.1 Speech intelligibility

Despite the apparent significance of speech intelligibility in L2 acquisition, to date there is no uniformly accepted definition or even conceptualization of intelligibility across the field (Jenkins, 2000, 2002; Munro, 2008; Nelson, 2011; Pickering, 2006; Sewell, 2010; Smith & Nelson, 1985), nor a standard method of measuring intelligibility. The purpose of this section is not to unify all disagreements so that a consensus can be reached on the definition and measurement of intelligibility, but to explore the nature and multitude dimensions of L2 speech intelligibility in hope of obtaining a more thorough understanding of what it means and entails.

2.1.1 Definition and conceptualization

To simply put, speech intelligibility refers to the match between a speaker's production intention and a listener's response to the speech (Schiavetti, 1992). In this sense, speech intelligibility is regarded as perfect when all the words that a speaker intends to produce are completely understood by a listener. On the opposite, if none of the words that the speaker intends to produce is correctly recognized by the listener, intelligibility is reduced to zero. A continuum of intelligibility is thus developed between the two extremes of zero and perfect, where the key for determining the degree of intelligibility resides in the matching process, i.e., to what extent the words uttered by a speaker is accurately responded by a listener. It is crucial to understand that speech intelligibility concerns the speaker's production and the listener's response, because an array of speaker and listener variables may influence the way speech intelligibility is defined and quantified. The following subsections discuss several different perspectives on what speech intelligibility is and how it may be linked to some relevant concepts.

Intelligibility and effectiveness of communication

Catford (1950) characterizes intelligibility in terms of what he referred to as the "effectiveness of communication". Specifically, intelligibility entails the recognizablility of the linguistic forms produced by the speaker, as well as appropriate response from the listener demonstrating understanding of the meaning by the speaker. He further states that "Intelligibility losses are due to defective selection or execution on the part of a speaker, or to defective identification or interpretation on the part of a hearer, or to a combination of these factors" (p.15). In other words, unintelligibility occurs either because the linguistic form of the utterance is unrecognizable to the listener, or because the utterance lacks effectiveness: Listener's response misaligns with speaker's intention, even if the linguistic form of utterance may be recognizable. An example from Catford (1950) illustrates what it means by ineffectiveness: A speaker intends to say "I dont like the *collar*" when referring to a shirt, but since he pronounces "collar" as /kʌlɔˠ/, the listener interprets the utterance as "I don't like the *color*". Here communication breakdown occurs because listener's response is not consistent with speaker's intention, and therefore it is not effective communication.

Since intelligibility and effectiveness cannot always be easily teased apart, Catford (1950) recommends that *intelligibility* should be used as a cover term to refer to utterances that are both intelligible (recognizable) and effective. In other words, it requires the speaker to produce reasonably good exemplars of the linguistic elements, while it also requires the listener to appropriately identify and interpret these linguistic elements that aligns with the intention of the speaker. With respect to communication in L2, Catford (1950) introduces the notion of *threshold of intelligibility*, which emphasizes the influence of linguistic experience and cultural context on L2 speech intelligibility. Specifically, listeners' familiarity with L2 varieties and speakers' cultural background can help lower the threshold of intelligibility and make L2 speech more accessible. Overall, Catford (1950) is among the earliest researchers whose work points

out the importance of speech intelligibility as a functional index of communicative performance.

Smith's paradigm of intelligibility

The idea that L2 speech intelligibility involves the responsibilities of both speaker and listener becomes especially attractive along with the rise of the sociopolitical theory on world Englishes (Kachru, 1985, 1986, 1992), in which one of the central concerns is how to achieve mutual understanding between speakers of different varieties of English, whether these varieties belong to the inner, outer, and expanding circles. The world Englishes paradigm contends that the traditional view, which places native speaker of English in the special position as the solely legitimate custodian who define and maintain the standards of English, does not reflect the current reality that non-native speakers of English far outnumber native speakers of English. Therefore, a new criterion should be established to make evaluations of a variety of English: It should not depend on how native it sounds, but how intelligible it is. Thus how to define intelligibility becomes even more urgent in the world Englishes context.

Larry Smith, together with his colleagues, has contributed ground-breaking work on the intelligibility of world Englishes (Smith, 1992; Smith & Bisazza, 1982; Smith & Nelson, 1985, 2008; Smith & Rafiqzad, 1979). He proposed a seminal paradigm that defines "intelligibility" (in a broad sense, which means understanding in general) by breaking it down to three components: 1) intelligibility (in a narrow sense), 2) comprehensibility, and 3) interpretability, which are defined as follows:

- Intelligibility in the narrow sense only refers to how listener recognize the linguistic form of the utterances produced by the speaker.

- Comprehensibility refers to listener's ability to understand the locutionary force, i.e., the meaning of the utterances.

- Interpretability refers to listener's ability to understand the illocutionary force, or in other words, what the speaker implicates by the words and utterances.

Among the three components, intelligibility serves as the foundation of comprehensibility and interpretability, while achieving interpretability depends on both intelligibility and comprehensibility.

This theoretical proposal of a three-layer structure of speech intelligibility has also been examined by a number of empirical studies (Smith, 1992; Smith & Nelson, 2008), where quantification of intelligibility, comprehensibility and interpretability is achieved through separate tasks. Typically, a cloze test is used for measuring intelligibility, a multiple choice test for comprehensibility, and a paraphrasing task for interpretability. Findings of these studies demonstrate that intelligibility scores were often higher than comprehensibility and interpretability scores for many L2 speech samples, which supports the argument that intelligibility is the basis for comprehensibility and interpretability, and that recognizing words produced by an L2 speaker does not necessarily guarantee understanding of what the L2 speaker intends to express.

Intelligibility and English as a lingua franca

Another approach that seeks to address the use of English as a language for international communication is the English as a Lingua Franca (ELF) movement (Jenkins, 2000, 2002, 2007; Seidlhofer, 2001, 2005; Seidlhofer, Breiteneder, & Pitzl, 2006). ELF proclaims that both native and non-native speakers of English are linguistically and politically equal members of the international community, and it is inappropriate to label anyone as a "*foreign* speaker of English", which seems to carry a negative connotation. Jenkins (2002) points out that an intrinsic implication of this claim is that instead of selecting a native variety of English as the standard model for L2 learners and users, it is necessary to develop an agreed international norm for all ELF members. The key element of such a norm of English hinges on pronunciation,

because quite often miscommunication among ELF speakers arise from the different phonological features (often L1-influenced) between the interlocutors.

For the purpose of promoting international phonological intelligibility, Jenkins (2000, 2002) proposed what she called the Lingua Franca Core (LFC), which includes a set of phonological features that are believed to be crucial for preventing miscommunication and safeguarding mutual intelligibility among L2 speakers of English. These features are:

- All consonant sounds, except for dental fricatives;
- Vowel length contrasts, such as between /ɪ/ and /i/;
- Initial and medial consonant clusters;
- Nucleus (tonic) stress.

Along with these cores features is the speakers' accommodation skills, i.e., whether L2 speakers are able to adjust the acoustic characteristics of their speech in order to make it more intelligible to L2 interlocultors. With sufficient accommodation skills, L2 speaker may show flexibility in adjusting the pronunciation of the core features directed to the expectations of the interlocutor, a strategy that is likely to enhance mutual intelligibility. The non-core features (such as substitution of the dental fricatives with stops), on the other hand, represent regional variation, and are reported to be less likely to endanger intelligibility and impede cross-cultural communication (Jenkins, 2000, 2002).

Finally, Jenkins (2002) points out that instead of a pronunciation model, LFC only serves as a set of guidelines for communication between English speakers, especially between non-native speakers of English. She also admits that the features listed in LFC needs constant fine-tuning. For example, Pickering (2009) offered experimental evidence that pitch variation plays a crucial role in ELF speakers' communication, and accordingly, Jenkins, Cogo, and Dewey (2011) suggests that pitch cues may also be incorporated as a core feature in LFC. This indicates that LFC stresses the

dynamics of communication where negotiation between interlocutors at the phonetic and phonological levels plays a key role.

Intelligibility, comprehensibility and accentedness

While many previous studies seem to conflate L2 intelligibility with L2 accent, other research suggests that intelligibility and accent are two independent, though related, concepts. While increasing intelligibility may be a goal for all speakers to enhance communication, accent is sometimes intentionally reserved for the purpose of identity preservation (Pennington & Richards, 1986). While there are occasions where accent may impinge on communication, it does not always do so (Derwing & Munro, 2009).

A series of work by Derwing and Munro (Derwing & Munro, 1997, 2005; Munro, 2008; Munro & Derwing, 1995a; Munro, Derwing, & Morton, 2006) examining L2 speech intelligibility have demonstrated that intelligibility is independent from accentedness and comprehensibility. In these studies, intelligibility is identified as the extent to which a speaker's utterance is understood by a listener, comprehensibility as listener's estimation of the difficulty in understanding the utterance, and accentedness as the degree to which the pronunciation of an utterance differs from listener's expectation. In other words, intelligibility is about the amount of understanding, comprehensibility is about the effort of listening, and accentedness is about differences. Methodologically, both accentedness and comprehensibility are subjectively measured on Likert scales, and intelligibility is objectively measured through tasks such as dictation or comprehension question. Among the major findings of these studies is the partial separation of intelligibility from accentedness, since L2 speakers were often rated as perfectly intelligible yet heavily accented. However, L2 speakers who received low intelligibility scores were always rated as heavily accented. Comprehensibility scores were typically correlated with intelligibility ratings, but tended to be lower. Accentedness ratings were usually significantly lower than the ratings

for intelligibility and comprehensibility, suggesting that accent does not necessarily interfere with listeners' comprehension of the content of the speech.

The way Derwing and Munro operationalize "intelligibility" apparently differs from Smith and his colleagues. In particular, while intelligibility and comprehensibility denote the understanding of linguistic units and meaning respectively in Smith' framework, the two concepts seems to conflate in Derwing and Munro's definition of intelligibility. At the same time, Derwing and Munro redefine comprehensibility such that it represents listeners' estimation or expectation of the difficulty of listening, while this cognitive aspect is not represented in Smith's framework (nor that of ELF). However, it is noteworthy that in the empirical studies carried out by Smith and colleagues as well as Derwing and Munro, intelligibility was often measured via a transcription task, such as a cloze test or a dictation task, suggesting that intelligibility in the two frameworks may share some similarity, and in particular that findings from empirical research may be comparable.

Relating intelligibility to language attitude

The fact that intelligibility is different from but related to accentedness (Derwing & Munro, 1997, 2005; Munro, 2008; Munro & Derwing, 1995a; Munro et al., 2006) suggests two possible causes of reduced L2 intelligibility: Either this difficulty solely arises from listeners' difficulty processing the phonetic and phonological characteristics of L2 speech, or it attributes to, at least partially, the interaction between accent and listeners' subjective attitudes towards L2 accents.

Previous studies investigating the effect of language attitude on L2 speech have demonstrated that L2 speakers are often labeled with various stereotypes because of their accent (Brennan & Brennan, 1981; Cargile, 1997; Nesdale & Rooney, 1996; Rubin & Smith, 1990). For example, Nesdale and Rooney (1996) reported that when Australian children were asked to evaluate native Australian English, Italian-accented English and Vietnamese-accented English, they assigned lower status to

the two accented varieties of English in comparison with their native variety. Rubin and Smith (1990) showed that English-speaking students not only held negative attitudes towards accented English spoken by international teaching assistants, but they also believed that the instructors who speak accented English were lacking desirable teaching skills.

Furthermore, bias against L2 English is observed not only among native English listeners but also among L2 listeners as well (Chiba, Matsuura, & Yamamoto, 1995; Matsuura, Chiba, & Yamamoto, 1994; McKenzie, 2008). In an investigation of Japanese students' attitude towards American English versus six Asian accents of English, Matsuura et al. (1994) reported that American English received much more positive attitudinal ratings than the L2 accents. Moreover, it seems that the more a participant aspired to acquire a native-like accent of English, the more they showed prejudice against L2 varieties, including Japanese-accented English.

Language attitude towards L2 speech, held by both native and L2 listeners, is found to have a impact on intelligibility and comprehensibility. In a study investigating the effect of L2 accent on listening comprehension, Major, Fitzmaurice, Bunta, and Balasubramanian (2002) noted that positive attitude towards L2 speech was associated with increased comprehensibility and negative attitude with decreased comprehensibility. The comprehension test yielded quite interesting results: Chinese-accented English received lower scores than Japanese-accented English by Chinese listeners, Japanese-accented English received lower scores than Chinese-accented English by Japanese listeners, while Spanish-accented English received a much higher score than other two accents by both listener groups. Major et al. (2002) argued that this result could be partially explained by Chinese and Japanese students' negative attitude of their own English accents.

However, van Rooy (2009) argues that positive attitude does not always guarantee better intelligibility. In her empirical study on intelligibility and perception of English proficiency, the South African listeners' positive attitude towards Korean-accented English did not translate to lower threshold of intelligibility. In contrary, Korean-

accented English received fairly low intelligibility scores even though listeners reported great ease understanding the Korean speakers was high. van Rooy (2009) thus drew the conclusion that positive attitude is a necessary but not sufficient condition for improved intelligibility and comprehensibility.

Intelligibility and familiarity

L2 speech intelligibility is also potentially subject to the influence of listeners' familiarity with L2 accents, as well as listeners' language background and prior linguistic experience (Gass & Varonis, 1984; Levis, 2006; Smith, 1992). For instance, Gass and Varonis (1984) found that listeners' familiarity with non-native speech in general has a positive effect on listeners' ability to comprehend a non-native speaker. Moreover, familiarity with a particular accent also has a facilitating effect on listeners' comprehension of the L2 speech in that particular accent.

The effect of familiarity also manifests when an L2 speaker addresses an L2 listener who share the same L1, as is suggested by Bent and Bradlow (2003) and Hayes-Harb, Smith, Bent, and Bradlow (2008) that listeners may gain "interlanguage speech intelligibility benefit" when their native language is the same or similar with that of the speaker's. For example, Bent and Bradlow (2003) and Hayes-Harb et al. (2008) reported that L2 listeners found the English speech samples produced by highly proficient L2 English speakers from the same language background to be as intelligible as the speech samples produced by native speakers of English. This is probably because these L2 listeners might speak English with the same accent, and were thus highly familiarized with the accent of the high-proficiency L2 English speech samples.

The interlanguage speech intelligibility benefit may also has a positive effect on L2 comprehension, as, for example, Smith and Bisazza (1982) reported that Japanese listeners could complete comprehension tasks better when listening to Japanese-accented English speech than when listening to native English speech. However, this benefit was not observed in Major et al. (2002), whose Japanese participants found

the English speech samples produced by Japanese speakers the least comprehensible, suggesting that interlanguage speech intelligibility benefit does not necessarily apply universally, and might be highly dependent on the proficiency levels of the L2 speakers. Nevertheless, one should remain cautious when interpreting the results of this study due to methodological restrictions. For example, the selection of speakers may be biased because of the strict criteria and it is likely that they did not best represent the accent carried by the populations.

Furthermore, Munro et al. (2006) examined how native and L2 listeners from different L1 background evaluate L2 speech samples produced by speakers who may or may not share the same L1 with the listeners. No intelligibility benefit was identified in L2 listeners' evaluation of speakers from the same L1, and it was suggested that whether interlanguage intelligibility benefit exists or not depends on a complex interaction of an array of listener and speaker variables that certainly needs more in-depth investigation.

Intelligibility: Who counts as the judge?

The study of Munro et al. (2006) leads to another well-debated issue in the literature of L2 speech intelligibility, that is, who should be the listeners making decision on L2 speech intelligibility? While native English listeners are often conveniently drawn as participants of empirical studies on L2 speech intelligibility, this approach has been criticized in particular because it seems to suggest the superior status of native English listeners as the judges of other English speakers (Smith & Rafiqzad, 1979). It also ignores the possibility that L2 English speech may be more intelligible to the ears of L2 listeners than to native listeners. Therefore, it is suggested that a more thorough understanding of L2 speech intelligibility depends on listeners with more diverse linguistic background, both native and non-native (Berns, 2008; Smith & Rafiqzad, 1979). Nevertheless, this approach may introduce new complexities especially with respect to experimental design. For example, how to select the best

representation of listeners? Should listeners' language background match that of the speakers'? Or should listeners characterize as diverse L1s as possible regardless of the L2 varieties represented in the speech samples?

Additionally, Munro et al. (2006) provide evidence that the significance of listeners' linguistic background may not be as pronounced as is assumed. In their study, native English listeners and non-native listeners (who are native speakers of Cantonese, Japanese and Mandarin Chinese) rated speech samples produced by L2 English speakers of Cantonese, Japanese, Polish and Spanish. Inter-rater reliability between the native and non-native listeners was high, which means that ratings of intelligibility, comprehensibility, and accentedness assigned by non-native listeners did not differ much from those assigned by native listeners. The benefit of a shared L1 was observed in some listener groups, but the effect was only minor. Munro et al. (2006) suggests that it is the intrinsic properties of the speech samples, rather than listener' linguistic background, that are of paramount importance in determining L2 speech intelligibility. That being said, this conclusion is only tentative, and more studies comparing native and non-native listeners' responses to L2 speech intelligibility are still needed.

Finally, it should also be noted that many previous studies differ in the context of English use when addressing L2 speech intelligibility (Derwing & Munro, 2005), some being set in an English as a second language (ESL) context versus some in an English as an international Language (EIL) context. This makes it inappropriate to compare across studies because different inferences could be drawn between the two contexts. In an ESL context, L2 English speakers need to make themselves understood by an audience of primarily native speakers of English, whereas in an EIL context, speakers encounter a wider range of interlocutors, the majority of whom are likely to be also L2 English speakers. While achieving intelligibility is certainly critical in both contexts, L2 speakers may adapt different speaking strategies in order to suit different audience and environments. Hence the selection of listeners when examining

L2 speech intelligibility should partially depend on the context. In the present study, focus will be placed on L2 English speakers living in an ESL context.

2.1.2 Measurements of intelligibility

Various techniques have been developed to measure intelligibility, which can be roughly grouped into three categories, targeting at the auditory perceptual, linguistic, and acoustic dimensions of speech intelligibility, respectively.

Perceptual approaches: the scaling procedure

The perceptual measures of intelligibility involve listeners' subjective evaluation of speech at a holistic and impressionistic level using a scaling procedure. It is perhaps the most straightforward way of measuring intelligibility, and has been implemented via different techniques. Each technique is based on certain assumptions and is designed for specific situations.

Stevens (1999, 1951) outlined four types of scaling measurement for assigning numerals to objects and events, which are: 1) Nominal, 2) ordinal, 3) interval, and 4)ratio measures. All four types of scales have been used to investigate intelligibility in different studies.

The nominal level of measurement is achieved through classifying objects and events into mutually exclusive categories, under the assumption that these categories are of equal status. When applied to measuring speech intelligibility, an example of the nominal scale would be to ask listeners to categorize speech samples as either "unintelligible" or "intelligible". Stevens (1999) points out that while the nominal scale is the least restrictive method of assigning numerals, it is the most restricted in terms of applying statistical operations. The only permissible statistical analyses include counting number of cases, obtaining mode, and in some conditions, using contingency correlation to test hypothesis on the distribution of cases among the categories.

The ordinal level of scaling differs from the nominal scale in that it involves the operation of rank-ordering during numeral assignment, where objects or events are mapped onto a hierarchy of descriptive labels from less to greater values or vice versa. An ordinal scale for measuring speech intelligibility is often based on the degree of intelligibility, such as "totally unintelligible, somewhat unintelligible, neutral, somewhat intelligible, totally intelligible". Because the ordinal scale characterizes order preservation, it not only allows statistical operations such as frequency and mode, but also median and percentiles. Other commonly used statistics such as mean and standard deviation are, strictly speaking, inappropriate for an ordinal scale simply because the successive intervals on such a scale are not necessarily equivalent. For instance, it is difficult to prove psychologically that the distance between "totally unintelligible" and "somewhat unintelligible" is the same as "somewhat intelligible" and "totally intelligible". Stevens (1999) warned that although computing mean and standard deviation with an ordinal scale seems to be a common practice and has yielded fruitful outcomes, researchers should be cautious, if not completely outlawing these statistics, of the possibly inaccurate inferences drawn from them.

The interval scale, argued by Stevens (1999) as the true quantitative scale, is operationalized by assigning objects and events to a scale of equal intervals. Classical examples of the interval level of measurement include the scales of temperature and time, where linear transformation can be applied to change a value from one scale to another (such as from Fahrenheit to Celsius). Theoretically, many descriptive statistics can be applied to an interval scale, such as mean and standard deviation, although sometimes it is unclear whether a so-called interval scale is truly interval. For example, in speech-related research, interval scales (such as a 7-point Likert scale) are typically used to quantify perceptual evaluation of speech intelligibility, despite of the extreme difficulty of partitioning human perception along a scale of equal sizes. Indeed, Stevens (1999) claims that most of the widely used psychological scales are essentially ordinal, while only in a few occasions the attempt of equalizing units of a perceptual scale succeed, mostly because the characteristic of the object or event

follows a normal distribution. A typical example is the human intelligence index scale: Although fundamentally ordinal, it mathematically approximates an interval scale since human intelligence is normally distributed.

Finally, the ratio level of measurement is assessed by the estimation of ratios between objects or events on the value of the property that is measured. The major advantage of a ratio scale is that it permits almost all types of statistical operations (Stevens, 1999). In intelligibility research, ratio scale is also often experimentally implemented as the direct magnitude estimation procedure. Different from the other scales, this method does not confine listeners with defined points or intervals, but requires them to directly judging a speech sample by estimating its perceived magnitude of intelligibility compared to other speech samples. It typically provides listeners with a standard or modulus speech sample, which may be numbered 10 or 100, and listeners are required to scale each of the subsequent speech samples with a number that is proportional to the intelligibility of the modulus. The experimenter is responsible for deciding on a modulus, which can represent either the high, mid, or low range of the intelligibility continuum. An alternative is to offer no modulus but ask listeners to assign any number to the first speech sample they hear. As the speech samples accumulate, listeners make judgment of a newly heard speech sample based on its perceived ratio to all previous ones. The major disadvantage of this approach is that it can expose experimental results to vast intra- and inter-rater variability, and make it difficult to interpret the estimated ratios given the large individual variability among listeners (Southwood & Flege, 1999).

From a statistical point of view, the ideal level of scaling measurement is the ratio scale, and when practical limitation bars its implementation, an interval scale is preferred. The least desired are the ordinal and nominal scales, primarily because of the restricted options of statistical tests and potential problems with inference of the results. Essentially, reliable application and interpretation of the rating scales depends they yield a normal distribution. On the other hand, feasibility is a major concern when researchers design an experiment, especially that circumstances may

sometimes preclude the possibility of using the interval and ratio scales. In general, the rule of thumb for using the scaling procedure to measure speech intelligibility is: Whenever practical applicability permits, the higher the level of measurement is, the better.

Linguistic approaches

The linguistic approach to measuring intelligibility is based on the assumption that computing the quantity of information transfer can reliably reflect the degree of intelligibility (Schiavetti, 1992; Yorkston & Beukelman, 1984). By providing a percentage of speech information that listeners understand by means of identification or transcription tasks of linguistic units such as words or sentences, this approach my offer more insight into how linguistic units relate to speech perception / production than the scaling technique. In practice, the linguistic approach can be implemented using various testing designs.

Intelligibility tests using isolated words

The rationale of intelligibility tests at the word level is based the assumption of speech perception theories that speech signals are processed in a "bottom-up" manner: Phoneme recognition tends to precede word identification and sentence comprehension in the listening process. Therefore, many tests are designed to evaluate segmental intelligibility at the word level. These tests are usually administered by having listeners hear audio-recorded target words either in isolation or embedded in semantically neutral carrier sentences such as "Please write (target word) now" (Institute, 1989), and responses are elicited in the forms of word identification ("what word did you hear?") or verification ("Did you hear rake or lake"), or word transcription ("Write down all the words that you hear"). Intelligibility scores are subsequently computed as the count or proportion of correct responses.

Word identification or transcription tests are often used as diagnostic tests for evaluating the quality of speech synthesis or text-to-speech systems, since segmental accuracy is one of the primary criteria for assigning the quality of synthetic speech. Commonly used tests include the Harvard phonetically balanced (PB) words, the modified rhyme test (MRT), the diagnostic rhyme test (DRT), consonant Identification (CI), and the polysyllabic and polymorhphemic word test (Francis & Nusbaum, 1999; Schmidt-Nielsen, 1995). Some of the tests have established standard word corpora, while many researchers also compile their own lexical inventories that are tailored to specific research purposes. The following introduces two lexically oriented tests that could potentially be adapted for L2 intelligibility study.

The *Harvard Phonetically Balanced (PB) word test* was first proposed by Egan (1948) for general hearing tests. It features phonetically balanced words in the sense that each phoneme occurs with approximately the same frequency, so that the test is not biased due to differences in informational load of the phonemes (the more frequently a phoneme occurs, the less informational load it carries). The corpus of the test consists of 1000 words, which are divided into 20 lists and each list contains 50 words. All are monosyllabic English words in the same phonotactic structure, namely, the consonant-vowel-consonant (CVC) pattern. Prior to the test, listeners are typically provided with a chance to familiarize with the 1000 words, although it is argued that excluding the training session does not significantly affect experimental results (Francis & Nusbaum, 1999).

Another word-level intelligibility test is the *Minimal Pair Test*, which was designed on the reasoning that minimal pairs can help identify pronunciation problems related to intelligibility. While the test has primarily been used to assess disordered speech (Ansel & Kent, 1992; Boothroyd, 1985), it can be easily adapted to evaluate L2 speech intelligibility, as long as the adaptation take into account the major difference between these two types of speech. Specifically, reduced intelligibility in speech disorders has a pathological origin, but less intelligible L2 speech may be caused by very different factors, such as onset of learning, exposure to target norms, L1 transfer, to name a

few. These factors should be addressed when researchers design words of minimal pair for assessing L2 intelligibility.

One example that illustrates proper adaptation of the minimal pair test to study L2 speech intelligibility is Rogers and Dalby (2005). In this study, segmental intelligibility of L2 English speech produced by Mandarin Chinese speakers was examined using minimal pair tests by focusing on differences between the L2 production and a specific L1 norm (American English). Prior to composing the minimal pair lists, the researchers first investigated the English speech samples produced by a group of Chinese speakers and identified a list of acoustic-phonemic contrasts that typically deviated from the phonemic representation of American English. It included an array of production difficulties with both consonants (place of articulation, manner of articulation, voicing) and vowels (height, tenseness, and presence of diphthongs), based on which a list of minimal pairs was developed. Another group of Chinese learners were recorded reading the minimal-pair word list, and a panel of native listeners of American English were recruited to transcribe the words. Results showed that word intelligibility scores accounted for 76% of the variance of the same speakers' sentence intelligibility scores, suggesting L2 speech intelligibility may be partially explained by phonemic differences between speakers' segmental production and listeners' expectations. This study illustrates one way to incorporate the minimal pair test into research on L2 speech intelligibility, especially when the focus is at the segmental level.

Finally, a note about word-level intelligibility tests is that they are usually carried out using monosyllabic words of relatively simple phonotactic structure, such as the CV and CVC patterns, but multisyllabic words are not well represented. Nevertheless, an intelligibility score derived from a test that consists of only monosyllabic words will not reflect how successfully an L2 speaker is at producing multisyllabic words, especially how they produce appropriate cues to lexical stress in combination with the segments. Moreover, multisyllabic words may be processed by listeners differently from monosyllabic words, because more contextual cues are provided as the word un-

folds, which can facilitate word recognition. Future studies comparing the processing of monosyllabic and multisyllabic words can certainly extend our understanding of word intelligibility in L2 speech.

Intelligibility tests using words in sentences

Recognizing words in isolation is considered to be more difficult than recognizing words in sentences, because in addition to the segmental and prosodic features of the target word, more contextual cues are also provided to assist to word identity (Nusbaum & Pisoni, 1985). However, when it comes to L2 speech, an L2 speaker who can produce intelligible words in isolation is not necessarily capable of producing good sentential prosody that listeners expect. Indeed, L2 intelligibility may even suffer from abundant yet misleading contextual cues, especially when sentential prosody is so poor that it may direct the listeners to identify wrong words. Therefore, testing L2 intelligibility at the sentence level may be more informative than testing intelligibility at the word level.

The common practice to test speech intelligibility at the sentential level is to record speakers who read aloud a set of pre-designed sentences, which are then transcribed by a group of listeners. Similar to word-level tests, intelligibility is quantified as the percentage of correctly transcribed words.

One widely used method to measure sentence-level intelligibility is the *Key Word Identification Test*, which builds on the assumption that speech intelligibility primarily relies on the recognition of key words (mostly content words) instead of every single words produced by the speaker. This is particularly true in real life communication, where listeners are often under time constraint or in sub-optimal listening condition (e.g. with background noise or over the phone) and processing key words becomes a more effective way of listening. There are two famous sentence corpora for assessing sentence-level speech intelligibility: 1) the Harvard sentences, which were developed together with the PB word list by Egan (1948); and 2) the Haskins sentences (Nye & Gaitenby, 1973).

The Harvard sentences consist of a group of single-clause sentences that are meaningful but unpredictable. This means that listeners can only understand the meaning of the sentence after hearing it entirely, but understanding a fraction of the sentence does not help predict the overall meaning. For example, a sentence like *These days a chicken leg is a rare dish* can be easily understood by normal-hearing listeners, but a phrase extracted from it such as *These days a chicken leg* will not help listeners to anticipate the content of the rest of the sentence. In other words, the Harvard sentences are not as semantically predictable as a sentence such as *He likes his coffee with cream and sugar*, where it is easy to predict *with cream and sugar* after *He likes his coffee* (Francis & Nusbaum, 1999). The goal of the Harvard sentences is to control listeners' real-world and semantic knowledge so that it does not confound with speakers' actual level of intelligibility. Each Harvard sentence typically contains five key words, and intelligibility score is calculated as the percentage of correctly transcribed key words.

Different from the Harvard sentences, the Haskin sentences aim to eliminate any influence from real-world knowledge in the process of speech recognition. These sentences are completely grammatically acceptable but semantically uninterpretable, such as *The old corn cost the blood.* The purpose of the design is to ensure that listeners' performance is solely based on the acoustic characteristics of the speech signal as well as listeners' knowledge on morphology and syntax. Like the Harvard sentences, intelligibility is computed as the percentage of correctly transcribed content words.

Key word identification technique seems useful for studying L2 speech intelligibility at different levels, in addition to the convenience of the readily availability of exiting sentence corpora. However, this technique is not without problems. First of all, speech materials are usually recorded via highly controlled production experiments where speakers are instructed to read pre-structured sentences instead of formulating natural utterances. This approach may introduce certain clear speech effect because speakers in reading tasks tend to show more prosodic variation than

in normal speech (Munro, 2008), and tend to make more exaggerated or even unnatural articulatory movement than they would normally do (Kwiatkowski & Shriberg, 1992). As a result, it may guise some pronunciation problems in casual speech such as final consonant deletion, syllable deletion, stopping and vowel neutralization (Dyson & Robinson, 1987; Klein, 1984; Morrison & Shriberg, 1992). Meanwhile, speech materials recorded by reading tasks may be perceptually unnatural and unrealistic to the listeners, especially when presented with semantically illogical sentences. It is unclear whether these uninterpretable sentences will exert negative influence on the word recognition process. Finally, orthography may also interfere with reading tasks, as it may lead to artificial errors such as mispronunciation or hypercorrection (Munro, 2008).

Intelligibility tests using natural speech

Methodological problems associated with intelligibility assessment using speech materials in citation form or elicited by reading tasks give rise to the proposal of using spontaneous speech materials from more natural scenarios (Kwiatkowski & Shriberg, 1992). Such speech materials are typically elicited by engaging speakers in tasks such as picture description, personal narratives (Munro & Derwing, 1995a), or by recording natural speech in lectures or interviews (Brodkey, 1972). Intelligibility measurement using these speech materials is typically carried out by instructing a panel of listeners to transcribe the recordings sentence by sentence. The transcriptions are then compared with the intended utterances and intelligibility score is computed as the ratio of correctly transcribed words over the total number of words.

A more challenging type of natural speech is conversation, which presents difficulties for reliable measurement of intelligibility since it is much more unpredictable and less manageable. In particular, conversation often features incomplete sentence structure, turn taking, interruption, question and confirmation, and sometimes overlapping by multiple talkers, all of which add extra difficulty to the transcription task and may underestimates the actual level of intelligibility of the speakers.

Nevertheless, analyzing conversational speech may prove useful in certain circumstances. For example, Kwiatkowski and Shriberg (1992) argues that from a clinical perspective, conversational speech captures the momentary variability of intelligibility that is typical in children with developmental speech disorders. Flipsen (2006) proposed to use syllable count to estimate the level of intelligibility in the population of children with speech delays. Since these children often produce long strings of unintelligible utterances, it is difficult to calculate the number of word they produced. Instead, it is much easier to obtain syllable count because syllable nuclei can be easily detected from acoustic signals that characterized peak of sonority or relative loudness. Once the number intelligible syllables and the number of unintelligible syllables are obtained, researchers can count the number of syllables per word (SPW) in the intelligible portion of the speech sample, and use it to yield an estimate of the number of words in the unintelligible portion. This approach designates every speaker as his or her own control, and is particularly useful for diagnosing problems of unintelligibility on an individual basis.

The SPW technique seems most appropriate for assessing low-intelligibility speech where contextual cues are limited, which makes it possible to adapt the measure for diagnosing L2 speakers of low intelligibility. However, caution also arises because L1 and L2 acquisition are known to differ in many ways. For example, whereas children with speech disorders often produce unrecognizable segments, many unintelligible L2 speakers characterize clear segmental production yet poor prosody. It is unclear how effective SPW may be for the evaluation of L2 speech intelligibility, and future research addressing this issue is certainly warranted.

Acoustic approaches

In contrast to both perceptual and linguistic measures of intelligibility, the acoustic measures of speech intelligibility focus on fine-grained acoustic-phonetic properties of single segments or simple acoustic correlates that are associated with enhanced or

reduced intelligibility. It is based on acoustic analysis of speech characteristics using waveforms and spectrograms by acoustical analysis software such as Praat (Boersma & Weenink, 2013).

Examining the physical properties of speech sounds is important for understanding speech intelligibility because phonetic categories are cued by an aggregate of interrelated acoustic correlates, and each correlate has a different weight in cuing phonetic categorization (Coleman, 2003; Francis, Baldwin, & Nusbaum, 2000; Jongman, Wayland, & Wong, 2000; Abramson & Lisker, 1964). For example, the perception of English stop voicing is related to 16 acoustic parameters, among which voice onset time (VOT, defined as the length of time that passes between the release of a stop consonant and the onset of voicing) and fundamental frequency at the onset of voicing (onset $f0$) are the two primary cues for the categorization of stop voicing (Lisker, 1986). The relative weighting of these cues are known to be language-specific. Research on the relationship between VOT and onset $f0$ suggests that while non-tonal language listeners tend to employ VOT and onset $f0$ as the primary and secondary cues for voiceless stop identification, tone language speakers suppress the use of onset $f0$ as a cue to stop consonant voicing, because they tend to prioritize $f0$ information for tonal identification (Francis, Ciocca, Wong, & Chan, 2006; Xu & Xu, 2003). Native speakers of a language typically have acquired these L1-specific cue-weighting patterns in earlier stage of life and are thus able to produce phonetic segments that align with listeners' expectation.

Deviation from native representations of acoustic cue-weighting patterns is often observed in L2 speech. For instance, Zhang, Nissen, and Francis (2008) examined the acoustic cues of English stress produced by L2 speakers of Mandarin Chinese, and found that compared to native English speaker, the L2 speech shows significantly higher $f0$ and different vowel formant. Trofimovich and Baker (2006) also illustrated that the production of several suprasegmentasl features by Korean learners of English differed from native English speakers, such as stress timing, peak alignment, speech rate, pause frequency, and pause duration. These differences may potentially affect

the intelligibility of L2 speech, since listeners are extremely sensitive to the acoustic cue-weighting patterns in their native language and are hence acute to fine acoustic variations (Idemaru & Holt, 2011).

How acoustic attributes of segments are related to overall intelligibility of disordered speech has been extensively investigated in the field of speech pathology, where a range of acoustic and articulatory features are found to be associated with speech production deficits with dysarthric and hearing-impaired patients (Weismer, Martin, & Kent, 1992). One of the major findings is that vowel quality plays a crucial role in predicting intelligibility: Low intelligibility is typically accompanied by reduced vowel space. More specifically, dysarthria patients' speech usually characterizes reduced or even collapsed vowel space (Zlegler & Von Cramon, 1983), difficulty with the front-back distinction of vowels (R. Kent & Netsell, 1978), reduction of the range and slope of formant transition (R. D. Kent, Weismer, Kent, & Rosenbek, 1989), and inappropriate length of vowel duration (Caruso & Burton, 1987). Other commonly reported problems include hypernasality (R. Kent & Netsell, 1978), difficulty with the distinction of the manner and place of obstruent consonants (R. Kent & Netsell, 1978), and reduced formant transition in the production of glide and liquid (Weismer, Kent, Hodge, & Martin, 1988).

Studies on normal speech intelligibility and speaker variability have also revealed the contribution of acoustic-phonetic properties to speech intelligibility, including not only segmental attributes but also a number of suprasegmental properties. For example, Bond and Moore (1994) reported that word duration, vowel duration, and vowel space were the major acoustic-phonetic properties that differentiated between speakers of high and low intelligibility. Bradlow, Torretta, and Pisoni (1996) found that $f0$ range and vowel space expansion (especially F1 range) were highly correlated with overall speech intelligibility. Focusing on the intelligibility of vowels in clear speech, Ferguson and Kewley-Port (2007) demonstrated that vowel intelligibility was significantly improved when vowel space was expanded and vowel duration lengthened.

Similar to findings from the speech disorder research, these studies also collectively suggest that vowel space is critical for speech intelligibility.

However, divergent results were reported in other studies. For example, Hazan and Markham (2004) investigated speech intelligibility of normal-hearing children and adults using not only acoustic measures shown to influence intelligibility by previous studies such as $f0$ range, word duration and vowel space, but also two new variables: Long term average spectrum and consonant-vowel intensity ratio, which were believed to reflect voice dynamics and articulatory precision respectively. Contrary to findings in prior studies, Hazan and Markham (2004) reported that vowel space was not correlated with intelligibility, but the most predictive variables were word duration and the total energy in the 1- to 3-kHz frequency band. Hazan and Markham (2004) attributed the difference between their study and others partially to the difference in the speech materials used as stimuli: Single words were used in their study in contrast with sentence-length materials used in previous studies, which may characterize more contextual variation.

While most studies on phonetics characteristics and intelligibility focus on segments such as vowels and stop consonants, a few have attempted to explore the relationship between fricative production and speech intelligibility. For example, Todd, Edwards, and Litovsky (2011) compared the spectral peaks and means of /s/ and /ʃ/ produced by children with cochlear implants (CIs) and normally hearing children, finding that children with CIs typically exhibited reduced contrast between /s/ and /ʃ/, which partially contributes to their reduced intelligibility. Maniwa, Jongman, and Wade (2009) examined how fricative production relates to the intelligibility of clear speech and conversational speech using 14 spectral, amplitudinal, and temporal parameters. Results showed that fricatives in clear speech featured longer duration, higher spectral peaks, and higher spectral means and skewness, suggesting that the production of clearly intelligible fricatives involves systematic acoustic-phonetic modifications. The major shortcoming of studies on the intelligibility of fricatives is that they focused exclusively on fricatives, and therefore it remains unclear how fricatives

are related to other segmental and suprasegmental characteristics. Exploring this question, in particular how acoustic variables collectively influence speech intelligibility, is certainly of both theoretical importance and pedagogical significance.

Statistical analyses of acoustic measures

Studies examining speech intelligibility often use acoustic variables to predict holistic intelligibility ratings. A potential problem undermining this approach is that speech signals often abound in redundant acoustic attributes, and therefore it is difficult to infer whether the predictive power of acoustic correlates to intelligibility ratings is independent between different variables, or whether it is the aggregation of multiple acoustic properties that collectively influences speech intelligibility. Indeed, high intercorrelations are often observed among acoustically-measured variables associated with intelligibility (Liu, Tseng, & Tsao, 2000; Monsen, 1978; Nickerson & Stevens, 1980).

One way to deal with the intercorrelation problem is to compare the relative weighting among the variables using the regression technique. For example, in a study on the speech intelligibility of deaf adolescents, Monsen (1978) measured nine acoustic variables related to both consonants and vowels (the VOT difference between /p/ and /b/, /t/ and /d/, /k/ and /g/, presence of nasal and liquid production, spectral range of F1 and F2 and F2 variation associated with the diphthong /ai/, mean f0, and sentence duration). Pearson correlation analysis showed that many of the nine acoustic variables were highly correlated with each other, and they were subsequently submitted to a multiple regression analysis with intelligibility score as the dependent variable. Three of the nine variables (VOT difference between /t/ and /d/, F2 difference between /i/ and /o/, and the ability to produce nasals and liquids) were found to account for 73% of the total variance of intelligibility scores, meaning that they had the greatest impact on speech intelligibility. A more recent study by Liu et al. (2000) examined a list of acoustic features (F1 frequency, F1 and F2 frequency locations, the VOT difference between /p^h/ and /p/, /t^h/ and /t/, /k^h/

and /k/, frication duration, possibility of initial burst, nasality, and burst spectrum) of speech samples produced by Mandarin-speaking young adults with cerebral palsy in relation to intelligibility. The study also identified high intercorrelation (ranging from 0.60 to 0.92) among these acoustic variables, which were then regressed on overall intelligibility scores. Results showed that that F1 and F2 frequency locations, VOT differences and the presence of initial burst explained 74.8% of variance in subjective intelligibility scores.

Metz, Samar, Schiavetti, Sitler, and Whitehead (1985) further pointed out that intercorrelation among acoustic variables may reflect the operation of some underlying fundamental dimensions of the speech mechanism. For example, the high correlation among VOT differences between voiced and voiceless stops at three places of articulation in English arise because the same speech-motor control mechanisms govern the timing of laryngeal and supralaryngeal articulation in all three cases, and therefore it may be expected that these VOT variables do, indeed, individually have similar predictive power for intelligibility. If the three variables are entered into a regression analysis simultaneously, it is highly likely that only one of them would emerge as a significant predictor, but not the other two because of their shared variance with the significant VOT measure. This might lead one to wrongly conclude that the statistically significant VOT variable strongly predicted speech intelligibility while the other two did not.

A better approach suggested by Metz et al. (1985) is to use the factor analysis procedure, a statistical technique that can identify mutually uncorrelated latent variables that represent the fundamental dimensions of the original acoustic attributes. Using factor analysis with a set of 12 acoustic measures (including VOT differences, vowel formants, and suprasegmentals such as pure-tone average and sentence duration) extracted from hearing-impaired speech materials, Metz et al. (1985) identified a few mutually uncorrelated latent variables which were argued to be better representations of the underlying dimensions of speech mechanism that may affect intelligibility than the original acoustic measures. Subsequent multiple regression analysis suggested

that the first two factors could significantly predict intelligibility. This factor was interpreted as reflecting the temporal and spatial control for segmental events, and the secondary factor as reflecting speech prosody and the stability of production for the temporal aspects of certain segmental events.

In short, previous studies suggest that speech intelligibility might depend on a combination of different acoustic-phonetic characteristics, but that the evaluation of individual acoustic-phonetic properties may not tell the whole story. While these findings may be further extended to the investigation of L2 speech, there still exists a gap in the investigation of the relationship between fine-grained phonetic features and global evaluation of L2 speech quality. Lastly, it should also be kept in mind that acoustic description of speech intelligibility also marks a critical step towards automated speech recognition and intelligibility evaluation.

2.1.3 Summary

A review of the literature on speech intelligibility reveals that intelligibility is not simply about how many segments or words produced by a speaker can be recognized and understood a listener, but it is a multi-dimensional entity that involves an array of factors, which are even more intricate with respect to L2 speech. The complex nature of speech intelligibility justifies the various approaches of measurement, despite that they may sometimes produce divergent or even conflicting results. Moreover, the application of these measures in L2 studies is still limited. An attempt to triangulate different measuring techniques may be a first step towards building an explanatory model for L2 speech intelligibility.

2.2 Fluency

While L2 speech tends to differ from L1 speech at the phonetic level, it is also typically less fluent than L1 speech. For normal-speaking adults, L1 speech production is such an automated procedural ability all linguistic information can be processed

rapidly and without much effort (Levelt, 1993; Schmidt, 1992). This allows speakers to concentrate on the planning of the speech content rather on the linguistic forms of the production, hence resulting in fluent utterances. In contrast, achieving L2 fluency is a more difficult task because the phonological and syntactic processes are not as automatically encoded in L2 speech production as in L1 speech production. L2 speakers often have to spend much more time and effort formulating the linguistic structure prior to speaking, not to mention the effort spent on content planning and speech monitoring. All these additional processes may slow down their speech rate and may cause various disfluency problems that can ultimately affect intelligibility. Therefore, it is important to examine the nature and properties of L2 speech fluency as an avenue to further enhancing our understanding of L2 speech production and perception. The following sections will review the definitions and measurements of L2 fluency, as well as the various theoretical frameworks that are proposed to model the cognitive and psycholinguistic process. Findings from previous research on L2 fluency is also briefly surveyed and summarized.

2.2.1 Definition of L2 fluency

Fillmore (1979) described fluency as one's overall ability to speak a language. Specifically, he conceptualized fluency in terms of four dimensions. The first dimension refers to speakers' ability to produce utterances with minimum amount of pauses, hence representing the quantitative or temporal aspect of fluency. The second dimension refers to the syntactic and semantic coherence of production, which clearly reflects the qualitative aspect of fluency. The third dimension deals with a speaker's ability to appropriately use language in various contexts, whether familiar or unfamiliar. The fourth dimension points out that a fluent speaker should be able produce speech not only at ease but also creatively, such as making jokes, expressing humor, describing ideas in metaphor, and so on. Despite of the comprehensiveness of

Fillmore's definition of fluency, the major drawback is that it confuses fluency with overall language proficiency, while it is also difficult to quantify the four dimensions.

Based on Fillmore's work, Lennon (1990b) and Lennon (2000) proposed to define fluency in a broad sense as well as a narrow sense. Broadly speaking, fluency is a cover term for a speaker's global oral proficiency, including syntactic complexity, discourse coherence, semantic appropriateness, pausing patterns, lexical choices and language creativity. The narrow sense of fluency, on the other hand, focuses on the temporal facets of speech production. Specifically, fluency is defined as the amount of speech produced in a given time as well as the smoothness and rapidity of the utterances (Lennon, 1990b; Towell, Hawkins, & Bazergui, 1996; Wood, 2001). Another working definition of fluency proposed by Lennon (2000) specifies fluency as "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (p. 26). In recent L2 literature, fluency is operationalized in its narrow sense, because 1) it is measurable, and 2) the broad definition is too easily identified with general language proficiency which in turn remains vague and difficult to quantify (Chambers, 1997; Fulcher, 1987). Following the lead of previous literature on L2 speech, the present study approaches fluency in the narrow sense.

While the term fluency is typically used to refer to capabilities of the speaker, it can also be conceptualized from the standpoint of both the speaker and the listener. Segalowitz (2010) proposed a three-pillar structure for fluency, including cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency refers to the cognitive operation that governs speech production. Utterance fluency is the actualization of cognitive fluency, or in other words, it describes the articulatory features that are reflective of cognitive fluency. Perceived fluency represents listener's inference of the speaker's cognitive fluency through perception of utterance fluency. Methodologically, Segalowitz (2010) proposed that utterance fluency can be acoustically measured in terms of temporal variables, and perceived fluency can be inferred by listener's subjective ratings, while the combination of utterance and perceived fluency measures

are assumed to estimate cognitive fluency. This approach of viewing fluency from the angles of both production and perception has been gaining popularity over the years (Bosker, Quene, Sanders, & de Jong, 2014).

2.2.2 Temporal measures of L2 fluency

Temporal measures provide a gateway to evaluate L2 fluency objectively and offer insight into the cognitive mechanism underlying L2 acquisition. An array of temporal measures has been developed to investigate L2 fluency (de Bot, 1992; Towell, 1987; Towell & Hawkins, 1994), many of which are adapted from studies on L1 acquisition (Goldman-Eisler, 1958, 1968; Grosjean, 1980). Roughly speaking, L2 fluency measures can be grouped into three types, focusing on the 1) quantity, 2) rate, and 3) disruption of speech, respectively (Blake, 1996; Ginther et al., 2010).

Temporal measures on the quantity of speech

L2 fluency measures on the quantity of speech include total response time, articulation time, speech time, and speech time ratio (or phonation time ratio). Total response time refers to the total length of a given speech sample, including both meaningful speech and disruptions such as repetition and hesitation, and pauses. Articulation time is the duration of time when the speaker is articulating sounds, whether meaningful or un-meaningful. Speech time refers to the time that is used in producing meaningful information. Comparing the three measures, articulation time differs from total response time in that it excludes all silent pauses, and speech time further excludes all pauses and disruptions. Together the three measures reflect the quantity of speech production at different levels.

Speech time ratio refers to the percentage of speech time over total response time. As an indicator of the proportion of fluent speech, speech time ratio is considered to be more reliable than measures on absolute speech quantity in predicting a speaker's level of fluency. High speech time ratio is believed to reflect great ease of language

formulation and articulation, and low speech time ratio indicates difficulties in speech production (Lennon, 1990b; Towell et al., 1996; van Gelderen, 1994).

Temporal measures on the rate of speech

Temporal measures of fluency on the rate of speech include speech rate, articulation rate, and mean syllable per run. Speech rate is computed as the total number of syllables divided by total response time, and articulation rate is computed using the total number of syllables divided by articulation time. These two measures are considered reliable indicators of the speed and efficiency of speech. Previous studies have found that speech rate and articulation rate are often positively related to levels of L2 proficiency (Ginther et al., 2010; Kormos & Denes, 2004), and longitudinal studies have demonstrated that improvement in L2 fluency typically characterizes an increase in speech rate (Freed, 1995; Towell, 1987; Towell et al., 1996).

The measure of mean syllable per run is based on a different rationale compared to speech rate and articulation rate. It refers to the average number of syllables produced during a continuous stretch of speech between two silent pauses, and is computed using the total number of syllables divided by total number of runs in a given speech sample. A run is defined as a continuous chunk of utterances between two silent pauses, while a silent pause refers to a period of silence that is equal to or longer than 0.25 seconds. Previous studies have shown that an increase in mean syllable per run can reflect L2 speakers' improved ability to formulate complex syntactic structures with appropriate phonological encoding, as well as the improved ability to easily and effortlessly access to lexicon and activate vocabulary (Ginther et al., 2010; Towell et al., 1996).

A remaining question related to mean syllable per run is why the threshold for a silent pause is 0.25 seconds. As a matter of fact, this is only an arbitrary cut-off point, upon which previous research disagreed. While many studies settled on 0.25 seconds (Goldman-Eisler, 1968; Towell, 1987; Raupach, 1987), others used different

criteria, such as 0.2 seconds (Kormos & Denes, 2004), 0.28 seconds (Towell, 2002), 0.3 seconds (Raupach, 1980), and 0.4 seconds (Derwing, Rossiter, Munro, & Thomson, 2004). Furthermore, Griffiths (1991) picked 0.1 seconds and 3 seconds as the lower and upper limits respectively, and Riggenbach (1991) and Riggenbach (2000) used 0.5 seconds to 3 seconds as the two boundaries. Indeed, to determine the cutoff point for a silent pause is somewhat tricky. On the one hand, if the cutoff point is too low, then silences that are innate in speech, such as short breaks between syllables or words, may be identified as silent pauses originating from fluency problems. On the other hand, if the cutoff point is too high, then some silent intervals that are caused by speakers' difficulties in language processing and formulation may not be accurately captured. de Jong and Bosker (2013) examined various thresholds for silent pauses and demonstrated that acoustic measures based on the threshold of 0.25 seconds for silent pauses have the highest correlation with L2 proficiency. Therefore, the present study also adopts the 0.25-second criterion as the cutoff point for silent pauses.

A related measure to mean syllable per run is pruned syllable per second (Derwing et al., 2004; Derwing, Munro, Thomson, & Rossiter, 2009). Pruned syllables exclude all disfluencies, such as filled pauses, hesitation, and repetition, and therefore pruned syllable per second is essentially a measure on the speed of meaningful speech production. In this regard, pruned syllable per second is similar to mean syllable per run.

Temporal measures on the pauses

Pauses, or phenomena related to disruption of speech, such as silence and fillers (e.g., "uh", "um", and small words like "you know", "I mean"), as well as disfluencies such as repetitions, repairs, and restarts, are informative of a speaker's level of fluency and overall language proficiency (Fillmore, 1979; Lennon, 1990a). Although L1 speech is intuitively expected to be maximally pause-free, studies show that even fluent L1 speech abounds in pauses (Deese, 1980; Goldman-Eisler, 1968). In fact, pauses

constitute an indispensable part of L1 speech as they facilitate smooth transition between thought planning and speech production, and therefore the number or degree of pauses alone do not necessarily index level of L1 fluency. Part of the reason that L1 speech is often perceived to be fluent, despite the presence of pauses, may be that native speakers' pauses follow native language pausing conventions and are thus regarded as appropriate by listeners (Sajavaara, 1987).

Studies on pausology have identified two major types of pausing phenomena: Juncture pauses and non-juncture pauses. Juncture pauses refer to the type of pauses located at sentential or phrasal boundaries that are therefore syntactically predictable. Non-juncture pauses are those located within syntactic constituents (Deschamps, 1980; Lennon, 1984). Juncture pauses constitute an indispensable part of L1 speech production (Goldman-Eisler, 1968; Riggenbach, 1991). For example, in English, pauses are required at clausal boundaries accompanied by appropriate intonational contours, as long as they are shorter than about 2 seconds (Wood, 2001). That is to say, although L1 speech can be highly fragmented, these fragments are not random. The pauses produced by L1 speakers tend to be juncture pauses which not only allow the speaker sufficient time to plan the following speech unit, they also allow listeners to use the time to process the preceding unit. Therefore, juncture pauses usually do not interfere with listeners' perception of speakers' fluency. In contrast, L2 speakers tend to produce more non-juncture pauses (pauses within sentence and clause boundaries), which may reflect the need for more time to plan and process speech in an L2 (Clark & Tree, 2002; Ginther et al., 2010). Additionally, filled pauses at non-juncture positions may also reflect L2 speakers' affective state of anxiety (Goldman-Eisler, 1968).

The primary temporal measures for assessing disruption of speech include the duration, frequency, and distribution of silent pauses and filled pauses. Additionally, ratio measures such as pause duration over total response time are frequently used. Studies comparing L1 and L2 fluency have shown a higher ratio of pausing time over total response time in L2 speech as compared to L1 speech (Ginther et al.,

2010; Riazantseva, 2001). Moreover, L2 speakers of higher oral proficiency tend to display shorter pause duration and lower pause frequency than L2 speakers of lower proficiency (Freed, 1995; Lennon, 1990a; Riazantseva, 2001). Longitudinal studies further suggest that the percentage of silent pause time over total response time dwindled as L2 speakers became more proficiency in the second language (Lennon, 1990a).

Previous studies have shown that the distribution of pauses within and between clauses may distinguish L1 from L2 speech as well as proficiency levels within L2 speech. For example, non-juncture pauses are more common in L2 speech than in L1 (Lennon, 1984; Raupach, 1980), L2 speakers who paused more at clausal junctures and less within clauses were perceived to be of higher fluency than those who produced more pauses within clauses (Freed, 1995; Riggenbach, 1991), and improved proficiency reduces the prevalence of non-juncture pausing (Raupach, 1987). It is possible, however, that these effects pertain mostly to speakers with lower proficiency, because Riazantseva (2001) found that Russian speakers of both high and intermediate English proficiency produced similar numbers of silent pauses, and were also comparable to native English speakers in this regard.

Last but not the least, the duration and frequency of pauses displayed in L2 speech may also be influenced by the pause patterns in speakers' L1, since languages are characterized by different temporal and rhythmic patterns (Grosjean & Deschamps, 1975; Holmes, 1995; de Johnson, Oconnell, & Sabin, 1979). For example, cross-linguistic studies showed that native native speakers of French (Grosjean & Deschamps, 1975) and Spanish (de Johnson et al., 1979) typically produce more and longer silent pauses than native English speakers. These L1-specific pausing patterns may play a role in how speakers perform in an L2, at least during certain stages of language learning. In Riazantseva (2001), native Russian speaker of intermediate English proficiency produced longer silent pauses in both Russian and English than native English speakers, indicating that these L2 English speakers might be following their L1 pause-length conventions even when speaking in L2. In contrast, Russian speakers of high English

proficiency produced similar length of silent pauses in their English speech compared to native English speech, suggesting that the effect of L1 transfer may be minimized as L2 fluency improves. These results indicate possible interactions between pausing and language specific timing patterns.

2.2.3 Modeling L2 fluency

Current conceptualizations of L2 fluent speech production are primarily based on the speech production model proposed by Levelt (1993) and Levelt, Roelofs, and Meyer (1999), which was initially developed to describe the process of L1 speech production. Two types of knowledge are essential to Levelt's framework, namely, declarative knowledge and procedural knowledge. Declarative knowledge refers to the knowledge of the world, and procedural knowledge refers to the knowledge necessary for skilled behavior. Levelt points out that fluent speech production typically involves procedural knowledge due to the speed requirements.

Levelt's speech production model proposes three phases of speech production, namely, conceptualization, formulation, and articulation, all of which function independently yet collaboratively. The conceptualizer is responsible for generating the propositional pre-verbal content of the message. The formulator then accesses the lexicon and passes the message for syntactic construction and phonological encoding, while the proceduralization of declarative knowledge also takes place at this stage. The articulator actualizes the phonetic plan into overt speech. No feedback between these stages is allowed so that automaticity can be guaranteed to process the procedural knowledge. For L1 speakers of a language, the operations of the formulator are highly proceduralized so that speech is formulated extremely fast. For L2 speakers, the lack of automaticity may primarily attribute to less proceduralized operations of the fomulator. While the goal of Levelt's model is to describe the different stages of speech production, Towell et al. (1996) noted that it does not account for how fluency

may develop over the course of language acquisition, which limits its application to L2 acquisition.

The Adaptive Control of Thought (ACT) Model (Anderson, 1983), a model on cognitive development, was subsequently introduced to remedy the shortcomings of Levelt's model (Crookes, 1991; Raupach, 1987). The ACT model states that any cognitive performance involves three memory stores: Two are long-term memory capacities and the other is working memory, which is of limited capacity. It assumes that all knowledge is initially declarative, but it can be converted into procedural knowledge through learning. This conversion is necessary for speech performance because processing the two kinds of knowledge demands different level of cognitive effort. Processing declarative knowledge is effortful because it requires much attention, whereas processing procedural knowledge requires minimal attention and is thus rapid and efficient, without overloading the capacity of working memory. As a result, for cognitive behaviors that need to be rapidly performed, such as speech, the conversion from declarative knowledge to procedural knowledge is critical.

The ACT Model posits three stages that account for the learning process of this conversion. The first stage is called the "cognitive stage", which features slow and inefficient processing because it only contains declarative knowledge. At the second stage, or the "associative stage", access to the knowledge is faster because it has been partially proceduralized (although still partially declarative). In third or the "autonomous stage", the declarative knowledge is completely proceduralized such that it can be quickly accessed and autonomously processed by working memory.

Relating the ACT model to language acquisition, the ability to convert declarative knowledge to procedural knowledge directly determines a speaker's ability to produce rapid and smooth flow of speech. Most adults have arrived at the third stage in their L1 production over years of practice, which explains why L1 speech production if often an effortless task. In contrast, the different levels of fluency demonstrated by L2 speaker possibly suggest the different stages they are at in the process of learning the proceduralization skill. L2 speakers who characterize low fluency may still be

at the cognitive stage, and speakers of relatively higher levels of fluency may have entered the associative stage. An L2 speaker who has arrived at the autonomous stage should have mastered the ability of converting declarative to procedural knowledge, and may show native-like level of fluency.

Kormos (2014) proposed a comprehensive model on L2 speech production that built on Levelt's blueprint but also incorporated findings from other theoretical and empirical research. Similar to Levelt's framework, Kormos's model of bilingual speech production consists of three separate modules: the conceptualizer, the formulator, and the articulator. The model also adopts the proposal of ACT that there are three knowledge stores involved in L2 speech production, but Kormos postulates an additional knowledge store that also plays a role in L2 speech production, which is the declarative knowledge store of L2 syntactic and phonological rules. This knowledge store is not necessary for L1 speakers because all the syntactic and phonological rules of the language are already proceduralized and are thus highly automatic, requiring little effort for retrieval. For L2 speakers, especially for those of low and intermediate levels of proficiency, the knowledge on L2 grammatical and phonological rules is not entirely automatized. The fact that there are stored in the form of declarative knowledge introduces an additional proceduralization step, and may in part contribute to reduced fluency. If an L2 speaker masters full automation of the linguistic rules of the second language, then their fluency may not differ significantly from an L1 speaker.

Kormos's proposal that L2 speech fluency is affected by an additional declarative knowledge store of L2 linguistic rules is partially supported by neuroimaging studies. For example, Abutalebi, Cappa, and Perani (2001, 2005) illustrate that L1 and L2 speech processing essentially activates the same cerebral areas for early bilinguals (who are proficient in both languages), but slightly different regions for late bilinguals (who are not as proficient in L2 as compared to their L1). These results suggest that late bilinguals may store the declarative knowledge of L2 syntactic and phonological rules, which is not completely proceduralized, in a different region of the brain. In contrast, early bilinguals did not show any difference because they have autonomized

the syntactic and phonological rules of both languages. It remains unclear whether the neural activities for highly proficient late bilinguals may or may not behave similarly with early bilinguals.

To sum up, the three models introduced in this section enhance our understanding of the development L2 speech fluency in great depth. However, most of the models conceptualize fluency from the standpoint of the speaker, whereas none addresses perceived L2 fluency from the listener's perspective, especially how L2 fluency may affect listener's perceived effort and processing capability. More theoretical and empirical studies are in need to explore the mechanism through which L2 fluency interacts with speech perception.

2.2.4 Development in L2 fluency research

Over the past few decades, a body of literature has been devoted to examine L2 fluency from various perspectives, such as the relationship between temporal aspects of fluency and rater assessment, the development of L2 fluency, comparison between L1 and L2 fluency, and the effect of manipulation on fluency parameters. This section will review and summarize representative major works in the field in light of the measures and models discussed in previous sections.

Among the most groundbreaking research on L2 fluency is Towell et al. (1996), which aimed to apply Levelt's model to explain development in L2 fluency. Specifically, it compared the level of fluency of a group of advanced L2 learners of French before and after a study-abroad program in a French-speaking country. Acoustic analysis of the learners' French narratives showed that they did become more fluent at producing spontaneous speech in French after attending the program, especially in terms of temporal measures such as speech rate, mean syllable per run, and speech time ratio. Quantitative and qualitative analyses revealed that learners' fluency improvement mostly attributed to increased length and complexity of utterances between pauses (i.e., runs), suggesting that it is the formulator, rather than the conceptual-

izer and the articulator, that may have the greatest impact on L2 fluency. However, note that although the three measures increased over time, they nevertheless failed to achieve comparable levels with the learners' L1 (British English) utterances.

Based on Towell et al. (1996),Towell (2002) further expanded the investigation by examining the same population for a longer period (three years) and by dividing participants into two groups of low and high proficiency levels. Similar to Towell et al. (1996), both groups demonstrated increased speech rate, mean syllable per run, and speech time ratio over the span of three years, but the low-proficiency group appeared to show greater improvement in all three measures than the high-proficiency group. Moreover, while all participants more or less managed to reduce the amount of silent pause time, only the magnitude of pause reduction in the low-proficiency group reached statistical significance. More interestingly, qualitative analysis on pause distribution revealed that the improvement observed in the low-proficiency group primarily originated from changes in pause location: Fewer pauses were produced within clauses and more at clausal boundaries. However, since the qualitative analysis was based on the speech performances of only two participants, the reliability of this result remains questionable. Finally, Towell (2002) also pointed out that the divergent individual performances observed in this study may be partially accounted for by differences in speakers' short term memory capacity, which governs their ability of formulating and processing L2 utterances. This suggestion is certainly worthy of further investigation.

While Towell's studies mainly focused on examining fluency from the angle of speaker characteristics, it is unclear how the acoustically measured features of L2 fluency are perceived by listeners. To explore this question, Kormos and Denes (2004) asked a panel of teacher raters (both native and non-native speakers of English) to evaluate the fluency level of L2 English speech samples produced by native speakers of Hungarian of high and low English proficiency. They also measure an array of acoustic features of fluency, as well as a few other linguistic variables such as the stress (measured as the number of stressed words produced per minute), and accuracy

of production (measured as the ratio of error-free clauses over the total number of clauses). The main finding of this study was that speech rate, mean syllable per run, speech time ratio and the number of stressed words produced per minute were all reliable predictors of subjective fluency ratings by both native and non-native raters. This result is consistent with Towell's findings (Towell et al., 1996; Towell, 2002), suggesting that listeners may rely primarily on speed-related features when assessing an L2 speaker's level of fluency. Another informative finding was that although accuracy of production did not significantly predict perceived fluency in the overall analysis, its importance emerged when individual performances were analyzed. For a few individuals, the effect of accuracy even outweighed all the other measures, suggesting that at least for some L2 speakers, increasing fluency depends greatly on the production of more accurate syntactic structures. With respect to pausing patterns, none of the acoustic measures on silent and filled pauses as well as other disfluency phenomena directly affected fluency ratings, nor did the L2 speakers of low and high proficiency differ in terms of pause duration and frequency. Thus the study concluded perhaps the speed of message delivery has a greater influence on listeners' perception of L2 fluency than phenomena related to the disruption of speech.

In line with this argument, Munro and Derwing (2001) focused primarily on the effect of speech rate and investigated how speech rate may predict listeners' assessment of the accentedness and comprehensibility of L2 speech. A curvilinear relationship between speech rate and the subjective ratings was identified in the study. Specifically, it demonstrated that speech rate that is too high or too low both led to decreased ratings of accentedness and comprehensibility, and it also established that the optimal L2 speech rate should be somewhat faster than the average speech rate generally used by the speakers. According to Munro and Derwing (2001), this curvilinear relationship could be explained as follows. L2 speakers may gain benefit from reasonably accelerating speech rate, because this moderate increase may help listeners to overlook the divergent phonetic and phonological differences and expedite speech processing. However, if speech rate becomes too fast, it may overtax listener's

processing resources and hence reduce comprehensibility. In contrast, very slow and tedious speech may also add extra processing load on listeners' short term memory and draws listeners' attention to the L2 phonetic features that differ from L1 representations. Pedagogically, this study suggests that a reasonable increase of L2 speech rate can contribute to improved intelligibility.

Different from the approach using speech rate as a primary indicator of fluency, Ginther et al. (2010) argued that a better understanding of L2 fluency should include examination of not only speech rate but also other aspects related to fluency, such as speech quantity and disruption patterns. To this end, Ginther et al. (2010) examined the relationship between L2 speakers' holistic proficiency scores from a local oral English proficiency test and temporal measures of quantity, rate, and pausing. It was based on a relatively large pool of speakers (150) from three native language backgrounds, including Chinese, Hindi, and English (native English speakers served as the control group). Results show that the holistic testing scores, assigned by trained raters, were strongly or moderately correlated with various temporal measures, such as speech rate, articulation rate, mean syllable per run, and speech time ratio. The holistic ratings were also correlated with silent pause duration and silent pause ratio, but none of the filled pause measures showed statistically significant correlation with overall proficiency rating. Moreover, while the temporal measures of fluency could differentiate L2 speakers of high proficiency from those of low proficiency, these measures failed to distinguish speakers of adjacent proficiency levels. This indicates that still other factors may play a role influencing listeners' evaluations of L2 speech quality, especially when it comes to speakers whose proficiency levels are not easily distinguishable. Therefore, the cluster of variables needs to be further expanded so that we can gain a more thorough understanding of the relationship between utterance fluency and perceived fluency.

Another direction to extend L2 fluency research is to delve into the relationship between L1 fluency and L2 fluency development. Most of the current literature has been concentrating on examining L2 speech fluency alone, unlike in other domains

of L2 acquisition (i.e., segmental acquisition) where L1 transfer is often considered an influential factor. To address this question, Derwing et al. (2009) conducted a longitudinal study by tracing a group of L2 English speakers with Mandarin Chinese and Slavic language backgrounds over a period of two years, during which both L1 and L2 speech samples were collected at different stages. These speech samples were analyzed using three temporal measures: Speech rate, number of pauses, and pruned syllables per second. Consistent with previous studies, all three measures were found to reliably predict subjective ratings of L2 speech samples assigned by trained raters throughout the research project. However, only pruned syllable per second was shown to be a strong predictor of ratings on L1 speech samples. More interestingly, while a strong correlation between L1 and L2 fluency was identified at the earlier stages, the strength of the correlation became weakened along with speakers' improved L2 fluency. In terms of the effect of native language, Derwing et al. (2009) reported that the correlation between L1 and L2 fluency at earlier stages was stronger for Slavic language speakers than for Mandarin speakers. Moreover, the association between temporal measures of fluency and subjective ratings were also stronger for L2 speech samples produced by Slavic language speakers and those by the Mandarin speakers. According to Derwing et al. (2009), these results may suggest certain linguistic benefit for the Slavic language speakers due to parallel syntactic structure between English and Slavic languages, which may ease the load of the formulator and facilitate speech production. Linguistic similarity between speakers' L1 and L2 may also benefit listeners due the match between received speech information and their expectations.

Bosker et al. (2014) examined how L1 and L2 fluency are weighted from L1 listeners' perspective by means of manipulating pause patterns and speech rate. In the first experiment, they manipulated pause frequency and duration of both native and L2 Dutch speech samples, which resulted in three sets of speech samples: The first set excluded all silent pauses; the second set characterized only short silent pauses (250-500 ms); and the third set only contained long silent pauses (750-1000 ms). Additionally, the number of silent pauses was matched up between native and L2 speech. These

manipulated speech samples were assessed by native Dutch listeners. As expected, irrespective of the method of manipulation, native Dutch speech was consistently rated as more fluent than L2 Dutch speech. Interestingly, the effect of manipulation was not different across native and L2 speech: Listeners perceived the speech samples without pauses to be more fluent than those with pauses, and perceived the speech samples containing shorter pauses as more fluent than those containing longer pauses.

In the second experiment, Bosker et al. (2014) manipulated speech rate and articulation rate of the speech samples collected in the first experiment. One the one hand, the articulation and speech rates of the L2 speech samples were increased to match the speed of the original native speech samples. On the other hand, the articulation and speech rates of the native Dutch speech samples were slowed down to match the speed of the original L2 speech samples. Both the original and manipulated speech samples were evaluated by native Dutch listeners, who were instructed to focus exclusively on the temporal dimension. Results, again, demonstrated similar effects between perception of native and L2 speech: The manipulated L2 speech samples received significantly higher fluency ratings, and the manipulated native speech samples received much lower fluency ratings.

Taken together, this study suggests that native listeners may weigh temporal aspects of native and L2 speech in a similar manner, although the fact that they showed consistent preference to the native speech samples indicates that perhaps there are still other factors that are also influential. Finally, this study pointed out that methodologically, the advantage of using speech manipulation to studying L2 fluency lies in that it allow researchers to hold constant all acoustic parameters except the manipulated one/ones, so that any differences between the perception of manipulated and original speech samples could only attribute to the acoustic characteristics that are manipulated. In the present study, we also follow this line of reasoning to justify our speech manipulation (see more details in Chapter 4).

2.2.5 Summary

To sum up, research on L2 fluency converges on the view that acoustic measures of fluency can account for listener ratings to a large extent. However, agreement has not been reached on which specific acoustic variables have the greatest impact on listener judgment. Generally speaking, previous studies have established that L2 speakers tend to speak at a slower rate compared to L1 speakers, and improvement in L2 fluency tends to be accompanied by an increased rate of speech. L2 speakers also tend to show lower speech time ratio (the ratio of meaningful speech production over total response time) than L1 speakers. Finally, studies comparing L1 and L2 fluency have revealed the general tendency of a higher ratio of pausing time over total response time in L2 speech as opposed to L1 speech, and longitudinal studies showed that the percentage of silent pause time over total response time dwindled as L2 speakers became more proficient. Between L2 speakers of high and low oral proficiency, the higher-level speakers tend to display shorter pause duration and a smaller number of pauses than low-level speakers. Nevertheles, generalizations from these interesting findings are obviously limited. More effort is needed to explore the multiple facets of L2 fluency and to relate it with other aspects of L2 speech production and perception.

2.3 Listening Effort

A crucial step for listeners to successfully recognize and understand speech is to map the incoming acoustic signal to a mental representation of linguistic elements (phonemes, words, phrases) stored in long-term memory. One source of effort may be the need to resolve mismatches between phonetic properties of the signal and those of long-term linguistic representations (Nusbaum & Schwab, 1986). When there is a little or no mismatch between features of the speech being heard and those of the listener's long-term memory traces of linguistic categories, then speech perception is accomplished with seemingly little or no effort. On the other hand, L2 speech contains

many more segmental and suprasegmental features that differ from the prototypical L1 patterns of the language, and these differences may be more significant than in native speech. Resolving the more severe and more frequent mismatches between L2 speech and native L1 prototypes increases the effort of L2 speech perception (Rogers, Dalby, & Nishi, 2001; van Engen & Peelle, 2014). In this section, I will briefly introduced the conceptualization and measurement of listening effort, and discuss how it may benefit L2 speech research.

2.3.1 Definition

Listening effort measures the ease or difficulty of listening (Downs, 1982; Hicks & Tharpe, 2002). Specifically, it refers to the amount of cognitive resources that are required to process and understand speech (Gosselin & Gagne, 2011; Zekveld, Kramer, & Festen, 2011). Listening effort has been extensively studied in the context of listening in sub-optimal conditions such as when listening to an accented talker, and in adverse conditions such as in the presence of background noise, reverberation, or hearing impairment. Such non-ideal conditions typically require listeners to allocate more cognitive resources such as working memory or attention to the listening task, a process that is perceived as an increase in listening effort (Kramer, Zekveld, & Houtgast, 2009; Pichora-Fuller et al., 2003).

2.3.2 Measures of listening effort

A range of methods have been proposed to quantify listening effort, which can be generally grouped into three types: subjective, physiological, and psychophysical or behavioral measures.

Subjective measures aim to tap listeners' perceived level of mental effort in a listening task (Feuerstein, 1992; Hicks & Tharpe, 2002), and are typically based on verbal assessments of workload in which listeners explicitly rate how much effort is required to understand or recognize the speech in question (Hällgren, Larsby, Lyxell,

& Arlinger, 2005). Evaluations may have many sub-scales given the multidimentional nature of listening effort, often including ratings of multiple types of demand and assessments of the participant's subjective responses to these demands, and are typically reported to be fairly sensitive to changes in listening effort (Hällgren et al., 2005).

Recently, speech researchers have begun to employ the NASA Task Load Index (TLX) (Hart & Staveland, 1988) for these purposes. The TLX was originally designed to measure the mental workload of user-control interfaces, and asks participants to assess a task using six dimensions: Mental Demand (how mentally demanding was the task); Physical Demand (how physically demanding was the task); Temporal Demand (how hurried or rushed was the pace of the task); Performance (how successful were you in accomplishing what you were asked to do); Effort (how hard did you have to work to accomplish your level of performance); Frustration (how insecure, stressed, or annoyed were you). The categorical extremities of these scales are "very low" on the left end, and "very high" on the right end, except for Performance, for which the scale is labeled as "perfect" on the left and "failure" on the right. These dimensions represent at least somewhat independent clusters of variables selected on the basis of an extensive analysis of factors that may affect subjective workload for different individuals performing tasks of various difficulty and complexity. Subsequent research suggests that specific combinations of these dimensions reliably and validly predict individual workload experience in a broad array of tasks (Hart, 2006; Rubio, Díaz, Martín, & Puente, 2004). For example, Mackersie and Cones (2011) used NASA-TLX to examine listening effort in a competing-talker task, where they identified systematic increase in subjective ratings of the two categories "mental demand" and "effort" along with increased difficulty of task demand. These systematic increases were also consistent with the increased stress experienced by the participants during the experiment. Within the field of speech perception, the Physical and Temporal demand dimensions are typically left out, but response patterns on the remaining items have been repeatedly shown to be closely related with listeners' performance

(Bologna, Chatterjee, & Dubno, 2013; Mackersie & Cones, 2011; Mackersie, MacPhee, & Heldt, 2015; Pals, Sarampalis, & Başkent, 2013).

Subjective ratings of listening effort are reported to be fairly sensitive in detecting changes in listening effort (Hällgren et al., 2005), but this method has its pros and cons. The major advantage of subjective measures is the simplicity of implementation, both laboratorially and clinically. How, these measures are highly susceptible to individual biases, but may also mislead listener to confuse perceived effort with perceived performance. For example, when listening to speech, listeners may not necessarily distinguish between perceived effort and perceived speech intelligibility. Moreover, listeners sometimes report increased listening effort even it is still early in the test session where the task is supposed to incur minimal effort, either because they have not encountered any difficult tasks yet, or because of factors unrelated to the experimental design such as that they just had a long tiring day. The consequence is that as listeners eventually hear the difficult tasks, they may have to re-calibrate their evaluation of listening effort and hence resulting in inconsistent ratings across experimental sessions. All these issues may potentially undermine the validity of subjective measures of listening effort, although not completely discounting their importance.

More recently, objective measures are developed in an attempt to more reliably quantify listening effort associated with changes in processing load, and these measures are mostly physiologically or psychophysically based. Commonly used physiological measures include cortisol level (Hicks & Tharpe, 2002), pupil response (Koelewijn, Zekveld, Festen, & Kramer, 2012), heart rate, skin conductance and temperature, and electromyographic (EMG) activity (Mackersie & Cones, 2011). These measures characterize different levels of sensitivity to subtle variations in listening effort and cognitive demand. For instance, Mackersie and Cones (2011) compared the four measures of heart rate, skin conductance, skin temperature, and electromyographic (EMG) activity, and found that skin conductance and EMG showed a significant positive correlation with changes in listening demand. In contrast, heart rate and skin temperature failed to predict listening effort. Furthermore, analysis of individual

participants showed that the majority of participants exhibited systematic changes in skin conductance along with changes in listening task demand, but none of the individual's EMG data revealed a similar pattern. This suggests skin conductance is likely to be the most sensitive to instantaneous changes in cognitive effort.

Another group of physiological measures on listening effort is related to pupil responses, such as peak dilation amplitude, peak latency, and mean pupil dilation (Beatty, 1982; Kramer, Kapteyn, Festen, & Kuik, 1997; Verney, Granholm, & Marshall, 2004). In particular, pupil response is highly sensitive to fine differences in language complexity: The more complex the listening condition is, the more it will evoke top-down processing, and as a result, the more pupil dilation will be observed (Kramer et al., 1997; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006). Pupil response was also found to be associated with reduced speech intelligibility for both young (Zekveld, Kramer, & Festen, 2010) and old adults with and without normal hearing (Zekveld et al., 2011). Additionally, pupillary measures can also reflect increasing listening effort associated with speech perception in the presence of a single-talker masker (Koelewijn et al., 2012).

The other type of objective measures on listening effort is the psychophysical or behavioral measures, among which the most common technique is the dual-task paradigm (Gosselin & Gagne, 2011; Hicks & Tharpe, 2002; Sarampalis, Kalluri, Edwards, & Hafter, 2009). This paradigm typically requires listeners to perform a primary task (such as speech recognition) and a secondary task (such as visual recognition) concurrently. The underlying assumption is that given limited cognitive resources, listeners will prioritize processing capacity for the primary task, and then allocate the rest of the capacity to processing the secondary task (Kahneman, 1973). If the difficulty of the primary task is systematically manipulated, then monitoring performance of the secondary task will provide information about changes in listening effort. When the primary task features low cognitive load, such as listening in quiet, more cognitive capacity will be available for processing the secondary task, which can be easily completed by the listeners. However, when the primary task requires

high cognitive load, such as listening in noise, the primary task may engage much more processing resources and leaves significantly less capacity for the secondary task (Kahneman, 1973). This often results in poorer performance in the secondary task, and is interpreted as a consequence of the increased listening effort in the primary task (Fraser, Gagne, Alepins, & Dubois, 2010; Hicks & Tharpe, 2002; Sarampalis et al., 2009).

While both the subjective and objective measures of listening effort are commonly used techniques in speech sciences, their relationship remains ambiguous. As a matter of fact, more and more evidence points to the possibility that objective measures of listening effort are not consistent with subjective ratings, as correlation between the two types of measures appears to be either weak or absent (Fraser et al., 2010; Gosselin & Gagne, 2011; Hicks & Tharpe, 2002; Mackersie & Cones, 2011; Sarampalis et al., 2009; Zekveld et al., 2011). When subjective and objective measures are used in the same study, it is often found that they may both reflect systematic variation of performance levels, but in inconsistent ways (Gosselin & Gagne, 2011; Zekveld et al., 2011). This discrepancy suggests that subjective ratings of listening effort and objective physiological or behavioral measures may be assessing different aspects of listening effort (Gosselin & Gagne, 2011). In particular, subjective ratings appear to represent listeners' perception of the ease of listening, but may not be extremely sensitive to changes in cognitive load or processing resources (Feuerstein, 1992). In this sense, objective measures may tap "real" changes associated with listening effort due to high sensitivity to physiological changes. Nevertheless, both psychophysical and physiological methods are time-consuming and require specialized techniques or equipment (or both), while subjective measures are a more practical option. For the present study, subjective measures were deemed most likely to be effective, because they can be used quite simply with the speech samples available.

2.3.3 Listening effort and speech intelligibility

As listening condition becomes more challenging, listening effort tends to increase. Sometimes this increase in effort is accompanied by decreasing speech intelligibility (Mackersie & Cones, 2011; Zekveld et al., 2011), possibly because sub-optimal listening condition may affect listeners ability to decode linguistic information in the incoming speech signal and thus reduce the ease of listening. However, in other cases recognition performance may remain unaffected even while listening effort measures yield significant differences (Gatehouse & Gordon, 1990; Koelewijn et al., 2012; Pittman, 2011). This dissociation suggests that listening effort and intelligibility are different constructs. Even when sub-optimal or adverse listening conditions do not reduce understanding, they may still demand more cognitive resources to achieve the same level performance. One possible explanation is that in sub-optimal listening conditions where the received speech signal is insufficient, listeners are forced to make more use of contextual information as well as their own linguistic knowledge to identify segments and to disambiguate alternatives, which engage more attention effort than in listening conditions where speech signals are of better quality (Pichora-Fuller, 2006). Finally, note that previous studies on the relationship between listening effort and speech intelligibility have been primarily focusing on disordered speech or listening in adverse conditions, but what remains largely missing in the literature is how listening effort is related to L2 speech intelligibility.

2.3.4 Listening effort and working memory capacity

The concept of listening effort is commonly related to working memory capacity (WMC), which refers to one's ability to temporarily store and process information for complex cognitive tasks, such as language comprehension (Baddeley, 1999; Just & Carpenter, 1992). Working memory may be contrasted with long-term memory, which consists of previously learned knowledge that is stored for a long period of time and can be retrieved during performance of cognitive tasks (Baddeley, 1999;

Harberlandt, 1994). Working memory is thought to be actively involved in listening to speech in a challenging acoustic environment (Lunner, 2010). Under optimal listening conditions (such as listening to L1 in quiet), listeners can quickly retrieve information from long-term memory to match up with the incoming speech signal, resulting in a process that is fast and seemingly automatic and effortless and thus facilitating comprehension. On the other hand, when the received signal fails to match representations in the long-term memory (such as when listening to an L2 speaker whose pronunciations are significantly different from those expected by the listener), additional working memory resources must be employed to hold the information and infer its meaning, thus increasing the effortfulness of speech processing. Even when meaning is successfully decoded at a higher cognitive level, it may be accomplished at the cost of high mental effort as a result of overtaxing limited working memory resources (Lunner, 2010; Pichora-Fuller, 2006; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012; van Engen & Peelle, 2014). Furthermore, WMC also tends to vary from person to person, which may lead to variation in performances in complex cognitive tasks, especially those that may require processing resources exceeding listeners' WMC such as listening to L2 speech (Harberlandt, 1994). These individual differences have been found to be robust predictors of performance on demanding tasks (Conway, Kane, & Engle, 2003; Engle & Oransky, 1999; Kane, Hambrick, & Conway, 2005). Specifically, listeners with larger WMC may perform better in cognitively difficult tasks such as listening to L2 since they have more processing resources at their disposal, in contrast to listeners with relatively smaller WMC.

Because working memory resources in general are highly constrained, any recruitment of additional resources for processing L2 speech may have behavioral consequences. To resolve discrepancies between L2 speech patterns and L1 linguistic representations in long term memory, listeners may have to actively listen and think about listening. This engages working memory, which may eventually cause a shortage of capacity for other tasks. The situation is further exacerbated when the L2 speech features disfluencies, because these may force the listener to temporarily hold

incomplete fragments of speech in working memory in order to make sense of the information, further consuming limited WMC. Ultimately, these extra demands on WMC may leave the listener with insufficient capacity for other tasks such as message understanding and formulation of a response.

Interference with completing related tasks may affect the acceptability of unfamiliar speech in a given context, or listeners' evaluation of the overall quality of or preference for the speech (Francis & Nusbaum, 1999). For example, when listening to the synthetic voice of a global positioning system (GPS) device, one not only needs to understand the speech, but more importantly, must be able to drive along the correct route based on the understanding of the instruction delivered by the GPS voice. Increased WMC demand from a poorly intelligible synthetic voice may interfere with driving, making the voice unacceptable in this context. Likewise, when listening to L2 speech, listeners are most likely engaged in other operations, be it learning algebra with an instructor who is an L2 speaker, conducting business transactions with a partner who is L2 speaker, or filing a complaint about a product over the phone with a customer service representative who is also an L2 speaker. In such situations, increased WMC demand caused by the mismatch between L2 speech and listeners' long term linguistic representations also affect its acceptability. Until now, how listening effort affects the acceptability of L2 speech remains an open question.

Yet another unanswered question is what specific aspects in L2 speech may cause listening difficulty. One seemingly obvious answer is L2 speakers' ability to instantiate native-like production of speech sounds. Deviations from native phonetic and phonological patterns in L2 speech may increase effortfulness because listeners have difficulty mapping segmental information onto long-term linguistic representations. However, L2 speech also typically differs from native speech in terms of fluency. It is possible that properties of L2 fluency patterns may also contribute to the efforfulness of listening of L2 speech. Munro and Derwing (2001) have established that L2 speakers gain an intelligibility benefit from reasonably accelerating speech rate such that it is faster than average L2 speech rate. They have argued that this moderate increase

in speech rate may help listeners to overlook the divergent phonetic and phonological differences and facilitate information processing. If L2 speech is produced at a very fast speech rate, it may overtax listener's processing resources and hence reduce intelligibility, while very slow and tedious speech may also add extra load on listeners' working memory and may draw listeners' attention to the L2 phonetic features that differ from L1 patterns. Both speech that is too fast and speech that is too slow may increase listening effort and affect the intelligibility and acceptability of L2 speech.

It is unclear whether another aspect of fluency, i.e., pausing, may yield similar cognitive effects as speech rate. On the one hand, pauses may be a positive speaker strategy for assisting listeners to better process L2 speech by allowing them sufficient time to decode divergent segmental pronunciations. Pausing for this reason may free up working memory resources and facilitate subsequent linguistic and conceptual processing of the recognized speech. On the other hand, pauses in L2 speech might also exert a negative effect by driving listeners to commit working memory resources for temporarily storing large numbers of short runs of speech (many of which may be syntactically incomplete) and thus may leave less capacity for information processing, which in turn could deteriorate intelligibility. Whether the cognitive effect of pausing is positive or negative, listeners with relatively smaller WMC may be more easily affected by pauses when listening to L2 speech than are listeners with larger WMC, because those with smaller WMC have fewer processing resources at disposal to begin with. For these listeners, L2 speech may introduce greater demand on processing resources. If this is the case, then it leads to the question of whether reduction of pauses can affect working memory demand and listening effort, and for that matter, the overall subjective evaluation of the L2 speech.

2.3.5 Summary

Previous literature on listening effort has focused on listeners and listening conditions as the sources of differences in effort. Less is known about the cognitive effort

incurred by listening to unfamiliar speech. L2 speech provides an excellent context for such studies, as native listeners may engage additional cognitive support for understanding non-native speech (van Engen & Peelle, 2014). There are at least two possible, non-exclusive, mechanisms by which L2 speech might increase cognitive demand on listeners. On the one hand, listeners may have to work hard to deduce the speaker's intended words and phrases when the L2 speech includes pronunciations that are significantly different from the patterns that native listeners expect. On the other hand, disfluencies in L2 speech may also increases listening effort because listeners must hold information in memory for longer while waiting for interrupted phrases to be completed. While the benefit of matching L2 pronunciation to native listeners' expectations is well-known, less is known about the effects of fluency on assessments of speech quality. If the presence of disfluencies in L2 speech incurs greater cognitive demand, then improving fluency may free up processing resources, allowing listeners to spend more effort on understanding divergent pronunciations, and thereby improving both intelligibility and acceptability. Part of the goal of the present study specifically aims to determine how acoustic measures of pronunciation and fluency of L2 speech impact listening effort.

2.4 Speech acceptability and overall speech quality

2.4.1 Definition

Speech acceptability refers to listeners' preference of the overall quality of an utterance (Francis & Nusbaum, 1999; Hecker & Williams, 1966). It has primarily been used in the evaluations of synthetic speech quality and speech perception through assistive listening devices such as hearing aids. Compared to other basic speech quality measures such as intelligibility, acceptability is the most global assessment of speech quality, reflecting listeners' subjective opinion of the overall goodness of speech performance (Francis & Nusbaum, 1999).

2.4.2 Measurement

To elicit acceptability ratings, listeners are often presented with utterances and rate their quality either on a potentially interval scale (such as a 7-point Likert scale) or an ordinal scale with labels (e.g., Excellent, Good, Fair, Poor, Bad) (Schmidt-Nielsen, 1995).

More importantly, the assessment of acceptability is usually context-specific, with testing applications designed for various discourses. For example, one test of acceptability may require listeners to evaluate the voice quality of a voice-mail system, and the other may target at assessing the synthetic voice incorporated in a GPS. This variability in context means that acceptability ratings could be easily affected by subjective factors such as listeners' preconceived notion of different discourses (Francis & Nusbaum, 1999), which makes it difficult to compare acceptability ratings across speech systems or contexts.

2.4.3 Relationship between acceptability and intelligibility

Although acceptability is often correlated with intelligibility, the two measure different aspects of speech quality (Francis & Nusbaum, 1999; Schmidt-Nielsen, 1995). Intelligibility refers to the amount of information received by the listener, while acceptability refers to a broader assessment of overall quality, often in a specific context. The two constructs are often related, especially in cases of poor speech intelligibility, acceptability is also typically low. However, there are also cases where even if all words in a passage may be understood (high intelligibility), acceptability may still be poor because of other factors such as perceived foreignness or unnaturalness, or increased demands on listening effort. For example, a sample of highly intelligible synthetic speech that is nevertheless low in acceptability may be poorly incorporated into working memory (Luce, Feustel, & Pisoni, 1983), and in a multi-task situation can result in poor performance on competing tasks (Schmidt-Nielsen, Kallman, & Meijer, 1990). Similarly, hearing aid users may show a strong preference for certain

signal processing techniques even though these do not improve speech recognition, suggesting that such techniques make speech perception less cognitively demanding, and therefore more acceptable, even without improving intelligibility per se (Ricketts & Hornsby, 2005; Sarampalis et al., 2009). Extending this reasoning, assessing acceptability of L2 speech separately from intelligibility and listening effort may help distinguish the subjective factors that influence listener assessment of L2 speech.

2.5 Research questions of the study

Both fluency and pronunciation affect perceived speech quality, but both are highly multidimensional factors and less is known about the specific acoustic features that most strongly affect listeners' evaluations of L2 speech quality. The overarching goal of this study is to determine the role of fluency and phonetic pronunciation in listeners' evaluation of the listening effort, intelligibility, and acceptability of L2 speech. It is further decomposed into the following questions:

1. How does high- and intermediate-proficiency L2 speech differ in terms of fluency and phonetic intelligibility?

2. How do these differences affect listeners' evaluation of the listening effort, intelligibility, and acceptability of L2 speech?

3. How does improving the level of fluency by reducing pauses contribute to subjective ratings of L2 speech?

4. Are fluency-related differences in perceptual ratings of L2 speech dependent on listeners' working memory capacity?

To investigate these questions, the speech of twenty L2 speakers of English varying in proficiency (high and intermediate) and native language (Chinese and Korean) was evaluation by two experiments. Experiment I (Chapter 3) was designed to address the first two research questions. Four listener variables (word intelligibility, global

subjective intelligibility, acceptability, and listening effort) were obtained through assessment of the speech samples by normal-hearing native English listeners. The speech samples were also analyzed in terms of fine-grained acoustic measures of fluency and phonetic intelligibility. Experiment II (Chapter 4) aimed to address the last two research questions. To this end, the intermediate-proficiency English speech samples used in Experiment I was manipulated such that all inappropriate silent and filled pauses were removed to artificially improve fluency. These manipulated speech samples, together with the original high- and intermediate-proficiency speech samples, was evaluated by three groups of native English listeners in terms of listening effort, subjective intelligibility, and acceptability. Additionally, listeners' working memory capacity index was also measured to in order to examine how differences in individual processing capacity may affect subjective evaluations of speech quality.

3. EXPERIMENT I: ACOUSTIC FEATURES OF SECOND LANGUAGE SPEECH RELATED TO LISTENERS' EVALUATION OF SPEECH QUALITY

3.1 Introduction

Second language (L2) speech is typically less fluent than native speech, and differs from it phonetically. While the speech of some L2 English speakers seems to be easily understood by native listeners despite the presence of a foreign accent, other L2 speech seems to be more demanding, such that listeners must expend considerable effort in order to understand it. One reason for this increased difficulty may simply be the speaker's pronunciation accuracy or phonetic intelligibility, while L2 speakers also tend to differ from native speakers in terms of fluency.

This study hypothesizes that deviations from native phonetic and phonological patterns in L2 speech may increase mental effort because when listeners have difficulty recognizing divergent pronunciations, they will have to work harder to deduce the speaker's intended words and phrases. At the same time, disfluent speech makes it difficult for listeners to follow the thread of what a speaker is saying, thus increasing effortfulness as listeners must hold more information in memory for longer while waiting for phrases or sentences to be fully spoken. Improving fluency may free up processing resources, making it possible for listeners to more effectively employ native listening strategies to compensate for the non-optimal (accented) speech signals. Thus, more fluent speech allows listeners to spend more effort on understanding divergent pronunciations, which in turn contributes to greater intelligibility and acceptability. As a result, it is possible that speakers who are more fluent may require less listening effort and be more intelligible without actually producing more native-like speech sounds. Following this line of argument, Experiment I was specif-

ically designed to determine how acoustic measures of pronunciation and fluency of L2 speech impact listening effort, intelligibility and acceptability.

One important clarification before description of Experiment I is why this study did not adopt the measure of comprehensibility, which is frequently contrasted with intelligibility in L2 studies. One the one hand, there is even less agreement on the definition of this term compared to intelligibility. For example, Smith and Nelson (1985) argued that intelligibility refers to the linguistic decoding of words and utterances, which in turn serves as a basis for comprehensibility, defined as the understanding of meaning. On the other hand, Derwing and Munro (1997) used comprehensibility to refer to how easy or difficult it is for an utterance to be understood, a definition more comparable to the term "listening effort" as used in this study.

On the other hand, measurement of comprehensibility also varies. The construct is often rated objectively using comprehension questions or subjectively by rating scale, and is often compared with ratings of intelligibility. For example, Derwing and Munro (1997) and Munro and Derwing (1995a) used a transcription task to measure intelligibility and a 9-point Likert scale (1=extremely easy to understand; 9=impossible to understand) to measure comprehensibility. Their results showed a correlation between intelligibility and comprehensibility, but the relationship was not perfect. They observed cases in which highly intelligible L2 speech was nevertheless not rated as highly comprehensible, suggesting that intelligibility and comprehensibility may reflect different dimensions of L2 speech and thus should be considered distinctly.

Despite these considerations, comprehensibility was not directly assessed in the present study for two reasons. First, all speech samples had extremely similar content, precluding the use of questions about sample content to estimate comprehension, because listeners became increasingly familiar with the content with each successive sample. Second, some characterizations of comprehensibility are extremely similar to that of listening effort as used in the present study, and therefore the goal was to first investigate the utility of assessing a measure that could, in principle, be considered either listening effort or comprehensibility, with the expectation that future

work could be designed to distinguish between these concepts more specifically, if warranted.

3.2 Methods

3.2.1 Speech materials

The speech samples were drawn from a database of the Purdue University Oral English Proficiency Test (OEPT), a test designed to assess international graduate students' qualifications as prospective teaching assistants. It is a computer-based, semi-direct test, in which examinees respond to a variety of questions, present information and speak extemporaneously on various topics. The responses are recorded in a quiet testing room and are evaluated by at least two trained raters using a linear scale of proficiency ranging from 3 to 6, where 3 and 6 indicate lowest and highest level of acceptability, respectively. Examinees who receive a score of 5 or above are eligible for assignments of teaching assistantships.

Twenty de-identified OEPT samples were selected as speech materials of this study. They were produced by native speakers of Chinese (10) and Korean (10). These language groups were chosen because they represented the two largest subgroups of L2 examinees on the OEPT (36% and 11% of all test takes for Chinese and Korean, respectively, in 2013, the year from which the samples were selected). For each language group, five speakers were selected with a score of 5 (highly proficient) and the other five with a 3 (intermediate proficiency). To control for content homogeneity, all the speech samples were responses to the same test question. The duration of the speech samples varied from 83 to 120 seconds.

3.2.2 Listener assessments

Listener assessments of word intelligibility, listening effort, subjective intelligibility, and acceptability were obtained by two tasks. One task measured listening effort,

subjective intelligibility, and acceptability because they all involved rating the speech samples using interval scales. Word intelligibility was assessed in a separate task in which individual words were presented.

Listeners

Thirty native speakers of American English were recruited on a voluntary basis. All participants were undergraduate students in their first or second year of study and the experiments were conducted following a protocol approved by the Human Research Subjects Protection Program at Purdue University. Ten participants (6 women, 4 men; mean age =20.1) responded to a recruitment poster and completed the word intelligibility task. They were compensated at a rate of $10/hour for two hours of participation. The other 20 participants (19 women, 1 man; mean age =19.4) were recruited from an undergraduate course and they received extra credit as compensation for approximately 1 hour of participation in the subjective rating task. None of the participants had a history of speech or hearing disorder by self report.

All participants had studied or were currently studying at least one foreign language. Five participants had studied two foreign languages, one of which was Spanish, and another participant had studied three foreign languages, also including Spanish. Spanish was the most commonly studied foreign language (27 out of the 30 participants), followed by German (2), American Sign Language (2) and Arabic, Danish, Japanese, French, and Latin (all 1, each). None of the participant had studied Chinese or Korean. The average age of the onset of foreign language training was 12.9 years, and the average years of foreign language study was 4.5. All participants reported having experience interacting with L2 English speakers from different L1 backgrounds, including East Asian and European languages. These interactions were reported as having taken place either in a classroom context where the L2 speaker was a teaching assistant, or in social and everyday settings.

Testing methods

For both tasks, experiment sessions were conducted individually. Before beginning the experiment, each participant filled in a background assessment form, compiling variables related to language attitude, language experience, and other potentially relevant personal data.

Word Intelligibility.

Stimuli for the word intelligibility test consisted of all content words (nouns, verbs, adjectives, adverbs) extracted from all speech samples. Function words were excluded because they are generally less important to speech intelligibility. After extraction, all words were amplitude normalized. The stimuli were presented via an E-Prime 2.0 script (Schneider, Eschman, & Zuccolotto, 2002). Upon hearing each stimulus, the listeners typed the word they heard. Trials were self-paced and there was no time limit on responses, but each word was only presented once. Each listener heard and transcribed all content words produced by the 20 speakers. Words were blocked by speaker, but order of words within speaker and order of speakers for each participant were randomized. All transcriptions were scored automatically and results were also manually examined to ensure that obvious typographical errors and homophones were corrected to count as matches. Word intelligibility was subsequently computed as the percentage of correctly recognized words.

In order to familiarize listeners with this experimental paradigm, a practice session was provided at the beginning, where listeners transcribed 57 words produced by a Korean speaker with an OEPT score of 3 who was not among the 20 speakers formally tested. The same practice session appeared at the end of the task after participants finished listening to all twenty speakers, so that comparisons could be made for the same speaker before and after the experiment to determine whether listening to twenty L2 speakers of English yields a learning or un-learning effect.

The maximum length of the word intelligibility test was estimated to be approximately two hours according to pilot tests. To avoid any negative effect of fatigue on listener performance, this task was divided into 2 one-hour sessions and each listener visited our lab on two different days to complete the test. All experiment sessions were conducted individually.

Subjective assessments.

The second group of listeners (N=20) assessed listening effort, acceptability, and subjective intelligibility of the speech samples via six 20-point Likert scales. This included a modified version of the NASA-TLX consisting of only four of the six subscales (Mental Demand, Performance, Effort, and Frustration) with slightly modified questions (Table 3.1) similarly to the modifications introduced by Mackersie et al. (2015). The other two dimensions of the TLX, Physical Demand and Temporal Demand, were excluded because this listening task did not impose any physical or response time demand on participants. The scores assigned to the four questions on listening effort were averaged to provide a general index reflecting the amount of effort listeners estimated they spent to understand the corresponding speech sample.

The other two scales rated intelligibility and acceptability, asking listeners how well they understood the speakers and how willing they would be to accept the speaker as a course instructor, respectively. See Table 3.1 for questions and scale endpoints for all tasks.

Presentation of stimuli was blocked by speaker proficiency to increase the likelihood that participants would use the whole scale. To control for the possible effect of presentation order, proficiency type was counterbalanced across listeners. The entire task lasted about an hour in a single session.

Table 3.1.: Likert scale rating of listening effort, subjective intelligibility, and acceptability.

Assessment	Question	Left End (1)	Right End (20)
Listening Effort	MENTAL DEMAND: How mentally demanding was it to understand this person's speech?	Very Undemanding	Very demanding
	PERFORMANCE: How successful were you understanding the message in this speech?	Very Successful	Very Unsuccessful
	EFFORT: How hard did you have to work to understand the speech	Very Little	Very Hard
	FRUSTRATION: How insecure, discouraged, irritated, stressed, and annoyed were you?	Not Annoying At All	Very Annoying
Subjective Intelligibility	How clear was the person's speech?	Very Easy to Understand	Very Hard to Understand
Acceptability	How willing would you be to accept this speaker as your TA?	Very willing	

3.2.3 Acoustic measurements

Nineteen acoustic measurements related to fluency and intelligibility were obtained for each speech sample using Praat 5.3.51 (Boersma & Weenink, 2013). These included 13 measures related to fluency (Table 3.2), and six measures of segmental and suprasegmental acoustic-phonetic features (Table 3.3).

Acoustic measures related to fluency.

The 13 acoustic measures related to fluency (Table 3.2) were based on Ginther et al. (2010). To obtain these measures, the boundaries of speech units related to fluency, such as syllables, runs (a continuous chunk of speech between two silent pauses, where a silent pause refers to silence equal or longer than 0.25 seconds) and silent and filled pauses, were demarcated in Praat using simultaneous consultation of both the waveform and spectrogram. The beginning and ending points of the fluency

Table 3.2.. List of acoustic measures related to fluency

Speech Time Ratio	The ratio of speaking time (excluding pauses) over total response time.
Number of Syllables	The number of total syllables produced during the speaking time.
Mean Syllable Duration	Total speaking time divided by total number of syllables.
Speech Rate	Total number of syllables divided by total response time
Articulation Rate	Total number of syllables divided by the sum of speaking time and total filled pause time.
Mean Syllable per Run	Total number of syllable divided by total number of runs. Run is defined as a continuous chunk of speech between two silent pauses. A silent pause means silence longer than 0.25 seconds.
Mean Run Length	Total speaking time divided by the number of runs.
Number of Silent Pauses	Total number of silent pauses in a given speech.
Mean Silent Pause Time	Total silent time divided by the number of silent pauses.
Silent Pause Ratio	The ratio of total silent pause time over total response time.
Number of Filled Pauses	Total number of filled pauses in a given speech. Filled pauses included hesitation such as "hmm", "huh", and incomplete words.
Mean Filled Pause Time	Total filled pause time divided by the number of filled pauses.
Filled Pause Ratio	The ratio of total filled pause time over total response time.

units were decided based on the boundaries of the corresponding segments. The number and duration of all syllables, runs, and silent and fill pauses were extracted using a Praat script and served as the basis for the calculation of the other fluency measures.

Acoustic measures related to phonetic intelligibility.

Among the six acoustic measurements of phonetic features of pronunciation (Table 3.3), three were related to the production of stop consonants, two were related to vowels and one to overall pitch (*f0*) range, all of which have been previously found to correlate with subjective ratings of intelligibility (Bradlow et al., 1996; Liu et al., 2000; Maniwa et al., 2009). Segmentation of stops, fricatives, and vowels was

accomplished in Praat through simultaneous consultation of waveform and wideband spectrogram. Criteria for determining segmental boundaries are provided below.

The initial acoustic analysis also included 12 measures on the three fricatives /f, s, ʃ/ in terms of their four spectral moments: spectral mean, standard deviation, skewness, and kurtosis. While these measures have been shown in other studies to contribute to the recognition of fricatives in English (Jongman et al., 2000), they were nevertheless excluded from final analysis here primarily because in the present data set these measures exhibited standard deviations greater than group means, making them inappropriate for the statistical analyses to be employed here.

Table 3.3.. Acoustic measures related to phonetic intelligibility

Stop	Voiced VOT	Average duration of voiced stop VOTs
	Voiceless VOT	Average duration of voiceless stop VOTs
	VOT Diff	Difference between the average duration of voiced and voiceless VOTs
Vowel	Mean Vowel Duration	Average duration of vowels
	Vowel Space	Area of vowel space determined by ERB-transformed F1 and F2 values
Suprasegmental	*F0* Range	Difference between *f0* maximum and minimum over syllables

VOT was measured from the release of the stop burst to the initiation of glottal vibration of the following vowel (Abramson & Lisker, 1964). The three measures of stop consonants were selected because VOT was identified by previous studies to be the primary acoustic cue for distinguishing voicing contrast in initial stops cross-linguistically (Abramson & Lisker, 1964), and finer differences in VOTs could also cue places of articulation (Chao & Chen, 2008).

To obtain the vowel measure, all vowels in each speech sample were segmented such that vowel onset was identified at the point of stop release if there was one, otherwise at the beginning of voicing. Vowel offset was marked at the offset of voicing. Low-frequency voicing preceding a following consonant was not counted as part of the preceding vowel. The first two formants (F1 and F2) of each vowel were measured from formant tracks at the center of the vowel steady state, if any (otherwise peak intensity), and were converted to the perceptually motivated equivalent rectangular bandwidth (ERB) scale using the formula ERB=24.7*(0.0043*f+1), where f is the original frequency measured in Hertz (Glasberg & Moore, 1990). The ERB values of F1 and F2 were averaged for each type of vowel for each speaker. The vowel space of each speaker was plotted on a two-dimensional space where F1 represented the x-axis and F2 the y-axis. The Euclidian area of each vowel space was calculated using a custom-written javascript "Convex Hull Calculator", implementing Andrew's Monotone Chain Convex Hull Algorithm (Wikibooks.org, 2014).

F0 range was computed such that for each syllable produced by each speaker, the minimum and maximum *f0* were extracted from the voiced portion using the automatic pitch tracker in Praat. *F0* range was taken as the differences between maximum and minimum *f0*, and the arithmetic mean of all the *f0* differences by each speaker was used to represent the mean *f0* range of that particular speaker. Extreme values were checked by hand and re-computed based on the inverse of the period at the identified locations.

As a measure of reliability, the acoustic measures of fluency and intelligibility were re-measured and re-computed by two other experimenters on two randomly selected speech samples out of the twenty (10% of the speech data). The values of all the re-measured variables were strongly correlated across all experimenters, with the highest and lowest correlation coefficients appearing for F1 values (r=0.91, p<0.001) and pause duration (r=0.99, p<0.001), respectively. Due to overall high correlation, no re-measurements were made.

3.3 Results

3.3.1 Word intelligibility

To test for a learning effect in the word intelligibility task, a paired t-test was conducted using the intelligibility scores from both practice sessions (before and after the experimental session). Result showed a slight increase in the post-test score (71%) compared to the pre-test score (70%), but this difference was not statistically significant ($t(9)=1.10$, $p=0.30$). This suggests that the word intelligibility task did not introduce any learning effect.

A repeated measures mixed ANOVA was administered to determine whether the words produced by high-proficiency speakers were more intelligible than the words produced by intermediate-proficiency speakers. The dependent variable was word intelligibility score, and the factors included two fixed effects of proficiency (two levels: Intermediate, High) and L1 (two levels: Chinese, Korean), and a random effect of listener. Results (Table 3.4) showed a non-significant effect of listener ($F(9,9)=1.60$, $p=0.25$), but significant effects of proficiency ($F(1,9)=311.93$, $p<0.0001$) and L1 ($F(1,9)=62.94$, $p<0.0001$). Specifically, high-proficiency speakers received higher word intelligibility scores (81%) than intermediate-proficiency speakers (66%), and Korean speakers received higher scores (77%) than Chinese speakers (70%). The interaction of proficiency and L1 was marginally significant ($F(1,9)=5.14$, $p=0.05$).Post hoc Tukey HSD tests revealed that Korean speakers of high proficiency received the highest scores of word intelligibility (84%), followed by Chinese speakers of high proficiency (78%), Korean speakers of intermediate proficiency (70%), and lastly, Chinese speakers of intermediate proficiency (62%). All the pairwise comparisons were statistically significant ($p<0.05$).

While the proficiency effect was expected, it was unclear what caused the L1 effect, and listeners background information surveys suggested nothing consistent. It is possible that the L1 effect is attributable to the relatively small number of speakers in this study. Alternatively, Korean speakers of English may simply be

Table 3.4.: ANOVA table for word intelligibility.

	df	*F*	η	*p*
Proficiency	1	311.93	0.68	<0.0001
L1	1	62.94	0.38	<0.0001
Proficiency*L1	1	5.14	0.08	0.05

more intelligible to native English listeners than are Chinese speakers due to as-yet unstudied L1 effect, perhaps similar to the benefit enjoyed by speakers from a Slavic language background (Derwing & Munro, 2013). This seems less plausible in the present case, because Korean and Chinese are linguistically more similar to one another than are Sinitic and Slavic languages, but it is possible that the tonal properties of Chinese, which are not present in Korean, may play some role. Further research is necessary in this area.

3.3.2 Listening effort, subjective intelligibility, and acceptability

Three repeated measures mixed ANOVAs were performed to determine whether the ratings of listening effort, subjective intelligibility, and acceptability differed between high- and intermediate-proficiency speakers. Listening effort, subjective intelligibility, and acceptability were used as the dependent variable for the three ANOVA tests respectively, and each test included two fixed factors (proficiency, L1), and a random factor of listener. Prior to statistical analyses, subjective intelligibility and acceptability scores were adjusted by subtracting the raw scores from 21 such that higher scores corresponded to higher subjective intelligibility and acceptability.

The three ANOVAs showed a significant main effect of proficiency on listening effort ($F(1,19)=109.91$, $p<0.0001$) (Table 3.5), subjective intelligibility ($F(1,19)=169.93$, $p<0.0001$) (Table 3.6), and acceptability ($F(1,19)=228.66$, $p<0.0001$) (Table 3.7), but none of the three tests yielded a significant L1 effect and there were no significant

Table 3.5.: ANOVA table for listening effort ratings.

	df	*F*	η	*p*
Proficiency	1	109.91	0.70	<0.0001
L1	1	0.20	0.03	0.66
Proficiency*L1	1	1.15	1.05	0.32

Table 3.6.: ANOVA table for subjective intelligibility ratings.

	df	*F*	η	*p*
Proficiency	1	169.93	0.68	<0.0001
L1	1	0.15	0.03	0.70
Proficiency*L1	1	0.17	0.06	0.69

Table 3.7.: ANOVA table for acceptability ratings.

	df	*F*	η	*p*
Proficiency	1	228.66	0.70	<0.0001
L1	1	0.02	0.01	0.88
Proficiency*L1	1	0.27	0.04	0.61

interactions. This suggests that the listeners found the high-proficiency L2 speakers were more intelligible and acceptable than intermediate-proficiency speakers, and listening to the former group was less effortful.

3.3.3 Acoustic measures and listener assessment of fluency and intelligibility

The design of this study contained a mismatch between acoustic and listener variables: Each speaker received one value for each acoustic measure, but ten scores

of word intelligibility, and twenty each of listening effort, subjective intelligibility and acceptability. To tackle this problem, the four listener variables were averaged such that each speaker received a mean score of word intelligibility, subjective intelligibility, listening effort and acceptability. These mean scores, together with the acoustic variables, were used for the following statistical analyses. Descriptive statistics of the acoustic measures and listener variables are shown in Table 3.8.

Table 3.8.. Means and standard deviations of all variables.

Variables (unit)	High Proficiency Group: Mean (SD)	Low Proficiency Group: Mean (SD)
Acoustic measures		
Speech Time Ratio	0.68 (0.07)	0.57 (0.09)
Number of Syllables	284.90 (44.81)	244.70 (54.16)
Mean Syllable Duration (s)	0.23 (0.03)	0.25 (0.03)
Speech Rate (syllable/s)	2.79 (0.39)	2.17 (0.46)
Articulation Rate (syllable/s)	3.61 (0.47)	3.22 (0.45)
Mean Syllable per Run (syllable/s)	6.83 (2.12)	5.02 (0.96)
Mean Run Length (s)	1.70 (0.66)	1.33 (0.19)
Number of Silent Pauses	39.40 (10.07)	52.10 (7.40)
Mean Silent Pause Time (s)	0.60 (0.04)	0.72 (0.24)
Silent Pause Ratio	0.23 (0.05)	0.33 (0.10)
Number of Filled Pauses	22.40 (11.75)	24.90 (10.29)
Mean Filled Pause Time (s)	0.43 (0.08)	0.45 (0.08)
Filled Pause Ratio	0.09 (0.05)	0.10 (0.04)
*F*0 Range (Hz)	316.20 (69.19)	323.22 (69.83)
Vowel Space	3882.38 (1114.16)	3441.41 (1079.48)
Mean Vowel Duration (s)	0.11 (0.02)	0.11 (0.03)
Voiced VOT (ms)	21 (6)	24 (9)
Voiceless VOT (ms)	66 (13)	68 (18)
VOT Diff (ms)	45 (14)	44 (17)
Listener assessments		
(A) Word Intelligibility	0.81 (0.05)	0.66 (0.11)
(B) Listening Effort	7.87 (0.91)	12.65 (1.19)
(C) Subjective Intelligibility	12.17 (1.04)	6.98 (1.40)
(D) Acceptability	11.58 (1.22)	6.09 (0.96)

Note: shaded rows indicate significant differences between the two proficiency groups.

To determine whether high- and intermediate-proficiency L2 speakers differed in terms of any acoustic measures, a series of two-way ANOVA tests with Bonferroni adjustment were carried out with the acoustic measures as dependent variables, and with proficiency and L1 as the two fixed factors. Results showed a significant main effect of proficiency on several fluency measures but not on any acoustic

measures related to phonetic features of pronunciation. Compared to intermediate-proficiency speakers, high-proficiency speakers showed higher scores in speech time ratio ($F(1.16)$=15.50, p=0.0012), speech rate ($F(1.16)$=19.33, p=0.0004), and mean syllables per run ($F(1.16)$=7.03, p=0.0174), and lower scores in silent pause number ($F(1.16)$=13.62, p=0.0020), mean silent pause time ($F(1.16)$=5.42, p=0.0333), and silent pause ratio ($F(1.16)$=13.87, p=0.0018). This suggests that these fluency measures might be useful for differentiating between speakers of high and intermediate proficiency.

Table 3.9.. Correlation coefficients between acoustic and listener variables, and among listener variables.

Variables	Word Intelligibility	Listening Effort	Subjective Intelligibility	Accept -ability
Acoustic measures related to fluency				
Speech Time Ratio	0.14	-0.52*	0.45*	0.50*
Number of Syllables	0.13	-0.48*	0.39*	0.46*
Mean Syllable Duration	-0.25	0.39*	-0.33	-0.40*
Speech Rate	0.21	-0.59*	0.50*	0.58*
Articulation Rate	0.18	-0.39*	0.31	0.37*
Mean Syllable per Run	0.31	-0.45*	0.42	0.45*
Mean Run Length	0.24	-0.31	0.31	0.31
Number of Silent Pauses	-0.42	0.51*	-0.50*	-0.53*
Mean Silent Pause Time	0.10	0.34	-0.27	-0.32
Silent Pause Ratio	-0.15	0.55*	-0.49*	-0.54*
Number of Filled Pauses	0.03	0.00	0.05	0.04
Mean Filled Pause Time	-0.16	0.24	-0.20	-0.24
Filled Pause Ratio	-0.02	0.05	0.01	-0.00
Acoustic measures related to phonetic intelligibility				
*F*0 Range	0.26	-0.06	0.07	0.04
Vowel Space	0.18	-0.21	0.20	0.21
Mean Vowel Duration	0.21	0.05	0.02	-0.05
Voiced VOT	-0.28	0.37*	-0.36*	-0.32
Voiceless VOT	-0.03	0.07	-0.03	-0.09
VOT Diff	0.12	-0.11	0.15	0.07
Listener assessments				
Word Intelligibility	1	0.77*	0.81*	0.77*
Listening Effort	-	1	-0.99*	-0.99*
Subjective Intelligibility	-	-	1	0.99*
Acceptability	-	-	-	1

Note: * indicates significant correlation (p<0.05).

To further explore the relationship between acoustic measures and listener ratings, Pearson product-moment correlation coefficients were computed, as shown in Table

3.9. The four listener variables were significantly correlated with one another (r>0.7, p<0.0001), although the strength of correlations were much stronger among the three subjective measures of listening effort, intelligibility and acceptability than they were between any of these and word intelligibility. This may be an artifact of the experimental design, where word intelligibility was computed from a recognition task, while the other three scores were derived from 20-pt rating scales. The high correlation among the three subjective measures indicates the possibility that listeners were not able to tease apart the three concepts. However, it is also possible that these measures may tap slightly different aspects of subjective speech evaluation, which were nuanced yet informative.

There were significant correlations between the subjective measures and some fluency measures (speech time ratio, number of syllables, mean syllable duration, speech rate, articulation rate, mean syllable per run, number of silent pauses, and silent pause ratio). As suggested by the ANOVA results, subjective measures were not correlated with any of the acoustic variables related to phonetic features of pronunciation, although listening effort and subjective intelligibility were moderately correlated with voiced stop VOTs. Finally, word intelligibility was not significantly correlated with any of the acoustic variables, suggesting that word recognition might be partially independent from phonetic intelligibility (at least, the properties measured here).

3.3.4 Factor and regression analyses

To determine the functional relationships among acoustic variables of fluency and phonetic features of pronunciation, all acoustic measures were submitted to a factor analysis in SAS 9.3. This analysis was based on a 19x19 matrix containing the pairwise correlation coefficients of all acoustic variables, and the matrix was decomposed using the principal component method. This yielded four mutually uncorrelated factors of eigenvalues greater than 1.0, together accounting for 91% of the total variance of the matrix. These factors were rotated via the varimax procedure. Based on loadings (see

Table 3.10), these factors may be loosely identified with a join dimension of fluency and pronunciation (Factor 1), overall fluency (Factor 2), filled pausing (Factor 3) and vowel space (Factor 4).

Table 3.10.. Factor loadings represented in the rotated factor matrix.

	Factor1	Factor2	Factor3	Factor4
Speech Time Ratio	-0.26	0.80	0.02	0.49
Number of Syllables	-0.73	0.41	0.18	0.30
Mean Syllable Duration	0.88	-0.27	-0.10	0.21
Speech Rate	-0.69	0.69	0.07	0.17
Articulation Rate	-0.88	0.35	-0.25	-0.07
Mean Syllable per Run	-0.17	0.87	-0.31	0.18
Mean Run Length	0.17	0.81	-0.36	0.30
Number of Silent Pauses	-0.06	-0.89	0.19	0.18
Mean Silent Pause Time	0.25	-0.37	-0.61	-0.52
Silent Pause Ratio	0.15	-0.78	-0.43	-0.40
Number of Filled Pauses	0.01	-0.33	0.84	-0.11
Mean Filled Pause Time	0.87	0.06	0.05	-0.19
Filled Pause Ratio	0.27	-0.20	0.86	-0.27
*F*0 Range	0.26	0.24	-0.06	0.44
Vowel Space	0.07	0.08	-0.04	0.64
Mean Vowel Duration	0.78	-0.09	-0.12	-0.19
Voiced VOT	0.25	-0.06	0.16	-0.67
Voiceless VOT	0.89	0.22	0.21	0.07
VOT Diff	0.78	0.25	0.12	0.41

Note: * indicates factor loadings whose absolute value are greater than or equal to .5

Factor scores were used as predictors in four regression analyses using the stepwise selection method in SAS, where word intelligibility, listening effort, subjective intelligibility, and acceptability were the respective response variable. Results showed that only Factor 2 significantly predicted the subjective ratings: Listening effort ($F(1, 18)=6.36$, $p=0.02$), subjective intelligibility ($F(1, 18)=5.23$, $p=0.03$), and acceptability ($F(1, 18)=6.53$, $p=0.02$). Word intelligibility was not predicted by any factors.

Factor 2 is predominantly fluency-related, with high loadings on variables that are primarily related to speech quantity (speech time ratio, mean run length), speed (speech rate, mean syllable per run) and silence (number of silent pauses, silent pause ratio). Neither of the factors with high loading of pronunciation-related variables (Factors 1 and 4) were predictive of any listener variables. This suggests that, at

least for the speech samples used in the present study, fluency makes the greatest contribution to native English listeners' subjective assessment of L2 English speech.

3.4 Discussion

This study examined a set of acoustic measures related to the fluency and pronunciation of L2 English speakers of high and intermediate proficiency, and how these constructs may be linked to listeners' subjective evaluation of listening effort and L2 speech intelligibility and acceptability.

Examination of listeners' evaluation of L2 English speech in this study showed that listeners' were better at recognizing words produced by high-proficiency speakers, whose speech was also found to be more intelligible and acceptable, and less effortful to listen to, in comparison with that of intermediate-proficiency speakers. Listening effort was highly correlated with subjective intelligibility and acceptability, and to a lesser extent, with word intelligibility, potentially suggesting that the amount of cognitive effort listeners invest in listening to L2 speech is directly linked to the degree to which they understand and are willing to accept the speech. Alternatively, it may also reflect an intrinsic artifact of the experimental design: Because participants made these ratings in quick succession after listening to each speech sample, they may have failed to completely differentiate between the three dimensions. Still, the three subjective measures, taken together, can be seen as reflecting listeners' evaluation of the overall quality of L2 speech.

No difference was found between intermediate- and high-proficiency L2 speakers in terms of acoustic measures related to phonetic pronunciation, suggesting that these speakers might exhibit relatively uniform pronunciation accuracy, at least in terms of the acoustic correlates measured here. This may be partially attributed to the fact that, although every speaker had an easily discernible L2 accent, they were all relatively advanced learners of English who had satisfied the admissions requirements for graduate study in a major US university, including passing its TOEFL require-

ments. Such a profile suggests that these L2 speakers may have attained a relatively advanced level of English pronunciation. In particular, it is possible that their acquisition of English sounds may have arrived at a plateau where differences in segmental production between the two groups were not sufficient to differentiate between them or predict word recognition and subjective evaluations.

However, it remains a question why the acoustic measures related to phonetic pronunciation did not effectively predict word intelligibility scores nor subjective evaluations assigned by listeners. Thus we cannot rule out the possibility that other prosodic factors, which are also known to contribute to accentedness or affect L2 intelligibility, such as L2 syllable production (Major, 2001), stress timing (Trofimovich & Baker, 2006), and placement of lexical stress (Field, 2005), may be better at differentiating between the two groups of speakers in this study. While the two groups did not differed significantly in terms of $f0$ range, which was used as an indicator of intonational variability, this single suprasegmental measure may not provide adequate information to reflect overall prosodic patterns. Given that the distinction between fluency and prosody is not always well defined, the present study was designed to focus mainly on acoustic properties that could be ascribed unambiguously as corresponding either to phonetic intelligibility or to fluency, but future study should explore in more depth the potential impact of prosodic features on listeners' evaluation of L2 speech.

In this study, the two groups did differ in fluency: The high-proficiency speakers produced more speech, spoke faster, and paused less than intermediate-proficiency speakers. Given that both groups had clearly identifiable L2 pronunciations, these results suggest that improving L2 proficiency may be accomplished more readily by increasing fluency rather than by developing native-like pronunciation. Although acquiring more native-like pronunciation may be desirable, results of this study suggest that it may be fluency, not pronunciation, that most differentiates intermediate- from high-proficiency L2 speakers, at least after a certain stage of acquisition.

More proficient speakers also received higher subjective ratings of intelligibility and acceptability, and lower ratings of listening effort, and factor analysis showed

that these subjective evaluations were best predicted by a fluency-related factor. This suggests that fluency-related features have a stronger effect on listeners' subjective evaluations than do fine-grained phonetic properties, possibly because native listeners are better able to cope with divergent pronunciations in L2 speech if they appear in an otherwise more fluent context.

At this point, it is not clear why, or through what mechanism, fluency might affect subjective intelligibility and acceptability of L2 speech. One possibility is that greater fluency may contribute to L2 intelligibility by making listening less effortful. When L2 speech flows smoothly with few pauses, listeners do not have to exert much effort to keep track of the speaker's utterances, and therefore they may even be able to devote more cognitive resources to decoding those sounds that the speaker may be producing in a less native-like manner. Less fluent speech is more difficult to understand because listeners must hold in working memory everything that was said earlier in the sentence for a longer period of time since such speech is slower, has more and longer pauses, and perhaps often has pauses in unexpected locations within clauses (Riggenbach, 1991). These characteristics may introduce extra demand on processing resources because attention is needed not only for the primary task of understanding, but also for the additional (sometimes unexpected) task of holding in memory chunks of syntactically incomplete utterances. This may ultimately increase effortfulness of listening and thereby negatively affecting native listeners' ability to understand L2 speech. Moreover, disfluent L2 speech may cause native listeners to switch from a more automatic speech processing mode that is mostly used for listening to native speech to a more controlled top-down processing mode, which further increasing processing load and reducing ease of listening. If fluency affects intelligibility and acceptability by modulating listening effort, it is plausible that improving fluency, even while not changing pronunciation, may in and of itself reduce listening effort and improve the intelligibility and acceptability of L2 speech. We are currently implementing another experiment to test this hypothesis.

3.5 Conclusion

Results of the present study suggests that the importance of speech fluency may outweigh pronunciation accuracy in affecting the intelligibility and acceptability of L2 speech, at least for this population consisting of graduate-level students in an English-speaking context. These findings have direct implications for L2 instruction and assessment, suggesting that when working with advanced learners at university level, it may be more efficient to first improve fluency, rather than focusing scarce teaching resources on trying to further improve the accuracy with which specific speech sounds are produced.

4. EXPERIMENT II: HOW FLUENCY AFFECTS LISTENING EFFORT AND THE INTELLIGIBILITY AND ACCEPTABILITY OF L2 ENGLISH

4.1 Introduction

Experiment II addresses the questions of whether differences in L2 fluency exert a major influence on perceived speech quality. Specifically, it tests the hypothesis that the reduction of non-juncture pauses, which has already been shown to improve perception of fluency (Bosker et al., 2014), will also improve perceived intelligibility and acceptability of L2 speech by reducing listening effort, even though non-native phonetic pronunciation remains constant. In addition, the study investigates whether the perceptual benefit from improved fluency is related to listeners' working memory capacity.

4.2 Methods

4.2.1 Participants

A total of 60 native speakers of American English (48 women, 12 men; mean age =19.2) were recruited for the experiment on a voluntary basis via posters placed on public bulletin boards on campus. All participants were undergraduate students in their first or second year of study from the Purdue University community, and they were paid by $10 for participation. None of the participants had a history of speech or hearing disorder by self report. The study was conducted according to a protocol approved by the Human Research Subjects Protection Program at Purdue University.

All participants had studied or were currently studying at least one foreign language. Sixteen participants had studied two foreign languages, one of which was Spanish for 14 of them, and another participant had studied three foreign languages,

also including Spanish. Spanish was the most commonly studied foreign language (47 out of the 60 participants), followed by French (14), German (5), American Sign Language (4), Japanese (2), Latin (2), Italian (1), and Hebrew (1). None of the participant had studied Chinese or Korean. The average age of the onset of foreign language training was 14.63 years, and the average years of foreign language study was 3.33. All participants reported having experience interacting with L2 English speakers from different L1 backgrounds, including East Asian and European languages. These interactions were reported as having taken place either in a classroom context where the L2 speaker was a teaching assistant, or in social and everyday settings.

4.2.2 Stimuli

The stimuli of Experiment II consisted of thirty L2 English speech samples that were grouped into three sets: The first set consisted of naturally produced L2 speech of high proficiency, the second set consisted of naturally produced L2 speech of intermediate proficiency, and the third set consisted of manipulated speech derived from the intermediate-proficiency speech samples. Details of the manipulation are elaborated on below.

The first two sets of speech samples were the twenty speech samples used in Experiment I (see Chapter 3, Section 3.2.1). The third set of speech samples were derived from the ten naturally produced L2 speech samples with intermediate proficiency. The first step of the manipulation was too demarcate all silent pauses (pauses with duration longer than 0.25 seconds) and filled pauses ("um" and "uh") that occurred within clausal boundaries using Praat 5.3.51 (Boersma & Weenink, 2013), and they were annotated as "non-juncture pause". The boundaries of silent pauses at clausal junctures were also marked and were annotated as "juncture pause". The beginning and ending points of all pauses were determined based on the boundaries of the corresponding segments. Upon completion of pause annotation, all the non-juncture pauses were removed, while all juncture pauses remained intact. The resulting ten

speech samples lasted between 55 to 87 seconds. The three groups of stimuli will be hereafter referred to as: High-proficiency speech, intermediate-proficiency speech, and manipulated speech.

4.2.3 Procedure

The experiment was implemented in three conditions. In the first condition, 20 listeners participated and only listened to the high-proficiency speech samples. Likewise, in the second condition, another 20 listeners evaluated the low-proficiency speech samples, and in the third condition, yet another 20 listeners assessed the manipulated speech samples. The procedure in each condition was exactly the same, except that the speech samples were different. This design, instead of having each participant evaluate all the speech samples, was chosen based on two considerations: 1) Having participants rate only 10 samples avoided participant fatigue that might have arisen due to an otherwise excessively long listening session; 2) participants would not hear both the intermediate-proficiency speech and the manipulated speech, since the content of the two sets was the same, and hence listening to both might introduce a familiarity effect that could have affected listeners' evaluation.

Each participant in this experiment completed two tasks, both implemented via the psychology software tool E-Prime 2.0 (Schneider et al., 2002). In the first task, participants listened to the ten speech samples in a random order, and evaluated them in terms of listening effort, acceptability, and subjective intelligibility on the same six 20-point Likert scales implemented in Experiment I (See Table 3.1). Listeners could only hear each speech sample once, and made all evaluations for each sentence immediate after hearing it.

After each participant completed the first task, they were prompted to the interface of the second task, which consisted of the n-back paradigm for measuring working memory capacity. As one of the most popular experimental paradigms in studies of working memory, the n-back task asks participants to make decisions about

whether a stimulus in a sequence matches the stimulus that appeared n trials previously (Gevins & Cutillo, 1993; Owen, McMillan, Laird, & Bullmore, 2005), where n is a predetermined integer varying from 1 to 3. Additionally, the paradigm often includes a 0-back task, which typically asks participants to decide whether the stimulus in a sequence is the same as a predefined target. Since the 0-back task does not require much executive working memory, it usually serves as a control condition for comparison with other tasks. Overall, the n-back task requires temporary holding, on-line monitoring, and additional processing and manipulation of information and is therefore assumed to capture the executive skills that are essential to working memory.

The current experiment adopted the 0-, 1-, and 2-back tasks of the n-back paradigm. All participants first performed the 0-back task, and then proceeded to the 1- and 2-back tasks. Each task consisted of one or more practice sessions (based upon participants' choice) and two formal test sessions. Each practice session contained 20 trials, and each formal test session contained 30 trials. In the 0-back task, letters were randomly projected on the screen, and participants were instructed to press the numeric keypad key "1" for "YES" (the letter was an X) and the numeric keypad key "2" for "NO" (the letter was not an X). In the other two tasks, letters were also randomly projected, and participants were instructed to press the 1 numeric key for "YES" (the present letter matched the previous letter for 1-back task, or the letter before the previous one for 2-back task), and to press the 2 numeric key for "NO". In each formal test session, there were 10 YES trials and 20 NO trials. Across the tasks, the letters shown on a white screen used 30 pt text in Palatino Linotype font, in black. Each letter appeared on the screen for 500 milliseconds, and the inter-stimulus interval was 3 seconds. In all tasks, participants were told to press the keys as rapidly and as accurately as possible.

The entire experiment lasted for no more than fifty minutes. Each experiment session was conducted individually. Before beginning the experiment, each partici-

pant filled in a background assessment form, compiling variables related to language attitude, language experience, and other potentially relevant personal data.

4.2.4 WMC index computation

In order to compute the WMC index, four types of responses were obtained from 2-back tasks: hit, miss, false alarm, and correct rejection (Table 4.1). The 2-back task was selected because it was the most challenging and may thus better reflect the WMC differences of among the participants than the other tasks.

Table 4.1.. Response scheme of the n-back tasks.

	Response: YES	Response: NO
Stimuli: YES	HIT	MISS
Stimuli: NO	FALSE ALARM	CORRECT REJECTION

The sensitivity index d' was obtained to index participants' WMC, using the formula: $d'=Z^{HIT} - Z^{FA}$ (Macmillan & Creelman, 1990), where HIT denotes the proportion of YES trials that the participant responded "YES" (i.e., hits/(hits+misses)), and FA the proportion of NO trials that the participant responded "NO" (i.e., false alarms/(false alarms+correct rejections)). The two rates of HIT and FA were then z-transformed using the formula NORMSINV(HIT) - NORMSINV(FA) in a Microsoft Excel spreadsheet. Perfect scores were adjusted using 1-1/(2n) for HIT (1), and 1/(2n) for FA (0), where n is the total number of hit and false alarm trials. Results showed that the mean sensitivity index (d') was 3.05, with a median of 2.95, standard deviation of 0.93, and a range of 3.86 (from 0.40 to 4.26).

4.3 Results

4.3.1 Pause analysis

Table 4.2 shows the pause duration and ratio for individual speakers. Note that the speakers of the low-proficiency and manipulated speech were the same, who were different from the speakers of the high-proficiency speech. Overall, the average total pause duration of the high-proficiency speech samples (M=30s, SD=9.66s) was shorter than that of the intermediate-proficiency speech samples (M=47s, SD=12s), and also in terms of the average ratio of total pause duration over total sample duration (high-proficiency speech: M = 29%, SD =8%; intermediate-proficiency speech: M=43%, SD=10%). Comparison between these two sets of speech samples generally indicates that, other factors aside, the proportion of pause in speech may be an important aspect of speech performance that differentiates L2 speakers of relatively high and intermediate proficiency.

The average total duration of the manipulated speech was about 10s (SD=3.36) and the average pause ratio was 15% (SD=5%), which were considerably smaller than the intermediate-proficiency speech, suggesting that the manipulation did achieve substantial pause reduction. Since all the remaining pauses in the manipulated speech constituted silent pauses at syntactic boundaries, it led to the question of whether they were comparable with the juncture pauses in the high-proficiency speech. Table 4.2 showed that the average total duration of juncture pauses and the average ratio in the high-proficiency speech was 10.49 s (SD=2.63) and 10% (SD=2%). These descriptive statistics suggest that the intermediate- and high-proficiency speech samples were sufficiently similar in terms of the duration and ratio of juncture pauses, so the main reason that intermediate-proficiency speakers produced overall longer pauses and greater ratio of pause over total utterance than high-proficiency speakers is perhaps because the former produced more non-juncture pauses.

Due to variations in the total duration of the speech samples, statistical tests of pause ratio were more reliable and informative than absolute pause duration. Thus the

Table 4.2.. Pause analysis.

Speaker (L1)	High-Proficiency Speech		High-Proficiency Speech: Juncture Pauses		Speaker (L1)	Intermediate-Proficiency Speech		Manipulated Speech*	
	Duration	Ratio	Duration	Ratio		Duration	Ratio	Duration	Ratio
1 (Chinese)	12.30	0.15	5.91	0.07	11 (Chinese)	31.97	0.30	7.38	0.09
2 (Chinese)	22.86	0.23	9.87	0.10	12 (Chinese)	41.35	0.35	8.89	0.10
3 (Chinese)	35.61	0.34	10.52	0.10	13 (Chinese)	42.32	0.40	13.05	0.17
4 (Chinese)	28.75	0.33	11.00	0.13	14 (Chinese)	34.65	0.39	8.76	0.14
5 (Chinese)	23.44	0.24	8.48	0.09	15 (Chinese)	35.80	0.33	7.19	0.09
6 (Korean)	34.33	0.32	9.90	0.09	16 (Korean)	58.29	0.55	7.30	0.13
7 (Korean)	22.63	0.19	8.49	0.07	17 (Korean)	51.39	0.44	16.20	0.20
8 (Korean)	40.64	0.34	14.35	0.12	18 (Korean)	61.36	0.52	11.88	0.17
9 (Korean)	33.49	0.33	12.01	0.12	19 (Korean)	68.04	0.57	7.74	0.13
10 (Korean)	44.22	0.40	14.35	0.13	20 (Korean)	46.75	0.49	14.62	0.23
Average	**29.83**	**0.29**	**10.49**	**0.10**	**Average**	**47.19**	**0.43**	**10.30**	**0.15**

* Note: Manipulated speech only contains juncture pauses from the intermediate-proficiency speech.

ratios were submitted to a one-way analysis of variance (ANOVA) with speech/pause type as a fixed factor (four levels: intermediate-proficiency pause, manipulated pause, high-proficiency pause, and high-proficiency juncture pause). The test yielded a significant main effect for the factor ($F(3,36)=50.03$, $p<0.0001$, $\eta=0.89$). Post hoc Tukey HSD tests exhibited a significant lower pause ratio of the high-proficiency speech (29%) than the intermediate-proficiency speech (43%), and the pause ratio of the manipulated speech (15%) was significantly lower than both the intermediate- and high-proficiency speech. Meanwhile, the pause ratio of the manipulated speech was not statistically different from the juncture pause ratio of the high-proficiency speech (10%). Taken together, these results suggest that the stimuli manipulation implemented in this study had effectively improved the level of fluency from the intermediate-proficiency L2 speech. Therefore it is predicted that the manipulated

speech may receive better listener evaluation compared to the intermediate-proficiency speech, and that the manipulated speech may be perceived as equivalent to the high-proficiency speech in terms of listening effort, intelligibility, and acceptability due to improved fluency.

4.3.2 Subjective speech evaluation

In order to determine whether the three sets of speech samples received different subjective evaluations, three one-way ANOVAs were administered with listening effort, intelligibility, and acceptability scores as the dependent variable respectively (Figure 4.1). Each ANOVA included a fixed factor of condition (intermediate, manipulated, high). Prior to statistical analyses, subjective intelligibility and acceptability scores were adjusted by subtracting the raw scores from 21 such that higher scores corresponded to higher subjective intelligibility and acceptability.

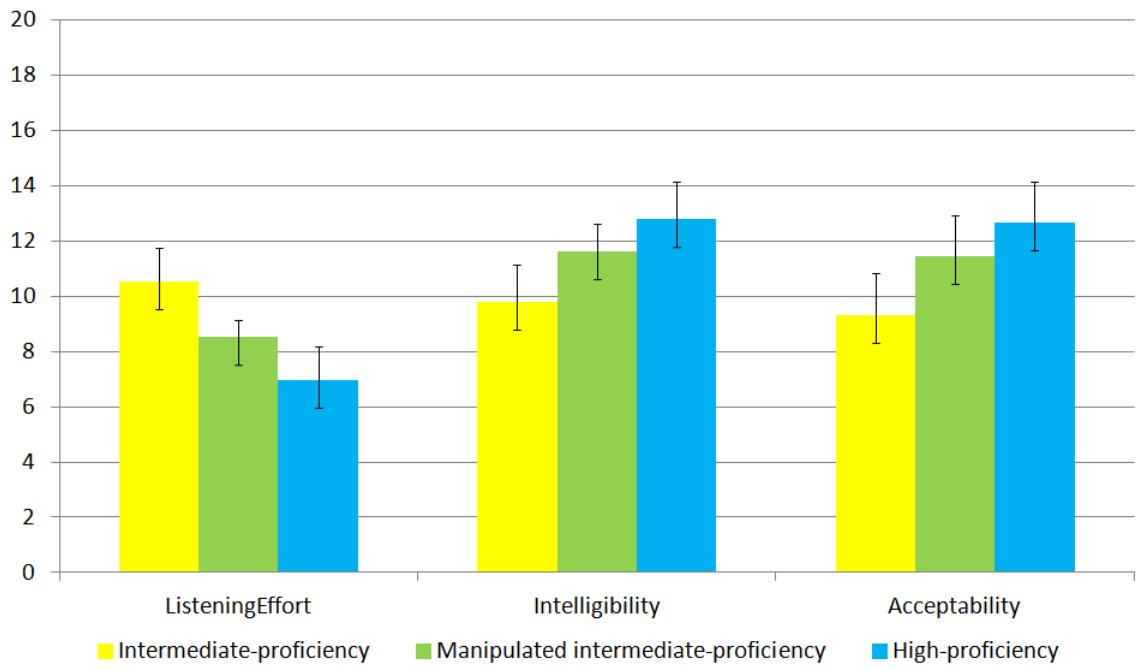

Figure 4.1. Subjective evaluation for the three types of speech (with error bars).

The main effect of condition was significant for listening effort ($F(2, 597)=33.31$, $p<0.0001$, $\eta=0.32$). Post hoc Tukey HSD tests showed that listening to the high-proficiency speech was significantly less effortful (M=6.96), than to manipulated speech (M=8.54), and both were significantly less effortful than listening to intermediate-proficiency speech (M=10.52).

The main effect of condition was significant for intelligibility ($F(2, 597)=19.91$, $p<0.0001$, $\eta=0.25$). Post hoc Tukey HSD tests showed that listeners found the high-proficiency speech (M=8.20) to be more intelligible than the manipulated speech (M=9.39), which was also more intelligible than the intermediate-proficiency speech (M=11.22).

The main effect of condition was significant for acceptability ($F(2, 597)=19.54$, $p<0.0001$, $\eta=0.25$). Post hoc Tukey HSD showed that, although the intermediate-proficiency speech was less acceptable (M=11.69) than both the manipulated speech (M=9.57) and the high-proficiency speech (M=8.36), the acceptability ratings of the latter two types did not differ significantly.

4.3.3 WMC as a covariate

The above ANOVA analyses, however, did not take into account the possibility that some of the variance in the subjective ratings could be explained by differences in listeners' WMC. To address this question, d' sensitivity scores were included as a covariate in three analyses of covariance (ANCOVA) with listening effort, intelligibility, and acceptability as the dependent variable respectively. Each ANCOVA also included a fixed factor of condition (Figure 4.2).

Results of the ANCOVA for listening effort (Table 4.3) showed a significant main effect of condition ($F(2, 596)=34.44$, $p<0.0001$), and a significant effect of WMC ($F(1, 596)=6.10$, $p=0.01$). Post hoc Tukey HSD tests showed the same results as the ANOVA: Listening effort ratings were significantly different between the three types of speech even when WMC was controlled for.

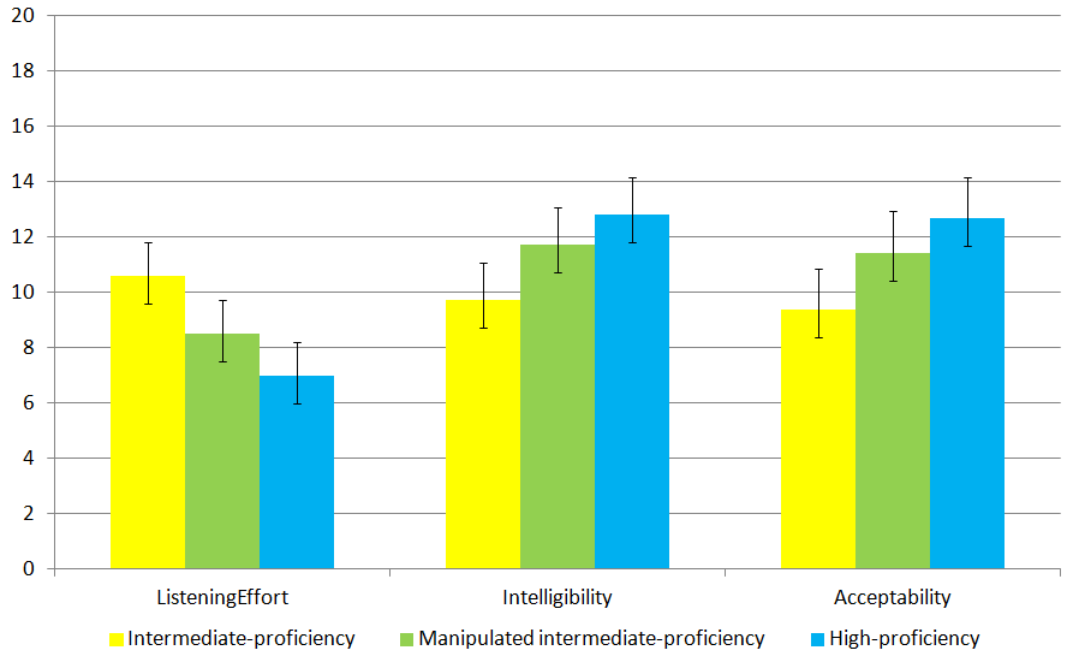

Figure 4.2. Subjective evaluation for the three types of speech with WMC as a covariate (with error bars).

Table 4.3.: ANCOVA table for listening effort ratings

	df	*F*	η	*p*
Condition	2	34.44	0.32	<0.0001
WMC	1	6.10	0.09	0.01

Results of the ANVOCA for intelligibility (Table 4.4) exhibited a significant main effect of condition ($F(2,\ 596)=20.76,\ p<0.0001$), and a significant effect of WMC ($F(1,\ 596)=5.41,\ p=0.02$). However, post hoc Tukey HSD tests yielded different results from the ANOVA: when WMC was controlled for, no significant difference was observed between the high-proficiency and manipulated speech, with both being more intelligible than intermediate-proficiency speech.

Table 4.4.: ANCOVA table for intelligibility ratings

	df	*F*	η	*p*
Condition	2	20.76	0.26	<0.0001
WMC	1	5.41	0.09	0.02

Results of the ANVOCA for acceptability (Table 4.5) showed a significant main effect of condition ($F(2, 596)=19.17$, $p<0.0001$), but WMC was not significant ($F(1, 596)=0.92$, $p=0.34$). Thus the results of the ANOVA are sufficient.

Table 4.5.: ANCOVA table for acceptability ratings

	df	*F*	η	*p*
Condition	2	19.17	0.25	<0.0001
WMC	1	0.92	0.04	0.34

Overall, the ANCOVA results demonstrated that after adjusting for WMC, the intermediate-proficiency speech was still more effortful to listen to than manipulated and high-proficiency speech, and was significantly less intelligible and acceptable. Moreover, while the manipulated speech was more effortful to listen to than the high-proficiency speech, the two were of similar degree of intelligibility and acceptability.

4.3.4 Comparison between listeners of high and low WMC

The significant effect of WMC in the ANCOVA raises the additional question of whether listeners' patterns of subjective ratings differed by individual processing capability, and if so, how. Our hypothesis is that listeners with high WMC may be better at processing and understanding intermediate-proficiency speech than listeners of low WMC, and therefore, low-WMC listeners may gain more benefit from pause elimination in the intermediate-proficiency speech. To test this hypothesis, each of the

60 listeners was assigned to one of two equally sized groups according to their WMC score. The 30 listeners with a WMC index above the group median score (2.95) were labeled as "high-WMC" and the other 30 listeners as "low-WMC". Three two-way ANOVA tests were performed with listening effort, intelligibility, and acceptability as the dependent variable respectively. Each ANOVA contained two fixed factors: condition (intermediate, manipulated, high) and WMC-level (high and low).

Table 4.6.: ANOVA Table Listening effort ratings by condition and WMC-level.

	df	*F*	η	*p*
Condition	2	39.25	0.32	<0.0001
WMC-level	1	27.93	0.03	<0.0001
Condition*WMC-level	2	5.64	0.10	0.004

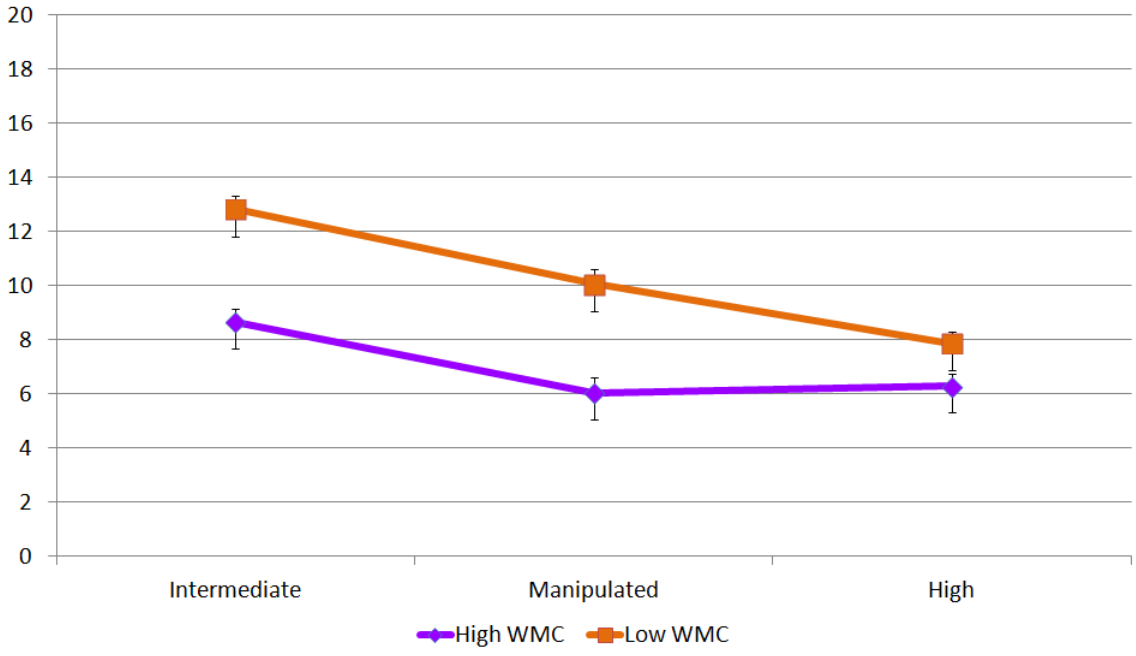

Figure 4.3. Listening effort ratings by condition and WMC-level (with error bars).

Results of the ANOVA for listening effort (Table 4.6, Figure 4.3) showed significant main effects for condition ($F(2, 594)=39.25$, $p<0.0001$) and WMC-level ($F(1, 594)=27.93$, $p<0.0001$), and the interaction term was also significant ($F(2, 594)=5.64$, $p=0.004$). According to Post hoc Tukey HSD tests, the low-WMC group expended significantly more listening effort than the high-WMC group for all three types of speech. Moreover, listening effort of both groups was reduced significantly when listening to the manipulated speech compared to the intermediate-proficiency speech. However, while the low-WMC group rated the manipulated speech as more effortful than the high-proficiency speech, the high-WMC group evaluated them as equally effortful, which may indicate a floor effect.

Table 4.7.: ANOVA Table Intelligibility ratings by condition and WMC-level.

	df	*F*	η	*p*
Condition	2	24.18	0.25	<0.0001
WMC-level	1	21.62	0.02	<0.0001
Condition*WMC-level	2	3.91	0.09	0.02

Results of the ANOVA for intelligibility (Table 4.7, Figure 4.4) showed significant main effects for condition ($F(2, 594)=24.18$, $p<0.0001$) and WMC-level ($F(1, 594)=21.62$, $p<0.0001$), and the interaction term was also significant ($F(2, 594)=3.91$, $p=0.02$). Post hoc Tukey HSD tests showed that the low-WMC group perceived all three types of speech to be significantly less intelligible than the high-WMC group. Moreover, both groups found the manipulated speech to be more intelligible than the intermediate-proficiency speech. However, while the low-WMC group found the high-proficiency speech to be more intelligible than the manipulated speech, the high-WMC group found them similarly intelligible, again indicating a floor effect.

Results of the ANOVA ratings (Table 4.8, Figure 4.5) for acceptability showed significant main effects for condition ($F(2, 594)=20.66$, $p<0.0001$) and WMC-level ($F(1, 594)=11.35$, $p=0.0008$), but the interaction term was not significant ($F(2,$

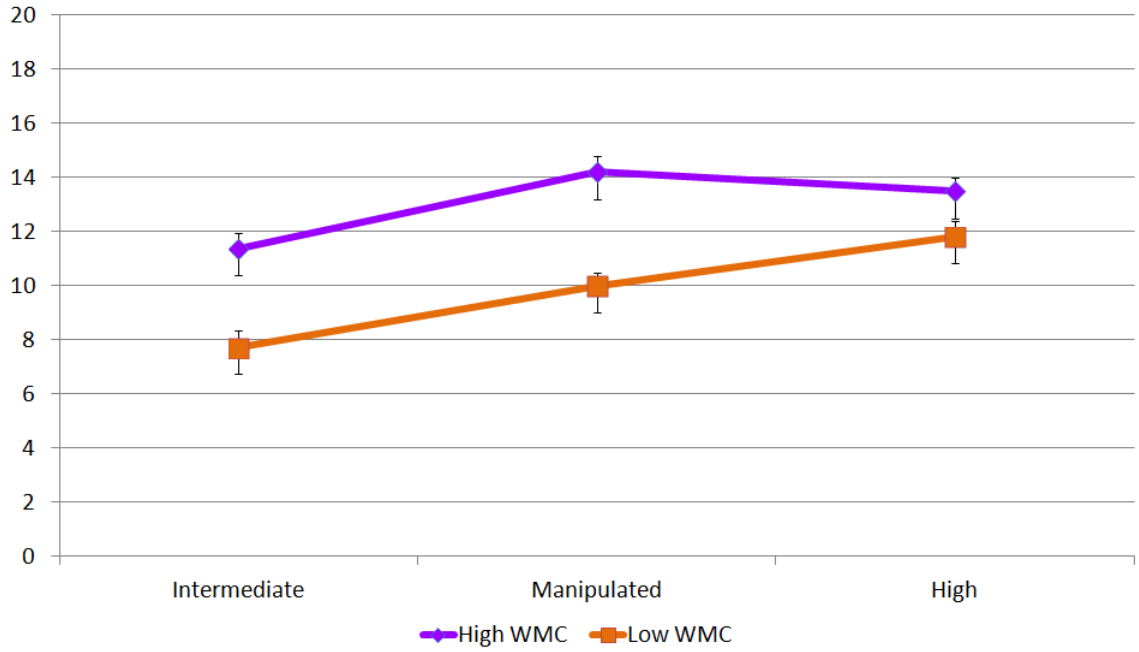

Figure 4.4. Intelligibility ratings by condition and WMC-level (with error bars).

Table 4.8.: ANOVA Table Acceptability ratings by condition and WMC-level.

	df	*F*	η	*p*
Condition	2	20.66	0.25	<0.0001
WMC-level	1	11.35	0.11	<0.0001
Condition*WMC-level	2	1.60	0.06	0.20

594)=1.60, p=0.20). Post hoc Tukey HSD tests exhibited that the intermediate-proficiency and manipulated speech was both seen as significantly less acceptable by the low-WMC group than the high-WMC group, but the two groups did not differ in their ratings of the high-proficiency speech. Moreover, both groups found the manipulated speech to be more acceptable than the intermediate-proficiency speech, but rated the manipulated speech to be similarly acceptable with high-proficiency speech.

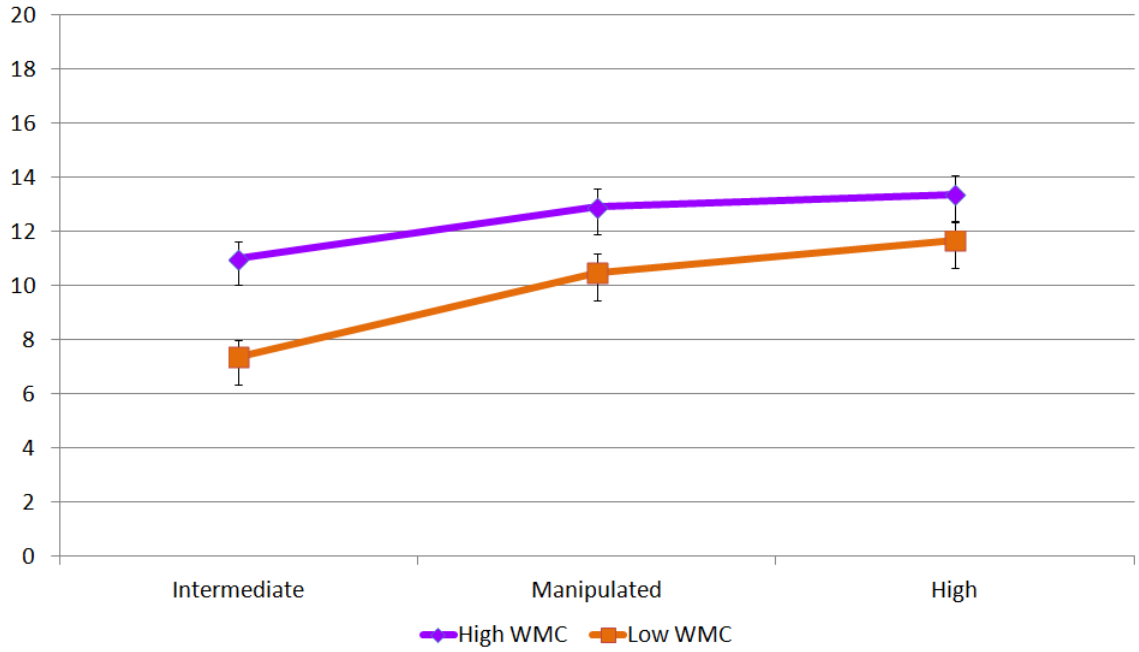

Figure 4.5. Acceptability ratings by condition and WMC-level (with error bars).

4.4 Discussion

4.4.1 Fluency and evaluations of high- and intermediate-proficiency L2 speech

Analysis of pause duration and pause ratio showed that high-proficiency L2 speakers devoted less time to pausing than intermediate-proficiency speakers. Specifically, the two groups of L2 speakers used similar amount of time on juncture pauses i.e., pauses located at syntactic boundaries but intermediate-proficiency speakers spent significantly more time pausing overall, suggesting that it is the non-juncture pauses that differentiates high-proficiency from intermediate-proficiency speakers in this data set. Generally speaking, high-proficiency speakers formulated longer runs of speech with fewer pauses, perhaps as a result of having learned to better plan L2 structures under time constraint.

These fluency differences between high- and intermediate-proficiency speech were expected to influence listeners' subjective evaluation of speech quality. Results of this study showed that intermediate-proficiency speech received substantially lower ratings across all three subjective measures compared to high-proficiency speech, suggesting that disfluency, especially non-juncture pausing, is perhaps one of the major negative influences on subjective evaluation of L2 speech. Nevertheless, there is insufficient evidence to claim that the amount and location of pause is the only factor affecting listeners' perception, since the high- and intermediate-proficiency L2 speech also differed in other aspects such as voice quality. To better address this question, non-juncture pauses were edited out, but the speech samples after manipulation did not differ from the original speech in any other way.

4.4.2 Effect of improved fluency on L2 speech evaluation

The manipulated speech obtained significantly higher ratings than the intermediate-proficiency speech in all three subjective measures, suggesting that eliminating non-juncture pauses did benefit listeners. Eliminating non-juncture pauses provides listeners with fewer, longer runs of speech for the same utterance, which might free up processing resources by reducing the number of syntactically incomplete fragments that must be stored in working memory. This suggests that ratings of L2 intelligibility and acceptability may benefit from simply improving the temporal flow of speech, especially by reducing non-juncture pauses, without changing any segmental or suprasegmental properties.

Despite this evidence for the benefit of eliminating non-juncture pauses, results showed that the manipulated speech was still rated as less intelligible and as requiring more listening effort than high-proficiency speech. One possible explanation for this finding is that the high-proficiency speakers may have adopted a variety of strategies to assist listeners' understanding beyond reducing non-juncture pausing: In addition to better controlling the pace and rhythm of their speech, they may

also have been more resourceful in other domains such as lexical and grammatical accuracy and syntactic complexity. On the other hand, the manipulated and high-proficiency speech did receive similar ratings of acceptability. This is an intriguing result, particularly because the original intermediate-proficiency speech was perceived to be much less acceptable than the high-proficiency speech. This suggests that the removal of non-juncture pauses may have improved listeners' attitude towards the intermediate-proficiency speakers and enhanced their overall preference for the manipulated speech, even though it was still somewhat more cognitively demanding than the high-proficiency speech.

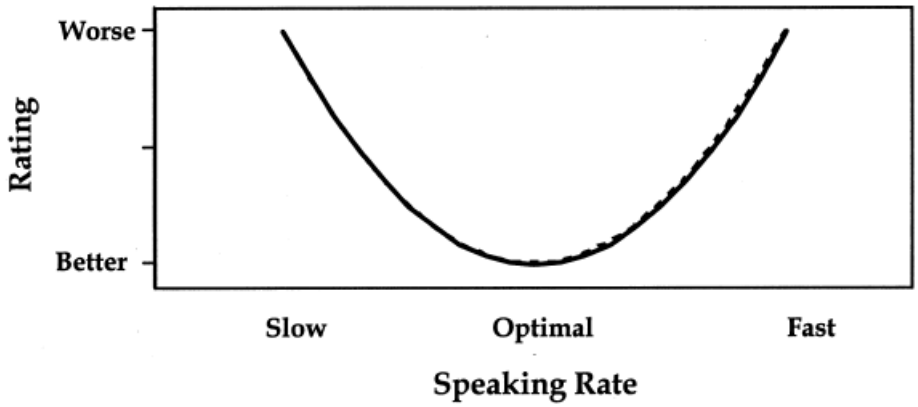

Figure 4.6. The hypothesized curvilinear effect of speaking rate on listeners' judgment (adapted from Munro and Derwing (2001))

.

Yet another possible explanation for the different perceptual effects between manipulated speech and high-proficiency speech may hinge on differences in pause ratios. Notice that the average pause ratio of the manipulated speech (15%) was even lower than that of the high-proficiency speech (29%), suggesting that perceptual evaluations and pause ratio of L2 speech does not necessarily exhibit a linear relationship, but possibly a curvilinear relationship instead. In other words, when pause ratio is too low or too high, both may be associated with decreased ratings, while the optimal pause ratio may situate somewhere in between. Such a curvilenear relationship

is previously identified by Munro and Derwing (2001) between speech rate and L2 speech comprehensibility (see Figure 4.6): Very fast speech and very slow speech both negatively affected listeners' processing of speech information. In a similar vein, it is possible that native listeners may prefer L2 speech that characterizes a certain level of pause ratio (such as represented by the high-proficiency speech samples in this study) because L2 speech is generally more difficult to process than L1 speech. When pause ratio is too low, it may place extra demand on the listeners who have to process considerably long runs, and therefore they may be more inclined to assign poor ratings. On the other hand, when pause ratio is too high, listeners may be also under high load because they have to allocate additional processing resources to memorize speech chunks between pauses. Nevertheless, this non-linear relationship between pause ratio and listener judgments is only speculative, and is certainly worthy further exploration by future studies.

4.4.3 WMC and L2 speech evaluation

The next question is whether ratings of L2 speech quality are subject to differences in listeners' WMC. Results of the ANCOVA (with WMC as a covariate) showed that WMC did have a significant effect on the ratings of listening effort and intelligibility, suggesting that some of the rating differences were not only attributable to L2 speech characteristics, but also to listeners' WMC. Most importantly, the manipulated and high-proficiency speech samples were found to be similarly intelligible in this analysis, as opposed to the ANOVA (without WMC as a covariate), in which the high-proficiency speech was found to be more intelligible. That is to say, had all the listeners been of the same WMC level, the intelligibility scores of the manipulated speech would have matched those of the high-proficiency speech. Since the manipulated speech received higher listening effort ratings than the high-proficiency speech, this indicates that the same level of intelligibility between the two speech types was achieved at different processing costs: listeners still had to allocate more cognitive

resources to understand the manipulated speech. Such a mismatch between cognitive effort and behavioral consequences is not uncommon, as there is abundant experimental evidence showing that listening to equally intelligible speakers may still incur different levels of processing cost (McLennan & Luce, 2005). For example, studies comparing L1 and L2 speech processing showed that highly intelligible L2 speech may still require more processing effort than L1 speech (Munro & Derwing, 1995b; Schmid & Yeni-Komshian, 1999), and it is processed more slowly (Munro & Derwing, 1995b; Floccia, Butler, Goslin, & Ellis, 2009). The current study shows that this mismatch also occurs in processing L2 speech at different levels of proficiency: The same behavioral performance (intelligibility) may mask different magnitudes of cognitive effort commitment. Future research is certainly warranted to further explore this question.

4.4.4 Individual differences in WMC

Results also showed that listeners with different WMC process L2 speech differently. High-WMC listeners reported significantly less effort for processing all speech types and perceived them to be more intelligible than did listeners with lower WMC. This is possibly because high-WMC listeners are better at temporarily maintaining and monitoring speech information, and may even be better at suppressing irrelevant information, either because they have more available processing capacity or because they are more efficient at designating available processing capacity to useful cues when dealing with cognitively challenging tasks such as listening to L2 speech.

High-WMC listeners also assigned higher acceptability scores to both the intermediate-proficiency and manipulated speech as compared to the low-WMC group, but the two groups gave similar scores to the high-proficiency speech. While this could simply be a ceiling effect, it could also reflect some other difference in opinion or attitudes exhibited by high- vs. low-WMC listeners. Further research is needed to explore the contribution of cognitive factors to individual differences in acceptance of L2 speech.

The two groups also differed in terms of the degree of benefit obtained from the different types of speech used here. Both groups rated the manipulated speech to be less effortful, more intelligible, and more acceptable than the intermediate-proficiency speech, suggesting that all listeners, regardless of WMC, benefited from a reduction in non-juncture pausing. However, the two groups differed in terms of their evaluation of the manipulated vs. high-proficiency speech. The low-WMC listeners found that the manipulated speech was still more effortful and less intelligible than the high-proficiency speech while the high-WMC listeners gave them similar ratings of listening effort, intelligibility, and acceptability. This suggests that WMC still constrains the ability to cope with whatever differences remain between high- and manipulated intermediate speech: Listeners with more WMC are able to accommodate those differences within their available capacity while those with less WMC are not. In short, these results suggest that evaluation of the subjective quality of L2 speech may depend at least in part on listeners' WMC. More work is certainly needed to thoroughly decipher the relationship between WMC and L2 speech processing.

4.5 Conclusion

To summarize, Experiment II examined listeners' evaluation of the intelligibility, acceptability, and listening effort of L2 English speech of relatively high and intermediate proficiency, as well as a set of manipulated samples in which all non-juncture pauses were removed from the intermediate-proficiency speech. It was found that at least given the experimental manipulation reported here, pause reduction largely decreased the effort listeners had to expend in order to understand the L2 speech, and also enhanced intelligibility and acceptability. While the intermediate-proficiency speech required more listening effort and was less intelligible and acceptable than the high-proficiency speech, the manipulation made the intermediate-proficiency speakers to be as intelligible and acceptable as the high-proficiency speakers, although at the cost of more processing effort. Additionally, Experiment II also found how listeners

evaluate L2 speech may be susceptible to individual differences in WMC, which is critical for online speech processing.

5. GENERAL CONCLUSIONS

5.1 Findings and implications

This study examined a set of acoustic, objective and subjective measures related to the fluency, intelligibility, and acceptability of L2 speech, and how these variables may be linked to listening effort and working memory capacity (WMC). Experiment I found that listening effort was highly correlated with subjective intelligibility and acceptability, and to a lesser extent, with word intelligibility. While listening effort, subjective intelligibility and acceptability were highly correlated with many fluency measures, acoustic measures related to phonetic intelligibility did not predict either word intelligibility or subjective ratings of speech quality. These results suggest that acoustic features related to fluency have a stronger effect on native listeners' subjective evaluations of acceptability, intelligibility, and listening effort than do those related to fine-grained phonetic properties, at least for the intermediate and advanced L2 learners.

These findings lead to the hypothesis that improving fluency, even while not changing pronunciation, may in and of itself reduce listening effort and improve the intelligibility and acceptability of L2 speech. To test this hypothesis, in Experiment II, I manipulated the intermediate-proficiency speech by artificially removing all the non-juncture silent and filled pauses. These manipulated speech samples, as well as the original two sets of high- and intermediate-proficiency speech samples, were evaluated by native listeners of American English in terms of listening effort, intelligibility, and acceptability. Results exhibited that pause reduction largely decreased the effort listeners had to expend in order to understand the L2 speech, which also increased the intelligibility and acceptability ratings. Moreover, while the intermediate-proficiency speech demanded more listening effort and was much less intelligible and acceptable

than the high-proficiency speech, the manipulation managed to transform the same intermediate-proficiency speech so that it became as intelligible and acceptable as the high-proficiency speakers, although still incurring more processing cost.

In addition to the finding that listeners' ratings of L2 speech quality can be predicted by speaker characteristics such as fluency, this study also demonstrates that listener judgment can be influenced by contribution of listener-related factors. Specifically, Experiment II investigated whether these subjective evaluations may also depend on listeners' WMC, which is critically important for online speech processing. It was found that listeners' evaluations of L2 speech did depend on individual WMC differences, in particular that listeners of relatively higher WMC seemed to be more effective at allocating their cognitive resources for processing L2 speech compared to listeners of lower WMC.

Taken together, these findings contribute to the body of literature on L2 speech fluency and intelligibility by extending our understanding of the role of pauses in L2 speech evaluation. Specifically, disfluent speech with many non-juncture pauses may impede listeners' smooth processing and understanding L2 speech, suggesting that at least for intermediate-proficiency speakers, improving fluency, without changing their accent other linguistic aspects, may effectively increase their speech intelligibility and acceptability. More pedagogical studies are needed to explore effective teaching practices in order to achieve this goal. It is also desirable to see whether these teaching practices can be applied to L2 learners of a broader spectrum of proficiency, ranging from beginning learners to more advanced ones. Finally, this study has implications for L2 speech rating in the context of testing and assessment. Currently the majority of English spoken tests are rated by human raters, who have different working memory capacities and it may or may not affect their ratings. Thus it would be interesting to examine the effect of individual WMC differences among trained raters and whether the magnitude of the effect mirrors the patterns observed among nave listeners, and further, how this effect may be compensated by specific training.

Last but not the least, the present study has broad implications in everyday settings where multiple tasks compete for limited capacity of cognitive resources. For example, in an algebra class taught by an L2 speaker, recognizing the accented speech produced by the instructor may have already consumed much of the available processing resources, although it is only the first step toward comprehending the lecture and learning the mathematical concepts, tasks which are in themselves cognitively demanding. If fluent yet accented speech is less effortful to understand than disfluent and accented speech, then by increasing his or her fluency level the instructor can effectively help students to free up scarce cognitive resources and thus to better understand the lecture content. Moreover, by reducing the effort necessary to accomplish the primary classroom goal (e.g. learning calculus), improving fluency may also introduce a positive attitude towards the instructor's accent, which may further exert positive effect on students' learning outcomes.

5.2 Limitations

The present study is not spare of limitations. First of all, the number of L2 speech samples (20) and the number of listeners participating in the two experiments (Experiment I: 10 for word intelligibility test, and 20 for subjective evaluation; Experiment II: 20 for each speech type) were both relatively small. Had the sample sizes been larger, more sophisticated measures and analytical methods (such as structural equation modeling) could have been used to allow the testing of more complex hypotheses.

Moreover, the L2 speech samples used in this study were of relatively simple L1 profile, since they were produced by native speakers of only Chinese and Korean, both of which are East Asian languages. If the study had included L2 speakers from more diverse L1 backgrounds (e.g., European, Middle Eastern, and African languages), then perhaps more fine-grained acoustic differences might have been found to be not only related to fluency but also to segmental and suprasegmental features. Additionally,

these speech samples were obtained from an oral test instead of a strictly designed experiment where speech content and prompts are meticulously controlled, and they were not recorded in sound-attenuated booths where noise is maximally reduced. The presence of noise may have some negative effect on the acoustic analysis of the spectral characteristics of segments, especially with vowels and fricatives.

The use of only native speakers of American English as the listeners (or the raters) in this study did not reveal the full picture of L2 speech quality evaluation. While it has shown how people would evaluate L2 speech in a typical English as a second language (ESL) context, it is not clear whether these findings may be extended to listeners who are non-native speakers of English. As prior studies have reported that L2 listeners tend to find it easier to understand L2 speakers of the same L1 background (Major et al., 2002; Smith & Bisazza, 1982), it is possible that the evaluations of listening effort, intelligibility, and acceptability of L2 speech by L2 listeners may differ from those of L1 listeners. Future studies are thus particularly needed to address L2 English evaluation in the English as an International Language (EIL) context by diversifying the linguistic background of listeners, including both L1 and L2 listeners.

It should also be noted that Experiment II only manipulated pausing of the intermediate-proficiency speech, and it remains unclear whether the same perceptual benefit would be observed had the high-proficiency speech been manipulated in the same manner. An interesting follow-up study would be to further explore the effect of improved fluency on the subjective evaluation of L2 speakers who have already attained high proficiency.

Finally, a note of caution is warranted with respect to the use of subjective measures of listening effort. Recent research suggests that such subjective ratings are not necessarily consistent with more direct measures of listening effort. Indeed, the correlation between subjective and objective measures of listening effort appears to be either weak or absent in some cases (Gosselin & Gagne, 2011; Zekveld et al., 2011), suggesting that the two types of measures may be assessing different aspects of listening effort. In the present study, the high correlation between listening effort ratings

and scores of subjective intelligibility and acceptability suggests that listeners may conflate perceived performance with perceived effort. Further research is necessary to investigate the relationship between actual cognitive demand and the subjective perception of effortfulness.

5.3 Directions for future studies

There are many directions that future studies could build upon the present study. One possible direction is to examine how fluency measures predict L2 speech samples that characterize a wider range of proficiency levels. In the current study, fluency, instead of phonetic pronunciation features, appeared to differentiate intermediate- from high-proficiency L2 speakers, which may partially attribute to the fact that these L2 speakers were all comparatively advanced English learners (US graduate students). A follow-up study could include L2 speakers of, for example, low English proficiency as well as those passing for native speakers, in order to examine whether there is some interaction between acoustic measures of fluency, segmental, and suprasegmental features, and what the relative weighting is for the contribution of the various speaker-related factors to listeners' perceptual evaluations of L2 speech.

The use of listener judgment in this study entails certain degree of subjectivity, in particular that subjective ratings may be affected by listeners' personal bias towards different L2 accents. Therefore, a second direction for future study is to tease apart the possible confounding effect of language attitude, and to investigate the possible interaction between L2 accent and attitude and how it may affect perception of L2 intelligibility and acceptability. For example, studies can address question such as what acoustic-phonetic properties may trigger different attitudes towards L2 speech, and whether these features are more contingent to fluency or segmental pronunciation. Identifying these features is not only of theoretical importance, but also has pedagogical applications. Specifically, it may help enhance the effectiveness of L2 pronunciation teaching by targeting at those high-value features (speech characteris-

tics that can more easily to trigger positive attitudinal changes). At the same time, it is also worthwhile to explore ways to teach listeners how to listen to and understand L2 speech while also building positive attitude towards it.

Furthermore, the finding that L2 speech evaluation may be susceptible to listeners' differences in WMC is worthy of more in-depth examination. It may have major impact on L2 speech rating and rater training in the context of testing and assessment.

Finally, future studies could address the limitations of this study by 1) expanding stimulus and participant sizes; 2) diversify the linguistic background of both listeners and speakers so as to investigate in more depth issues such as mutual intelligibility and its cognitive effect; and 3) implementing more complex manipulations to L2 speech samples. Moreover, physiological and psychophysiological measures can be introduced to directly measure cognitive load in comparison with subjective evaluations.

REFERENCES

REFERENCES

Abramson, A. S., & Lisker, L. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422.

Abutalebi, J., Cappa, S. F., & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism: Language and Cognition*, *4*(02), 179–190.

Abutalebi, J., Cappa, S. F., & Perani, D. (2005). What can functional neuroimaging tell us about the bilingual brain. In J. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 497–515). Oxford, UK: Oxford University Press.

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*(02), 81–96.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*(6), 3171–3183.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*(s2), 49–78.

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research*, *35*(2), 296–308.

Baddeley, A. D. (1999). *Essentials of human memory*. Psychology Press.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610.

Berns, M. (2008). World Englishes, English as a lingua franca, and intelligibility. *World Englishes*, *27*(3-4), 327–334.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Blake, C. G. (1996). *The potential of text-based internet chats for improving ESL oral fluency* (Unpublished doctoral dissertation). West Lafayette, IN.

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer (Version 5.3.51) [Computer software].* `http://www.praat.org`.

Bologna, W. J., Chatterjee, M., & Dubno, J. R. (2013). Perceived listening effort for a tonal task with contralateral competing signals. *The Journal of the Acoustical Society of America*, *134*(4), EL352–EL358.

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, *14*(4), 325–337.

Boothroyd, A. (1985). Evaluation of speech production of the hearing impaired: Some benefits of forced-choice testing. *Journal of Speech, Language, and Hearing Research*, *28*(2), 185–196.

Bosker, H. R., Quene, H., Sanders, T., & de Jong, N. H. (2014). The perception of fluency in native and non-native speech. *Language Learning*, *64*(3), 579–614.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255–272.

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language and Speech*, *24*(3), 207–221.

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning*, *22*(2), 203–217.

Cargile, A. C. (1997). Attitudes toward Chinese-accented speech: An investigation in two contexts. *Journal of Language and Social Psychology*, *16*(4), 434–443.

Caruso, A. J., & Burton, E. K. (1987). Temporal acoustic measures of dysarthria associated with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *30*(1), 80–87.

Catford, J. C. (1950). Intelligibility. *ELT Journal*(1), 7–15.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544.

Chao, K.-Y., & Chen, L.-M. (2008). A cross-linguistic study of voice onset time in stop consonant productions. *Computational Linguistics and Chinese Language Processing*, *13*(2), 215–232.

Chiba, R., Matsuura, H., & Yamamoto, A. (1995). Japanese attitudes toward English accents. *World Englishes*, *14*(1), 77–86.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, *51*(3), 175–198.

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*(3), 351–372.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552.

Crookes, G. (1991). Second language speech production research. *Studies in Second Language Acquisition*, *13*(02), 113–131.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

Cutler, A., Smits, R., & Cooper, N. (2005). Vowel perception: Effects of non-native language vs. non-native dialect. *Speech Communication*, *47*(1), 32–42.

de Bot, K. (1992). A bilingual production model: Levelt's "speaking" model adapted. *Applied Linguistics*, *13*(1), 1-24.

Deese, J. (1980). Pauses, prosody, and the demands of production in language. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech, studies in honour of Frieda Goldman-Eisler* (pp. 69–84). Berlin: Mouton de Gruyter.

de Johnson, T. H., Oconnell, D. C., & Sabin, E. J. (1979). Temporal analysis of English and Spanish narratives. *Bulletin of the Psychonomic Society*, *13*(6), 347–350.

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *The 6th workshop on disfluency in spontaneous speech (diss)* (pp. 17–20).

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition*, *19*(1), 1–16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379–397.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(04), 476–490.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163–185.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(04), 533–557.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679.

Deschamps, A. (1980). The syntactical distribution of pauses in English spoken as a second language by French students. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 255–262). The Hague, Netherlands: Mouton.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders*, *47*(2), 189–193.

Dyson, A. T., & Robinson, T. W. (1987). The effect of phonological analysis procedure on the selection of potential remediation targets. *Language, Speech, and Hearing Services in Schools*, *18*(4), 364–377.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope*, *58*(9), 955–991.

Engle, R. W., & Oransky, N. (1999). Multi-store versus dynamic models of temporary storage in memory. In R. Sternberg (Ed.), *The nature of cognition* (pp. 515–555). Cambridge, Massachusetts: MIT Press.

Ernestus, M., & Mark, W. M. (2004). Distinctive phonological features differ in relevance for both spoken and written word recognition. *Journal of Brain and Language*, *90*, 378–392.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern british english sentences. *The Journal of the Acoustical Society of America*, *115*(1), 352–361.

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*(6), 3814–3826.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1241–1255.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing*, *13*(2), 80–86.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423.

Fillmore, C. (1979). On fluency. In C. Fillmore & D. Kempler (Eds.), *Individual diflerences in language ability and language behavior* (pp. 85–102). New York: Academic Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Baltimore, MD: York Press.

Flipsen, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics*, *20*(4), 303–312.

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.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668–1680.

Francis, A. L., Ciocca, V., Wong, V., & Chan, J. (2006). Is fundamental frequency a cue to aspiration in initial stops? *Journal of the Acoustical Society of America*, *120*(5), 2884–2895.

Francis, A. L., & Nusbaum, H. C. (1999). Evaluating the quality of synthetic speech. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 63–97). Springer.

Fraser, S., Gagne, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, *53*(1), 18–33.

Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam: John Benjamins.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287–291.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65–87.

Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology*, *24*(1), 63–68.

Gevins, A., & Cutillo, B. (1993). Neuroelectric evidence for distributed processing in human working memory. *Electroencephalogr. Clin. Neurophysiol*, *87*, 128–143.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1), 103–138.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96–106.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Gosselin, P. A., & Gagne, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, *54*(3), 944–958.

Griffiths, R. (1991). Pausological research in an l2 context: A rationale, and review of selected studies. *Applied Linguistics*, *12*(4), 345–64.

Grosjean, F. (1980). Temporal variables within and between languages. In H. W. Dechert & M. Raupach (Eds.), *Towards a cross-linguistic assessment of speech production* (pp. 39–53). Frankfurt: Lang.

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de langlais et du français: vitesse de parole et variables composantes, phénomènes dhésitation. *Phonetica*, *31*(3-4), 144–184.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology*, *44*(10), 574–583.

Harberlandt, K. (1994). *Cognitive psychology.* Boston: Allyn and Bacon.

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics*, *36*(4), 664–679.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108–3118.

Hecker, M., & Williams, C. E. (1966). Choice of reference conditions for speech performance tests. *The Journal of the Acoustical Society of America*, *37*(5), 158-166.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45*(3), 573–584.

Holmes, V. (1995). A crosslinguistic comparison of the production of utterances in discourse. *Cognition*, *54*(2), 169–207.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

Institute, A. N. S. (1989). *Method for measuring the intelligibility of speech over communication systems (ANSI S3. 2-1989 R1995 / Acoustical Society of America Catalog No. 85-1989).* New York, NY: Acoustical Society of America.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals.* London: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83–103.

Jenkins, J. (2007). *English as a lingua franca: Attitude and identity.* London: Oxford University Press.

Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into english as a lingua franca. *Language Teaching*, *44*(03), 281–315.

Jongman, A., Wade, T., & Sereno, J. (2003). On improving the perception of foreign-accented speech. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 1561–1564).

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In Q. Randolph & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge: Cambridge University Press.

Kachru, B. B. (1986). *The alchemy of English: The spread, functions, and models of non-native Englishes*. Urbana and Chicago: University of Illinois Press.

Kachru, B. B. (1992). *The other tongue: English across cultures*. Urbana and Chicago: University of Illinois Press.

Kahneman, D. (1973). *Attention and effort*. Citeseer.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66–71.

Kent, R., & Netsell, R. (1978). Articulatory abnormalities in athetoid cerebral palsy. *Journal of Speech and Hearing Disorders*, *43*(3), 353–373.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499.

Klein, H. B. (1984). Procedure for maximizing phonological information from single-word responses. *Language, Speech, and Hearing Services in Schools*, *15*(4), 267–274.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300.

Kormos, J. (2014). *Speech production and second language acquisition*. Routledge.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *International Journal of Audiology*, *36*(3), 155–164.

Kramer, S. E., Zekveld, A. A., & Houtgast, T. (2009). Measuring cognitive factors in speech comprehension: The value of using the Text Reception Threshold test as a visual equivalent of the SRT test. *Scandinavian Journal of Psychology*, *50*(5), 507–515.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "Perceptual Magnet Effect". In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121–154). Baltimore, MD: York Press.

Kwiatkowski, J., & Shriberg, L. D. (1992). Intelligibility assessment in developmental phonological disordersaccuracy of caregiver gloss. *Journal of Speech, Language, and Hearing Research*, *35*(5), 1095–1104.

Lecumberri, M. L. G., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, *119*(4), 2445–2454.

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11), 864–886.

Lennon, P. (1984). Retelling a story in English as a second language. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions.* Tubingen, Germany: Gunter Narr Veriag.

Lennon, P. (1990a). The advanced learner at large in the L2 community: Developments in spoken performance. *International Review of Applied Linguistics in Language Teaching*, *28*(4), 309–324.

Lennon, P. (1990b). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Michigan: The University of Michigan Press.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01), 1–38.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken english, tesol and applied linguistics: Challenges for theory and practice* (pp. 245–270). New York, NY: Palgrave Macmillan.

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech*, *29*(1), 3–11.

Liu, H.-M., Tseng, C.-H., & Tsao, F.-M. (2000). Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *Clinical Linguistics & Phonetics*, *14*(6), 447–464.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and. natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *25*(1), 17–32.

Lunner, T. (2010). Designing ha signal processing to reduce demand on working memory. *The Hearing Journal*, *63*(8), 28–30.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indices of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113.

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing*, *36*(1), 145–154.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*(4), 381–400.

Major, R. C. (1987). Phonological similarity, markedness, and rate of L2 acquisition. *Studies in Second Language Acquisition*, *9*(01), 63–82.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Routledge.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*(2), 173–190.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962–3973.

Matsuura, H., Chiba, R., & Yamamoto, A. (1994). Japanese college students attitudes towards non-native varieties of English. In D. Graddol & J. Swann (Eds.), *Evaluating language* (pp. 52–61). Clevedon, England: Multilingual Matters.

McKenzie, R. M. (2008). Social factors and non-native attitudes towards varieties of spoken English: A Japanese case study. *International Journal of Applied Linguistics*, *18*(1), 63–88.

McLennan, C. T., & Luce, P. A. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 591–609). Malden, MA: Blackwell.

Metz, D. E., Samar, V. J., Schiavetti, N., Sitler, R. W., & Whitehead, R. L. (1985). Acoustic dimensions of hearing-impaired speakers' intelligibility. *Journal of Speech, Language, and Hearing Research*, *28*(3), 345–355.

Monsen, R. B. (1978). Toward measuring how well hearing-impaired children speak. *Journal of Speech, Language, and Hearing Research*, *21*(2), 197–219.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech, Language, and Hearing Research*, *35*(2), 259–273.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, *20*(02), 139–154.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193–218). Amsterdam / Philadelphia: John Benjamins Amsterdam.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(04), 451–468.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(01), 111–131.

Nelson, C. L. (2011). *Intelligibility in world Englishes: Theory and application*. New York, NY: Routledge.

Nesdale, D., & Rooney, R. (1996). Evaluations and stereotyping of accented speakers by pre-adolescent children. *Journal of Language and Social Psychology*, *15*(2), 133–154.

Nickerson, R., & Stevens, K. (1980). Approaches to the study of the relationship between intelligibility and physical properties of speech. In J. Subtelny (Ed.), *Speech assessment and speech improvement for the hearing impaired* (pp. 338–364). Washington, DC: AG Bell Association for the Deaf.

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, *17*(2), 235–242.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines* (pp. 113–157). San Diego: Academic Press.

Nye, P., & Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). *Haskins Laboratories Status Report on Speech Research, SR*, *33*, 77–91.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59.

Pals, C., Sarampalis, A., & Başkent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1075–1084.

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly*, *20*(2), 207–225.

Pichora-Fuller, M. K. (2006). Perceptual effort and apparent cognitive decline: Implications for audiologic rehabilitation. In *Seminars in hearing* (Vol. 27, pp. 284–293).

Pichora-Fuller, M. K., et al. (2003). Processing speed and timing in aging adults: psychoacoustics, speech perception, and comprehension. *International Journal of Audiology*, *42*, S59–S67.

Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics*, *26*, 219–233.

Pickering, L. (2009). Intonation as a pragmatic resource in elf interaction. *Intercultural Pragmatics*, *6*(2), 235–255.

Pittman, A. (2011). Childrens performance in complex listening conditions: Effects of hearing loss and digital noise reduction. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1224–1239.

Raupach, M. (1980). Temporal variables in first and second language speech production. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 263–270). The Hague: Mouton.

Raupach, M. (1987). Procedural knowledge in advanced learners of a foreign language. In J. Coleman & R. Towell (Eds.), *The advanced language learner* (pp. 123–157). London: AFLS/CILT.

Riazantseva, A. (2001). Second language proficiency and pausing a study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(04), 497–526.

Ricketts, T. A., & Hornsby, B. W. (2005). Sound quality measures for speech in noise through a commercial hearing aid implementing. *Journal of the American Academy of Audiology*, *16*(5), 270–277.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441.

Riggenbach, H. (2000). *Perspectives on fluency*. Ann Arbor, Michigan: The University of Michigan Press.

Rogers, C. L., & Dalby, J. (2005). Forced-choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research*, *48*(2), 306–322.

Rogers, C. L., Dalby, J. M., & Nishi, K. (2001). Effects of noise and proficiency level on intelligibility of Chinese-accented English. *The Journal of the Acoustical Society of America*, *109*(5), 2473–2473.

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, *14*(3), 337–353.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61–86.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology*, *23*(8), 577–589.

Sajavaara, K. (1987). Second language speech production factors affecting fluency. In H. Dechert & M. Raupach (Eds.), *Psycholinguistic models of production* (pp. 45–65). Norwood, NJ: Ablex.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, *52*(5), 1230–1240.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders* (pp. 11–34). Philadelphia, PA: John Benjamins.

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research*, *42*(1), 56–64.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(04), 357–385.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, & S. Greenspan (Eds.), *Applied speech technology* (pp. 195–231). Boca Raton, Ann Arbor, London, Tokyo: CRC Press.

Schmidt-Nielsen, A., Kallman, H. J., & Meijer, C. (1990). Dual-task performance using degraded speech in a sentence-verification task. *Bulletin of the Psychonomic Society*, *28*(1), 7–10.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime reference guide*. Psychology Software Tools, Incorporated.

Scovel, T. (1988). *A time to speak: A psycholinguistic investigation into the critical period for human speech*. New York: Harper and Row.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, *11*(2), 133–158.

Seidlhofer, B. (2005). English as a lingua franca. *ELT journal*, *59*(4), 339.

Seidlhofer, B., Breiteneder, A., & Pitzl, M.-L. (2006). English as a lingua franca in Europe: Challenges for applied linguistics. *Annual Review of Applied Linguistics*, *26*, 3–34.

Sewell, A. (2010). Research methods and intelligibility studies. *World Englishes*, *29*(2), 257–269.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Karhru (Ed.), *The other tongue: English across cultures* (pp. 75–90). Urbana, IL: University of Illinois Press.

Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning*, *32*(2), 259–269.

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, *4*(3), 333–342.

Smith, L. E., & Nelson, C. L. (2008). World Englishes and issues of intelligibility. In B. Kachru, Y. Kachru, & N. C. L (Eds.), *The handbook of world englishes* (pp. 428–435). Malden: Blackwell Publishing.

Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly*, 371–380.

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, *13*(5), 335–349.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stevens, S. S. (1999). On the theory of scales of measurement. *Clinical Linguistics & Phonetics*, *13*(5), 335–349.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detaila). *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099.

Todd, A. E., Edwards, J. R., & Litovsky, R. Y. (2011). Production of contrast between sibilant fricatives by children with cochlear implantsa. *The Journal of the Acoustical Society of America*, *130*(6), 3969–3979.

Towell, R. (1987). Approaches to the analysis of the oral language development of the advanced learner. In J. Coleman & R. Towell (Eds.), *The advanced language learner* (pp. 157–181). London: SUFLRA, AFLS.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *IRAL*, *40*(2), 117–150.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119.

Towell, R., & Hawkins, R. D. (1994). *Approaches to second language acquisition*. Multilingual matters.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(01), 1–30.

van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*, 1–4.

van Gelderen, A. (1994). Prediction of global ratings of fluency and delivery in narrative discourse by linguistic and phonetic measures-oral performances of students aged 11-12 years. *Language Testing*, *11*(3), 291–319.

van Rooy, S. C. (2009). Intelligibility and perceptions of English proficiency. *World Englishes*, *28*(1), 15–34.

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, *52*(1), 23–36.

Weismer, G., Kent, R. D., Hodge, M., & Martin, R. (1988). The acoustic signature for intelligibility test words. *The Journal of the Acoustical Society of America*, *84*(4), 1281–1291.

Weismer, G., Martin, R., & Kent, R. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders* (pp. 67–118). Philadelphia, PA: Benjamins Amsterdam.

Wikibooks.org. (2014). *Andrew's monotone chain convex hull algorithm.* `http://en.wikibooks.org/wiki/Algorithm_Implementation`.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *57*(4), 573–589.

Xu, C. X., & Xu, Y. (2003). Effects of consonant aspiration on mandarin tones. *Journal of the International Phonetic Association*, *33*(02), 165–181.

Yorkston, K. M., & Beukelman, D. R. (1984). *Assessment of intelligibility of dysarthric speech.* Tigard, OR: C. C. Publications.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage*, *32*(4), 1826–1836.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510.

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, *123*(6), 4498–4513.

Zlegler, W., & Von Cramon, D. (1983). Vowel distortion in traumatic dysarthria: A formant study. *Phonetica*, *40*(1), 63–78.

VITA

VITA

Mengxi Lin is a Ph.D. candidate at the Linguistics Program of Purdue University. Prior to joining the graduate program at Purdue, she obtained her M.A. and B.A. in English Literature and Language from the English Department of Peking University, China.

Her research focuses on the perception and production of second language (L2) speech and accented speech. She is particularly interested in exploring the acoustic and perceptual factors that affects speech intelligibility, as well as the cognitive mechanism that drives listeners to understand L2 and accented speech. Her recent research also examines the phonetic flexibility and adaptability of bilinguals' speech production. In addition, she has studied cross-linguistic speech perception between tonal and non-tonal languages, and have worked on sociolinguistic-oriented projects on language attitude towards world Englishes.

Areas of specialization: speech perception and production, listening effort, experimental and acoustic phonetics, second language intelligibility, second language acquisition