Pronunciation Pedagogy and Speech Perception: Three Studies

Jennifer Ann Foote

A Thesis

In the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Education) at

Concordia University

Montreal, Quebec, Canada

June 2015

CONCORDIA UNIVERISTY

School of Graduate Studies

This is to certify that the thesis prepared

By: Jennifer Ann Foote

Entitled: Pronunciation Pedagogy and Speech Perception: Three Studies

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Education)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____ Chair
Dr. Joanna White

____ External Examiner
Dr. John Levis

____ External to Program
Dr. Denis Liakin

____ Examiner
Dr. Kim McDonough

____ Examiner
Dr. Sara Kennedy

____ Thesis Supervisor
Dr. Pavel Trofimovich

Approved by ____
Dr. Richard Schmid, Department Chair

June 22, 2015 ____
Dr. André Roy, Dean of Faculty

ABSTRACT

Pronunciation Pedagogy and Speech Perception: Three Studies

Jennifer Ann Foote, Ph.D. (ABD)

Concordia University, 2015

This dissertation investigates how second language (L2) speakers perceive non-native speech and how language learners can be helped to better perceive differences between their L2 output and target language speech, thus facilitating improvements in pronunciation.

Study 1 investigated which dimensions underlie the perception of L2 speech by L2 listeners. Fifteen L2 listeners (and 10 native English listeners who served as a baseline group) rated 30 L2 audio-recordings from controlled reading and interview tasks for dissimilarity, using pairwise comparisons. Multidimensional scaling analyses revealed that fluency and global aspects of the speakers' pronunciation explained listener judgments but that there was little agreement across the L2 and native listener groups.

Study 2 investigated the role of language background in comprehensibility judgments. English speakers from Mandarin, French, Hindi, and English language backgrounds (10 per group) rated the speech of 30 L2 speakers from the same language backgrounds for comprehensibility and provided verbal reports about each rating. They then rated the speakers for segmental and word stress errors, intonation, and speech rate. Correlations between the speech measures and comprehensibility ratings for each L2 listener-speaker group and hierarchical regressions carried out for each L2 listener group revealed that different speech measures contributed to comprehensibility for different listener-speaker groups, with language background accounting for an additional six percent of the variance in comprehensibility ratings for the Mandarin listeners after the linguistic variables were taken into account. Analysis of the verbal reports for whether the listeners attributed their ratings to the speakers' language background showed only a moderate relationship to the quantitative data.

Study 3 investigated the efficacy of shadowing, a common pronunciation practice technique. Sixteen learners practiced shadowing with iPods for eight weeks. Two language tasks (shadowing task, extemporaneous speaking task) were administered as pre-, mid-, and post-tests, and were rated by 21 L1 English listeners. The shadowing task was evaluated for the learners' ability to imitate a speech model, and the extemporaneous speaking task was rated for comprehensibility, accentedness, and fluency. Results indicated that the learners improved significantly in all speaking measures apart from accentedness and were largely positive about the activities.

Acknowledgments

There have been many people who have helped me get to this stage of my journey. I have been fortunate to work people who have been not only excellent scholars but also unparalleled mentors. First and foremost I would like to thank my supervisor, Pavel Trofimovich, whose support and guidance went far beyond what is required of a doctoral supervisor. I will never cease to be grateful for having the opportunity to be his student. I would also like to thank my committee members, Sara Kennedy and Kim McDonough, both of whom have been helped me grow as a scholar and offered immeasurable assistance in the writing of this dissertation. I would like to thank my internal examiner, Dr. Denis Liakin, and external examiner, John Levis as well as the chair of my defense, Joanna White. I would also like to thank Tracey Derwing, whose official role as my supervisor ended many years ago when I completed my Master's degree, but whose role as a mentor and friend has continued throughout my academic career.

In addition to the support I have received in academia, I have also had incredible support from friends and family throughout my PhD. While I don't have space to thank everyone, I would like to mention my parents whose support was unwavering, Anglia Redding and Jihan Rabah who were always there when needed, and Tieja Thomas for offering assistance right when it was needed most.

This research was supported by the Social Sciences and Humanities Research Council.

Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada

Contribution of Authors

Study 1 of this thesis is coauthored with my supervisor, Dr. Pavel Trofimovich. As the first author, I took the lead on this project. The original idea for this manuscript was based on the research conducted during my M.A. degree. The speech samples used in this study came from a research project conducted by Dr. Trofimovich, but the listener data were collected especially for this study. I took the lead in collecting and analysing the data and training the RAs. I also conducted the statistical analyses, with assistance from Dr. Trofimovich, and was in charge of writing the manuscript. Studies 2 and 3 of this thesis are single authored.

Table of Contents

List of Figures

List of Tables

Glossary

Accentedness: The meaning of accent is generally well understood by most people, and its definition in SLA research does not differ dramatically from its common meaning. Accentedness usually denotes the degree to which listeners perceive L2 speech as sounding different from what is considered to be nativelike (Munro & Derwing, 1995b). Accentedness is related to comprehensibility and intelligibility, but it is distinct from these terms because research has established that an accent, even a strong accent, does not necessarily make a person difficult to understand (Derwing & Munro, 1995a). Accentedness is usually measured in a similar way to comprehensibility, using rating scales.

Fluency: In vernacular speech, fluency is often used similarly to proficiency. However, in applied linguistics research, this term refers to the extent to which speech sounds relatively smooth and effortless. Fluency is affected by a number of variables such pausing, speech rate, self-corrections, etc.

Comprehensibility: The term refers to how difficult listeners find second language (L2) speech to understand; if they struggle and must listen carefully to understand an utterance, then this utterance would be considered as being low in comprehensibility, even if ultimately the message is understood (e.g., Derwing, Munro, & Thomson, 2008). In research, comprehensibility is usually measured using rating scales (e.g., Munro & Derwing, 1995a; Isaacs & Trofimovich, 2012).

Intelligibility: Intelligibility refers to actual difficulty in understanding an utterance. Rather than using ratings, intelligibility is usually measured more objectively. For example, to derive a measure of intelligibility, listeners may be asked to transcribe an utterance rather than rate it (e.g., Munro, Derwing, & Morton, 2006). However, it should be noted that while this specific definition of intelligibility has become well accepted in second language acquisition (SLA) literature, it is also used more broadly to mean speech that is understandable, even by researchers in this field.

Chapter 1: General Introduction

Introduction

For adult second language (L2) speakers, pronunciation often poses a serious challenge. Fraser (2010) notes that for many learners, pronunciation is "simultaneously the most difficult of the language skills and the one they most aspire to master" (p. 358). This makes it surprising that pronunciation has been historically under-represented and overlooked in comparison to other areas of second language acquisition (SLA) research. This neglect has been widely discussed (e.g., Celce-Murcia, Brinton, & Goodwin; 1996; Derwing & Munro, 2005; Gilbert, 1994; Isaacs, 2009), and even documented through surveys of how well pronunciation has been represented in SLA research (Brown, 1991; Deng et al., 2009). This lack of interest in pronunciation research has been mirrored by a lack of presence in language teaching. There have been surveys of English L2 instructors from a number of countries including Canada (Breitkreutz, Derwing, & Rossiter, 2001; Foote, Holtby, & Derwing, 2012), Australia (Burns, 2006; Macdonald, 2002), Brazil (Buss, 2015), and Finland, France, Germany, Macedonia, Poland, Spain and Switzerland (Henderson et al., 2012), which showed that many instructors lack confidence in their abilities to teach pronunciation or, at the least, feel they would benefit from greater training in pronunciation pedagogy. Further, an analysis of a 112,595-word corpus taken from over 40 hours of observation of the ESL courses in Quebec, Canada, found that pronunciation accounted for only 10% of language related episodes in the observed classes (Foote, Trofimovich, Collins, & Urzúa, 2013).

Fortunately, pronunciation research and pedagogy are beginning to catch up to other language skills. A new journal of L2 pronunciation has recently been launched, a steady stream of pronunciation-related research is being published, and an increasing number of pronunciation materials are appearing on the market, all of which led Thomson and Derwing (2014) to declare that "the tide has shifted" in the area of L2 pronunciation (p. 1). Nonetheless, research has still not provided a comprehensive answer to some of the fundamental questions about pronunciation learning and teaching. Many of these questions relate to speech perception. The basic concept that perception generally precedes production is largely accepted (e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999) and interventions designed to improve L2 listeners' perception of sounds in their L2 have been demonstrated to lead to improvements in production, even when explicit training in production is not given (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Thomson, 2011). However, while studies targeting L2 speech perception have become more common, particularly those carried out within linguistic and psycholinguistic research traditions, and research looking at how first language (L1) speakers perceive L2 speech has made headway, far less is known about how L2 speakers perceive target L2 speech, especially the speech by other L2 speakers. There is also a need for more research investigating how language learners can be helped to better perceive differences between their speech and that of their target language, thus facilitating improvements in pronunciation. Without more evidence-based answers to these questions, language instructors and learners alike are too often forced to rely on intuition rather than evidence when making pedagogical choices related to pronunciation. Study 1 and Study 2 of this dissertation are primarily concerned with the first issue: how L2 speakers perceive English speech by other L2 speakers. Study 3 focuses on the second issue: how learners can be helped to better perceive differences between their speech and that of the target language in order to improve their pronunciation.

The first issue, how L2 speakers perceive L2 speech, is of interest because of an increasing recognition that many L2 speakers are using a shared L2 to communicate with other L2 speakers rather than native speakers (NSs), which is particularly true for English (Nelson, 2011). While pronunciation research focusing on L2-L2 interlocutors is increasing, many of the findings related to which elements of pronunciation are most important for learners still assume a NS listener (e.g., Hahn; 2004; Kang, Rubin, & Pickering, 2010; Munro & Derwing, 2006; Tajima, Port, & Dalby, 1997; Trofimovich & Isaacs, 2012). A better understanding of how the language backgrounds of speakers and listeners impact speech perception is needed in order to know how to set teaching priorities that will prepare learners to successfully communicate with a wide range of interlocutors, not just L1 users of the language. Study 1 addresses this issue by investigating which aspects of L2 speech are most salient to L2 listeners of English and are therefore more likely to be learned/acquired without explicit instruction. Two groups of listeners – a group of L2 speakers of English with mixed L1 backgrounds and group of L1 listeners – rated speech samples in two different tasks using a paired-comparison method, with listeners judging for each pair of speech samples how dissimilar they were from each other. A statistical technique called multidimensional scaling (MDS) was then used to plot the response patterns in an *n*-dimensional space. A wide range of speech features were then considered to understand which of these features could explain the dimensions underlying listeners' responses. One of the primary objectives of this study was to see which features of speech are salient to L2 listeners when listening to L2 speech, and thus likely to be noticed and attended to in L2-L2 interactions.

Study 2 addresses the issue of how L2 speakers perceive L2 speech by investigating the role of L1 background in listeners' understanding of L2 speech from English speakers from different language backgrounds. More specifically, this study examines whether L2 English speakers from different L1 language backgrounds (Mandarin, French, and Hindi) differ in which aspects of pronunciation (e.g., word stress, intonation, etc.) contribute to their understanding of L2 speech and whether the language background contributes to understanding above and beyond what can be explained by these specific aspects of pronunciation. Further, the study investigates whether L2 listeners attribute their ease or difficulty of understanding to the language backgrounds of the speakers.

Study 3 is primarily concerned with the second issue discussed: how learners can be helped to better perceive differences between their speech and that of the target language in order to improve their pronunciation. Specifically, this study investigates the utility of a popular, but under-researched pronunciation teaching technique called shadowing. This technique, which involves repeating and copying speech nearly simultaneously with a target recording, is commonly used in language classrooms. In this study, voice recorders are used with the shadowing activities to enable learners to analyze their own speech after shadowing. One of the most interesting aspects of shadowing in light of Study 1 and Study 2 is the ease with which it can be used to target different aspects of pronunciation at the same time; one shadowing activity can be used to focus on different problems with different learners, potentially allowing for differentiated instruction for learners with different pronunciation needs and communication goals.

In the overarching review of the literature that follows, I discuss some key issues in the areas of speech perception and pronunciation instruction. The first of these is a discussion of four core concepts in pronunciation research: accentedness, intelligibility, comprehensibility, and fluency. I then talk about the complexity that language background plays in the understanding of L2 speech and the challenges of setting teaching priorities in the classroom. I also discuss the problematic and uneasy relationship of pronunciation with broader theories of SLA research, and why it is often difficult to situate pronunciation studies within a theoretical framework. There is then a discussion of the friction between what are considered to be best practices in terms of general language teaching pedagogy and pronunciation pedagogy. The chapter concludes by explaining how the three studies that make up this dissertation tie into these broader concepts, thus allowing for a more in-depth discussion of the issues targeted by these studies than is possible within the confines of an article-length manuscript.

Overarching Review of the Literature

Accentedness, Intelligibility, Comprehensibility, and Fluency

When discussing pronunciation research and pronunciation instruction, it is important to contrast three related, yet distinct, concepts: accentedness, intelligibility, and comprehensibility. Accentedness has been defined differently by different researchers. Lippi-Green (1997) defines accent as "…the breakthrough of native language phonology into the target language" (p. 43), while Derwing and Munro (2009) define it as "the degree to which a speech sample differs from the local variety" (p. 476). These two definitions are similar; however, Derwing and Munro's definition focuses more on the listener's *perception* of differences between two varieties of English. In fact, in a recent publication, Derwing and Munro (in press) argue that accent "is, by definition, something that is noticed by listeners; therefore, there is no kind of accent other than a perceived accent" (p. 8). This suggests that research investigating accent should primarily focus on listeners' perceptions rather than on acoustic measurements of speech. Research investigating peoples' perception of accent has demonstrated that listeners are extremely sensitive to even small variations in accent. For example, Flege (1984) found that listeners could detect a foreign accent in speakers with fairly high accuracy, even when hearing only a portion of a /t/ sound. A wide range of speech characteristics contribute to the accentedness of speech (e.g., Anderson-Hsieh, Johnson, & Koehler, 1992) and Munro, Derwing, and Burgess (2010) found that even when more obvious aspects of accented speech, such as segmental cues, were removed by playing speech samples backwards and normalizing them for speech rate, listeners were still able to detect foreign accents at above-chance levels.

Accent, then, refers to any difference in phonology that a listener will notice between two speech varieties. Intelligibility, on the other hand, refers to whether "listeners can understand the speaker's intended message" (Derwing & Munro, in press, p. 1). In some cases, intelligibility measures try to assess whether listeners understood a message. For example, Munro and Derwing (1995b) had listeners assign truth values to statements by L2 speakers such as, "Elephants are big animals" and "Most people wear hats on their feet" (p. 292). However, in research studies, intelligibility is often operationalized more narrowly, equating *understanding* with simply being able to identify the actual words spoken rather than the message being conveyed. One of the most typical ways of measuring comprehensibility is to have listeners write down what they hear and count the number of words transcribed correctly (e.g., Bent & Bradlow, 2003; Derwing & Munro, 1997; Munro & Derwing, 1995a; Xie & Fowler, 2013). For this reason, intelligibility measures often do not capture understanding of a speaker's message, but rather of their words. Comprehensibility typically refers to listeners' perception of the ease or difficulty with which they comprehend an utterance. Accentedness and comprehensibility are typically measured using rating scales (e.g., Crowther, Trofimovich, Saito, & Isaacs, 2014; Derwing, Munro, Foote, Waugh, & Fleming, 2014; Derwing & Munro, 1997; Munro & Derwing, 1995a; Munro, Derwing, & Morton, 2006; Trofimovich & Isaacs, 2012).

While it is common to equate the presence of a strong accent to a person being difficult to understand, studies comparing accentedness to intelligibility and comprehensibility have found only a partial relationship (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995a, b). Generally, people who score low in terms of comprehensibility and intelligibility are also rated as highly accented, but the reverse is not necessarily true. People could have strong accents and still be very easy to understand. The relationship between intelligibility and comprehensibility is closer but the measures still do not correlate perfectly. This is not so surprising when the nature of the constructs is considered. In cases where intelligibility is measured using transcriptions, a person could identify all of the words a person speaks, but find either (a) that they must struggle to process the words or (b) that they do not understand what the person is saying, even though the words are clear. Interestingly, some studies with L2 rather than L1 listeners have found the relationship between intelligibility and comprehensibility to be weaker than those using L1 listeners (e.g., Kim, 2008; Matsuura; 2007).

Derwing and Munro (in press) clarify the relationship between these three constructs by breaking down the possible combinations of intelligibility/comprehensibility and intelligibility/accentedness an utterance could have for a particular listener. For intelligibility and comprehensibility, an utterance could have a high score for each, meaning it is "fully understood, little effort required," or have high intelligibility scores but low comprehensibility scores, meaning it is "fully understood" but that "great effort is required." (p. 6). It is also possible for both scores to be low, indicating that an utterance "is not fully understood" and "great effort is exerted" (p. 6). Finally, it is possible, though probably not common, that an utterance could be low in intelligibility but high in comprehensibility, meaning that the listener believed they understood something easily but actually missed the message. For intelligibility and accentedness, it is possible that a listener has no difficulty understanding the message, and the speaker has a strong accent, or that the message is easily understood and the accent is "barely noticeable" (p. 6). On the other hand, an utterance could be difficult to understand and be heavily accented. Derwing and Munro also note that a person could have speech that is not very intelligible but that is also not heavily accented. In this this case, pronunciation is not an issue, but the speaker may struggle with other aspects of the target language such as grammar or vocabulary choices. Of course, a similar set of comparisons could be drawn for comprehensibility and accentedness. While these distinctions may seem obvious, it is very common for people to assume that having a heavy accent automatically make a person difficult to understand. Further, it is important to note that being able to understand what a person says does not take into account difficulties that a listener may have in processing speech. For this reason, a speaker who has speech that is intelligible, but low in comprehensibility, may still benefit from instruction.

It should also be noted that some scholars, especially within *World Englishes* literature, discuss a fourth concept, *interpretability* (e.g., Kachru & Smith, 2008; Nelson, 2011). In this body of literature, comprehensibility is defined as the ability to understand a word's meaning in context, and interpretability is "the recognition by the hearer/reader of the intent or purpose of an utterance" (Kachru & Smith, 2008, p. 63). While I acknowledge that this contrast is both interesting and useful, interpretability falls outside the scope of research presented here. To avoid further confusion, the term *intelligibility* will be used in the broader definition given above of whether "listeners can understand the speaker's intended message" (Derwing & Munro, in press, p. 1) rather than through the narrower operationalized definitions which often refer to listeners' ability to identify words or transcribe utterance. Further, the definitions I have used here represent the constructs as they are used in my three studies, and as they are commonly used in research within this field. However, all three constructs have been defined and measured differently by different researchers at different times (see Isaacs, 2008, for an overview of different conceptualizations of intelligibility and comprehensibility).

Historically, there were two competing ideas about what the goal of pronunciation instruction should be, based on the *nativeness principle* or the *intelligibility principle* (for discussion of a full history of these two principles, see Levis, 2005). The former assumes both a desire and ability to achieve nativelike pronunciation, whereas the latter is concerned primarily with the aspects of pronunciation which are most likely to impede successful communication. However, despite the unfortunate fact that there are still people and products trying to sell the promise of *nativeness* to language learners, Isaacs and Trofimovich (2012) have noted that today "few L2 researchers and practitioners would disagree that intelligibility is the appropriate goal for L2 pronunciation instruction" (p. 477). This is due to the large amount of evidence demonstrating that regardless of whether or not learners *should* want to pursue nativelike accents, for those learning a language after puberty, it is very unlikely that nativelike pronunciation is possible (e.g., Flege & Frieda, 1995; Flege, Munro, & MacKay, 1995; Moyer, 1999).

Unfortunately, while researchers generally recognize that intelligibility should be the goal of pronunciation instruction, many studies targeting pronunciation interventions still fall under the nativeness principle. In a recent survey of 75 L2 pronunciation studies, Thomson and Derwing (2014) found that 63% implicitly aligned with the nativeness principle; in most cases, this meant that none of the measures used to test the success of a pronunciation intervention targeted intelligibility, but rather focused on features of accent which may or may not affect intelligibility or used acoustic measures of speech rather than listeners. Some researchers prefer computer-based measures, seeing them as more reliable (e.g., Hincks, 2003). However, by using such measures, it is impossible to know whether what was measured by the computer would have actually made the listeners easier to understand to human interlocutors. This is certainly not to say that studies that do not use measures of comprehensibility or intelligibility and do not use human raters are invalid. Any measure of improvement shows that a treatment has potential. However, improvements in intelligibility (or comprehensibility) measured with human listeners provide the strongest evidence that a treatment will lead to changes that will genuinely help a learner be better understood in real-world interactions.

Measures of comprehensibility are used in both Study 2 and Study 3 of this dissertation. In both cases, comprehensibility was chosen over intelligibility because comprehensibility captures the concept of *understanding* in way that is more useful for successful communication. A person who is able to speak words that can be identified, but whose speech is difficult to process and whose meaning does not come across easily, is likely to struggle with interlocutors despite having a high level of intelligibility. Study 3 also uses accentedness, in addition to comprehensibility, to evaluate L2 speakers although accentedness is given a lower priority in terms of judging the utility of the pronunciation intervention when compared to comprehensibility. Further, the measure used for accentedness is based on listener ratings, following Derwing and Munro's (in press) argument that a measure of accentedness is only meaningful in the context of listeners. Study 1 does not include measures of intelligibility, comprehensibility, or accentedness. However, this is because this study does not aim to understand what makes listeners intelligible or accented to each other, but rather seeks to uncover what is most salient to listeners when listening to L2 speech.

Finally, fluency is a construct that is used in some respect in all three studies in this dissertation. Fluency, as it is used here, does not refer to the broad commonly-used meaning which relates to overall proficiency (Lennon, 1990) but rather refers to "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (Ellis, 2003, p. 342). There are a number of speech characteristics that contribute to fluency, and it is often "operationalized through temporal measures such as speech rate, hesitations, and pausing" (O'Brien, 2014, p. 719). Studies also frequently use human raters to assess speech fluency (e.g., Derwing, Rossiter, Munro, & Thomson, 2004; Derwing et al. 2014; Lennon, 1990; O'Brien, 2014; Rossiter, 2009). Fluency has also been found to relate to comprehensibility (Derwing et al., 2004; Isaacs & Trofimovich, 2012).

All three studies in this dissertation involve fluency measure(s). Study 1 uses a number of temporal measures of speech including speech rate, filled and unfilled pauses, sample duration, and repetitions/self-corrections to capture various aspects of fluency. Study 2 relies on human raters to evaluate L2 speech fluency, through listener-based ratings of speech rate. Study 3 used L1 listener judgments of fluency.

The Importance of the L2 Interlocutor

As was mentioned above, research and pedagogy are moving away from the nativeness principle towards the intelligibility principle. However, the very notion of intelligibility raises an important question: *intelligible in which speaking/listening context*? As mentioned earlier, the majority of pronunciation research has assumed L1 users to be the target listeners. This is not necessarily a drawback in every study. For instance, Study 3 uses L1 listeners as raters to judge the effect of a pronunciation interaction because the intervention is relatively novel, and using an L1 group provided a higher level of control. However, too often assumptions are made that L2 learners are learning a language to speak primarily with NSs. In reality, with the spread of English around the globe, there are far more L2 users of English than there are NSs. To put this in perspective, Crystal (2003) estimates that approximately 329 million people speak English as an L1 while approximately 750 million speak English as an L2 – and this estimation is based on people who speak with at least a "medium level conversational competence" (p. 68).

The context in which English is used also varies widely. There is a vast literature documenting and describing World Englishes (see Kachru, Kachru, & Nelson, 2006, for an overview). One useful model for understanding the role of English in different parts of the world is Kachru's concentric circles model (1982, 1992, 1997). Kachru divides countries with English speakers into three groups: the Inner Circle, the Outer Circle, and the Expanding circle. The inner circle represents the traditional English bases of England and the original settler colonies (e.g., Australia, Canada, England, the United States, etc.). The Outer Circle represents countries that were colonized territories and where English plays a major role in society and government (e.g., India, Nigeria, Pakistan, etc.). The Expanding circle encompasses countries where English does not play a primary role, but where it is used as a foreign language (e.g., China, Brazil, Japan, etc.). Due to the complex nature of global interaction, speakers within these contexts are constantly moving and interacting. For example, an L1 Hindi-speaking student living in an Outer Circle country who has been schooled in English and has no intelligibility issues when speaking English with other Hindi speakers, may move to an Inner Circle country such as Canada to attend school. She may then work with classmates who are both L1 and L2 English speakers from Inner, Outer, and Expanding Circle countries. Levis (2005) discusses this when explaining the *intelligibility principle*, presenting first a four by four matrix, denoting the four speaker-listener combinations for NSs and non-native speakers (NNs) that can occur in interaction: NS-

NS, NS-NNS, NNS-NS, NNS-NNS. However, as Levis notes, when considering Kachru's full model, this actually expands to a nine-square matrix which includes speakers of English from Inner, Outer, and Expanding Circle countries.

There have been attempts to address this complexity in books such as Nelson's (2011) *Intelligibility in World Englishes: Theory and application* and Low's (2015) *Pronunciation for English as an International Language*. Perhaps the most well-known response has been the development of English as a lingua franca (ELF). The Vienna-Oxford International Corpus of English (www.univie.ac.at/voice) defines ELF as "an additionally acquired language system which serves as a common means of communication for speakers of different first languages" and this definition is commonly used by ELF researchers (e.g., Jenkins, 2012; Jenkins, Cogo, & Dewey, 2011). Jenkins (2002) has argued that because most speakers of English are now L2 speakers, using L1 speakers as a standard for intelligibly is obsolete. She proposes instead, a lingua franca core (LFC) syllabus for pronunciation that only considers aspects of pronunciation that are likely to cause problems between L2-speaking interlocutors. Jenkins (2012) also argues that ELF is not only for people who speak English in an Expanding Circle context. An English speaker from an Inner Circle country such as Canada or an Outer Circle country such as India, will not need ELF when talking to others who share their language background but may use ELF as an additional language when speaking with L2 users who do not share their language background. However, she also argues that "those for whom English is the L1 do not determine the linguistic 'agenda' of ELF" (p. 487). For this reason, most LFC researchers "specifically exclude" L1 speakers from their research (Jenkins, 2006, p. 160). The LFC has some differences from what research outside of ELF would suggest for successful communication. For instance, there is very little emphasis on suprasegmental aspects of pronunciation (e.g., rhythm, intonation, and stress), with all but contrastive stress (I want to BUY a car vs. I want to buy a CAR) eliminated from the LFC despite a general consensus that suprasegmental features are important to intelligibility (Derwing & Munro, 2005). For this reason, research on the LFC has not been without controversy. It has been criticized for a lack of sufficient evidence to support the specific components of the LFC (e.g., Dauer, 2005; Derwing & Munro, in press) and researchers have called into question many specific features of the LFC. It should be noted that teaching pronunciation as a lingua franca is about more than the LFC. Another major aspect of the LFC, is the idea of accommodation, whereby English language users adjust their speech to different interlocutors in different contexts in order ensure mutual intelligibility (e.g., Jenkins, 2012; Walker, 2010).

One of my main goals in this dissertation is to address the complexity of interaction between speakers from different language backgrounds speaking in a wide range of contexts. In this sense, the motivation for the current research is very similar to that which drives ELF research. However, L1 speakers are not "specifically excluded" from this dissertation, as L1 speakers also represent a group of speakers that L2 users are likely to interact with (Jenkins, 2006, p. 160). Further, the issue of L2-L2 communication is important in Inner Circle countries as well. For example, according to Statistics Canada (http://www12.statcan.gc.ca/), in 2011 Canada had a foreign-born population of 6,775,800 people and 72.8% of these people speak an L1 other than English or French. Study 1 and Study 2 of this dissertation try to capture at least a small piece of this diversity. Study 1 compares an L2 listener group with mixed L1 backgrounds to a group of L1 listeners. Study 2 investigates differences between listeners from specific L1 backgrounds, which represent language users from Inner, Outer, and Expanding Circle countries.

Choosing Instructional Priorities

Choosing instructional foci for pronunciation instruction is challenging even when the language backgrounds of the listeners are not taken into account. There has been research investigating the relative importance of segmentals (i.e., individual sounds) and suprasegmentals in L2 pronunciation (e.g., Anderson-Hsieh et al., 1992; Derwing, Munro, & Wiebe, 1998). There is now general agreement that both are important (Derwing & Munro, 2005, in press). A number of studies have attempted to look at how different speech characteristics can impact listeners' understanding of speech (e.g., Anderson-Hsieh & Koehler, 1998; Crowther et al., 2014; Hahn, 2004; Isaacs & Trofimovich, 2012; Kang et al., 2010; Munro & Derwing, 2001, 2006; Zielinski, 2008). However, less research has been conducted with L2 listeners. This situation is improving. There is the research related to the LFC which was already discussed in the previous section. Other studies have used L2 listeners when investigating such aspects of speech as the role of speaking rate (Derwing & Munro, 2001), lexical stress (Field, 2005) and fluency (Rossiter, 2009) in listener perceptions. There have also been a number of studies investigating the role of L1 in the intelligibility of L2 speech. However, the results have been somewhat contradictory, leaving unanswered questions about how much the specific language backgrounds of learners may impact their difficulty in understanding speech from a variety of L2 interlocutors (see Study 2 for a more in-depth review of this literature).

All of this leaves language instructors in a somewhat difficult place when trying to decide what to teach, and how to judge whether their own impressions of their learners' intelligibility would compare to an interlocutor from a different L1 background. Study 1, Study 2, and Study 3 all address this issue, albeit in different ways. Study 1 investigates which speech features L1 and L2 listeners attend to when judging speech samples only in terms of global similarities and differences between them; this research aims to target which features of the language are the most salient to listeners, making those features more likely to be noticed by learners, and thus have the potential to be acquired with explicit instruction. Study 2 focuses on specific L2 backgrounds and also investigates the role of language background using the construct of comprehensibility in targeting specific pronunciation features that are commonly taught in the classroom. Study 3 attempts to find a solution for helping learners focus on, and acquire, different aspects of pronunciation depending on their own difficulties and their communicative purposes.

Theoretical Frameworks

One the biggest challenges pronunciation researches face is the lack of theory to guide research. The theories, models, and frameworks that have been employed in pronunciation research to date are generally too limited in scope to have the potential to take a central role as a theory of L2 pronunciation learning. For example, a large number of studies have been carried out using models such as the Perceptual Assimilation Model (Best, 1995), the Speech Learning Model (Flege, 1995), and the Ontogeny-Phylogeny Model (Major, 1986). However, the scope of these models is narrow; they really only explain how a learner's L1 influences speech acquisition in an L2, making their overall utility to the learning of pronunciation limited. Recently, Fraser (2006, 2007, 2010) has attempted to move beyond such models, arguing that they serve "more to explain why learners can't learn pronunciation than to offer them practical help with errors" (2010, p. 360). She advocates a theoretical framework inspired by Cognitive Grammar (for an overview, see Langacker, 2008). At the heart of this framework is the idea that language learners do not transfer different rules and/or sounds from their L1s, but rather entertain different ways of conceptualizing pronunciation. Instruction should, therefore, be aimed at helping learners develop new concepts. This framework is much more useful for pedagogy, as it provides a framework for developing ways to help learners conceptualize a new phonological system, but largely remains limited to helping learners with speech perception and has yet to become a dominant theory. The Willingness to Communicate Framework (WTC) (MacIntyre, 2007; MacIntyre, Dörnyei, Clément, & Noels, 1998), though not intended specifically for pronunciation, has been used to explain variable outcomes in L2 pronunciation development that are related to language use rather than L1 interference (e.g., Dewing, Munro, Thomson, 2008). WTC provides a comprehensive explanation for why an individual chooses to engage in a communicative act with an interlocutor. While WTC in an L2 is certainly an important part of L2 pronunciation development, WTC is not a learning theory, and is ultimately limited in how much it can account for in L2 pronunciation development.

For these reasons, much of the published research on pronunciation is not based on a clear SLA framework or theory. This is the case for Study 2 and Study 3 of this dissertation. Both of these studies are conducted within the *intelligibilityprinciple*, discussed above, but otherwise, while these studies contribute to our knowledge of speech perception and pronunciation learning, they do not directly contribute to theoretical knowledge within a specific theory or framework of SLA. However, in some cases, broader theories of SLA can be used successfully in pronunciation studies to interpret research findings. The motivation for Study 1 lies within the principles of the Interaction Hypothesis. At the core of the interaction approach is the idea that input and interaction with interlocutors can give L2 learners the ability to *notice the gap* between their interlanguage and the language of their speaking partner (Gass & Mackey, 2006; Schmidt & Frota, 1986). This noticing usually occurs through negotiation of meaning and often happens due to feedback and within language related episodes (LREs) (Gass & Mackey, 2006). Gass and Mackey (2007) define LREs as instances when learners talk about the language they are using in an explicit way. For example, learners may discuss which verb tense they should use or where the stress should be placed on a particular word. Study 1 does not investigate interaction between learners, but rather embraces the principle that in order for a feature to be noticed and, therefore, potentially acquired, a learner needs to attend to it in some way. The study investigates what learners are attending to when making judgments about speech samples, and based on the interaction approach, those features that are able to be noticed, may require less explicit attention from an instructor because learners will have the opportunity to notice and attend to these features through interaction with other L2 speakers. This is not to say that noticing guarantees that a feature will be acquired, rather that noticing may assist in the acquisition of a feature.

Pronunciation Teaching Practices

It may seem odd to discuss pronunciation classroom practices in this introduction. While Study 3 is a pronunciation intervention study, it is not situated in a classroom, with learners completing shadowing activities in their own time. However, I felt this issue should be discussed because the lack of authentic communication and the drill-like nature of the activities used in Study 3 might run counter to what would be considered best practices in the broader scope of L2 teaching. Generally, the ideal language classrooms of today involve authentic communication, often initiated by way of a language task. There is some focus on form, but it is secondary to the focus on meaning (Ellis, 2001, 2005). It could be argued that pronunciation instruction is often the exact opposite. There is a strong focus on form, often by way of repetition, and while there is some attention to authentic communication, it is secondary to form (e.g., Gilbert, 2005; Grant, 2010; Hewings, 2004). There have been attempts to make pronunciation instruction more communicative in nature (e.g., Isaacs, 2009) but at the same time, it is often very difficult to address pronunciation adequately in a communicative classroom (Levis & Grant, 2003). While there are good reasons to include authentic communication in pronunciation instruction, it is unlikely that the form-focused orientation of pronunciation instruction will shift dramatically. Studies of instructional interventions that have been able to demonstrate improvements in pronunciation have generally used form-focused instruction of some sort (e.g., Couper, 2003, 2006; Derwing, Munro, & Wiebe, 1997; Saito & Lyster, 2012). Pronunciation is taught, and subsequently learned, differently than other skills and vice versa.

Tying it Together

While this dissertation is comprised of three studies and each has its own specific goals and objectives, there are common threads running through the studies which, when taken in their entirety, bring a deeper understanding to how L2 learners perceive English speech, and how to help learners' improve their perception of L2 speech, regardless of their specific goals and difficulties, thus fostering speech that is more intelligible. Study 1 investigates which aspects of speech underlie judgments of L2 speech. This study compares a group of L2 listeners from a variety of language backgrounds to a group of L1 listeners. This helps us to better understand which aspects of L2 speech listeners attend to when making speech judgments, suggesting which features are more or less likely to be attended to in interaction, and to see whether L1 and L2 listeners judge speech in similar ways. Study 2 investigates whether language background impacts judgments of the comprehensibility of L2 speech. In order to address the complexities of interaction in the context of *Global Englishes*, this study uses language users from very different language backgrounds and from different contexts in Krachu's World Englishes model, namely, listeners from Inner, Outer, and Expanding Circle countries and speakers from Outer and Expanding Circle countries. Study 3 offers an attempt to find a partial solution to the issues raised in Study 1 and Study 2. When approaching language learners as a heterogeneous group of people with different needs and different communication goals, it can be difficult to know how to advise learners and how best to instruct them. The instructional activity that is tested in Study 3 allows for learners to focus on different aspects of pronunciation, thus maximizing the value of instruction for learners with different learning needs and goals.

Introduction to Study 1

The first study in this dissertation investigated which aspects of speech are most salient to L2 listeners when rating L2 speech and examined what underlies those ratings. This is the only study in this dissertation that does not have a focus on comprehensibility, as the intent with this manuscript was to see what listeners attend to when not being asked to focus on a specific construct. Study 1 targeted the construct of "saliency," defined broadly, because by uncovering which aspects of speech are perceptible to listeners, and perhaps more importantly, which are not, we can better understand which dimensions of speech may require more explicit instruction and which are more likely to be acquired incidentally.

One of the goals of this manuscript was to investigate speech perception in a way that is unique methodologically, addressing the issue of L2 speech perception in a new way, namely, through the use of multidimensional scaling. This study and Study 2 both use listener ratings to make judgments about speech, but try to approach this issue from different angles in order to achieve a better overall understanding of L2 perception of non-native speech.

Chapter 2: Study 1

Do you hear what I hear? A multidimensional scaling study of native and non-native listeners' perception of second language speech

Submitted to Journal of Perceptual and Motor Skills

By Jennifer A. Foote & Pavel Trofimovich

Abstract

Second language (L2) speech learning is predicated on learners' ability to notice differences between their own language output and that of their interlocutors. Because many learners interact primarily with other L2 users, it is crucial to understand which dimensions underlie the perception of L2 speech by learners, compared to native speakers. For this study, 15 L2 learners and 10 native English speakers rated 30 L2 audio-recordings from controlled reading and interview tasks for dissimilarity, using all pairwise combinations of recordings. PROXSCAL multidimensional scaling analyses revealed fluency and aspects of speakers' pronunciation as components underlying listener judgments but showed little agreement across listeners. Results contribute to our understanding of why L2 speech learning is difficult and provide implications for language training.

Introduction

Second language (L2) speech development is predicated on the idea that learners notice differences between their own speech and that of their interlocutors. Simply put, learners need to attend to language features in some way for input to be processed and lead to development (Schmidt, 2001). Language teachers and researchers are, therefore, often looking for ways to help learners notice the "gap" between their own output and the language that they are exposed to. Traditionally, the idea of noticing the gap has referred to differences between the speech produced by learners and the target-language speech, as spoken by native speakers. However, the reality for most learners is that the majority of their interlocutors are other L2 users. It is important then to consider not only how learners hear their own speech in relation to native-speaker models, but also how learners compare their speech to the speech of other non-native speakers. Therefore, this study used multidimensional scaling to understand which dimensions underlie the perception of L2 speech by learners as compared to native speakers.

Background

One framework which can accommodate the development of L2 speech is the Interaction Hypothesis (Long, 1996). Underlying this view is the idea that language learning takes place when linguistic issues cause communication to break down during interaction involving L2 speakers. Interlocutors attempt to repair communication by "negotiating for meaning," using such behaviors as clarification requests, confirmation checks, and feedback to resolve misunderstanding. Negotiation for meaning is believed to promote L2 development through opportunities for learners to hear and produce language and to receive feedback on their (non-target) production. In particular, negotiation for meaning ultimately leads speakers to notice the discrepancy (the gap) between the target linguistic system and their own conception of it (Schmidt, 2001), which in turn facilitates language development (Mackey & Goo, 2007). Thus, the idea of noticing similarities or differences between speakers' own linguistic performance and their interlocutors' language is a key component of interaction-driven learning.

If learners interact with native-speaking interlocutors, learners would have at least some opportunities to notice how their speech differs from native-speaker output. Some perceived differences might include various linguistic dimensions, such as individual segments or specific aspects of prosody, dysfluencies, poor word choice, and grammar errors, all of which have been linked to listener perception of L2 speech (Derwing, Rossiter, Munro, & Thomson, 2004; Isaacs & Trofimovich, 2012; Kang, Rubin, & Pickering, 2010). Thus, the experience of interacting with native-speaking interlocutors might gradually help learners align their speech – through sustained input and output practice – with the speech of their interlocutors (e.g., Derwing & Munro, 2013; Saito & Lyster, 2012). However, what if learners are not exposed to native-speaker language, but instead interact with other L2 users, largely in the absence of the types of feedback, interactional modifications, and focus on language typical of language classrooms? Do learners notice how their speech differs from the production of other L2 users and, if they do, which speech dimensions underlie these perceived differences?

There are a number of studies investigating the role of speech variables in listener judgments of speech; however, these studies have primarily focused on native rather than L2 listeners. For instance, numerous linguistic dimensions, including individual segments (e.g., Munro & Derwing, 2006), aspects of prosody and fluency (e.g., Kang et al., 2010), speech rate (e.g., Munro & Derwing, 2001), and various combinations of pronunciation, lexis, grammar, and discourse variables (e.g., Anderson-Hsieh, Johnson, & Koehler, 1992; Crowther, Trofimovich, Saito, & Isaacs, 2014; Isaacs & Trofimovich, 2012; Munro & Derwing, 2001; Saito, Trofimovich, & Isaacs, 2015), have been shown to factor into native-speaking listeners' perceptions of L2 speech.

While there is less research focusing on L2 listeners, interest in this area of research is increasing. For example, there have been studies investigating L2 perceptions of speech rate and fluency (e.g., Derwing & Munro, 2001; Kormos & Dénes, 2004; Rossiter, 2009), phonetic parameters (e.g., Riney, Takagi, & Inutsuka, 2005), and the role of segmentals versus suprasegmentals in judgments of speech (e.g., Kashigawa & Snyder, 2010; Winters & O'Brien, 2013). For instance, Field (2005) investigated the role of lexical stress in intelligibility, both for native and L2 listeners. Both listener groups, who evaluated words with correct or misplaced stress, found stress to be important for word intelligibility, with both groups performing in essentially similar ways. There have also been studies investigating which aspects of speech L2 listeners believe influence their judgments of L2 speech (e.g., Jun & Li, 2010; Wilkerson, 2013). However, there is still much that is unknown about which aspects of L2 speech are most salient to L2 listeners, especially when listeners are not directed (through research design or explicit instructions) to respond to or reflect on particular dimensions of speech.

The Current Study

To summarize, there is a need for more research investigating which linguistics variables are likely to be attended to when listening to L2 speech. As argued previously, a focus on L2 listener perception of non-native speech is crucial because, in the majority of contexts, L2 users tend to interact with other L2 users, with the consequence that non-native speech often represents the *only* input that learners receive. To address this issue, 15 non-native university-level students as well as a comparison group of 10 native speakers of English were asked to listen to short excerpts of L2 speech recorded as part of two tasks (reading, interview), with each excerpt presented for comparison against all other excerpts. The listeners rated how dissimilar each pair of speakers sounded, and these judgments were subsequently analyzed using multidimensional scaling, a procedure which uses similarity or dissimilarity responses to plot stimuli (L2 speakers, in this case) in an n-dimensional space. To interpret the dimensions underlying listener judgments, the dimensional coordinates for each speaker were compared against background characteristics of the speakers as well as several coded measures of pronunciation, fluency, lexis, and grammar, based on each speaker's excerpt. The research question asked, "Which dimensions underlie L2 listeners' perception of non-native speech in controlled reading and extemporaneous interview tasks, in the absence of directions for listeners to attend to any specific speech elements?"

Method

Participants

The non-native participants were 15 L2 speakers of English, with a mean age of 25 years (19.9-30.0), recruited from an English-medium university in Montreal (Canada). The speakers were enrolled in various undergraduate (3) and graduate (12) degree programs and represented a range of backgrounds, including Farsi (5), Telugu, Chinese, French (2 each), Akan/Twi, Arabic, Bengali, and Kinyarwanda (1 each). The speakers had studied English for a mean of 12.4 years (2-19) and had resided in Canada for a mean of .7 years (.2-2.5). All speakers were males, to ensure that gender did not factor into listener judgment. The speakers, recruited during the first semester of their studies, had reported recent TOEFL iBT standardized English proficiency scores, with a mean of 89.3 (79-104) and individual subscores of 21.1 (17-26) for speaking, 21.5

(17-25) for writing, 22.8 (16-30) for reading, and 24.0 (20-30) for listening. For native-speaker baseline, 10 native English listeners, with a mean age of 25 years (20-36) were recruited from the same university. These participants (4 females, 6 males) were born and raised in English-speaking homes and were exposed to English from birth, with one (6) or both (4) parents being native English speakers. Because the native listeners resided in Montreal (a multicultural, bilingual French-English city with a large population of immigrants), they were familiar with accented English as spoken by speakers from diverse L1 backgrounds.

Materials

The materials included 30 audio samples recorded by the L2 speakers as part of two speaking tasks, which differed in degree of formality (controlled reading vs. spontaneous speaking in response to an interview question). The task variable was manipulated because L2 speakers differ in accuracy and fluency of L2 output by task, such that read-aloud tasks often elicit more accurate production of L2 segments and prosody than more spontaneous tasks, such as storytelling and interviews (Rau, Chang, & Tarone, 2009). All recordings were made in a quiet location, with the order of tasks counterbalanced across speakers. The reading task, based on a short paragraph from an ESL textbook (Grant, 2001), elicited speech samples that were identical and therefore maximally comparable in content. After removing initial hesitations and dysfluencies, the first two sentences (*Have you noticed that some people interrupt conversations more than other people? All cultures do not have the same rules governing these areas of communication*) were extracted from each recording and saved as separate files. The resulting audio samples (25 words), which were on average 9.8 s in duration (8.3-12.2), were used as target audio samples from the reading task. The free-response task was based on an interview question from the IELTS English proficiency test, *Describe a job you would like to do in the future* (Jakeman & McDowell, 2008). Unlike the reading task, the interview task elicited spontaneous speech, allowing speakers to have control over their linguistic output while keeping thematic content (future employment) constant. After removing initial dysfluencies, the first few complete ideas from each speaker's response (15-36 words), with a mean duration of 9.6 s (8.3-12.2), were excised from the recordings and used as target samples from the interview task.

Similarity Rating

Roughly three months after recording the audio samples, the same 15 L2 speakers returned to participate in individual listening sessions (about 2 hours in total) to evaluate the recordings, which were followed by similar sessions completed by the 10 native English listeners. Listeners were informed that their task was to help researchers evaluate audio recordings by judging how similar or dissimilar each pair of recordings sounded to them. They received no guidance as to how they should judge the recordings, nor what they should attend to. The rating of '1' was reserved for audio recordings that sounded very similar, while the rating of '9' designated very dissimilar recordings. Listeners then performed a brief practice task, judging the similarity of three recordings (containing the same sentence, *Knowing when to take turns in a conversation in another language can sometimes cause difficulty*) spoken by three additional L2 speakers, with each pairwise combination of the recordings presented in a unique randomized order. Listeners then proceeded to evaluate the target audio samples, organized in two separate blocks by task (reading, interview), with the order of tasks counterbalanced across listeners. At the beginning of each block, listeners were reminded about the directionality and the endpoints of the scale and were encouraged to use its entire range. For the reading task, they were told that each recording featured the same two sentences, which were then shown to them. For the interview task, they were informed that the content of each recording was different but that all speakers described their future job.

Within each block, the 15 audio samples from the reading task and the 15 samples from the interview task (22,5 kHz, 16-bit resolution) were presented to each listener in all possible pairwise combinations (for a total of 105 pairs per block). The experiment was controlled by E-Prime (Schneider, Eschman, & Zuccolotto, 2002), and each listener used a headset and clearly labeled 1-9 keys on a computer keyboard to record their judgment. Each trial started with a warning which read "Next pair…" and stayed on the screen for 1.5 s, followed by two audio samples played in sequence with a .25 s interval. All pairs were presented in a unique random order, which included random designation of each recording as the first or second in each pair, and each trial terminated when a response was logged, which initiated a subsequent trial. All listeners took a short break after completing the first block of comparisons.

Speech Analysis

To relate psychological dimensions underlying similarity judgments to specific properties of speech, the 30 target recordings from the reading and interview tasks were analyzed for several pronunciation, fluency, lexis, and grammar variables:

(1) segmental errors: number of phonemic substitutions (e.g., *the* spoken as *da*);

(2) syllable structure errors: number of phonemic insertion/deletion errors (e.g., *would* without the final /d/), with both error counts divided by the total number of words;

(3) word stress errors: number of misplaced or missing stresses in polysyllabic words (e.g., *com-PU-ter* spoken as *COM-pu-ter*) over the total number of polysyllabic words;

(4) total sample duration (in seconds) as a coarse measure of fluency;

(5) unfilled pauses: total number of silent pauses lasting longer than .4 s (e.g., *I think in the future I will still I will still be* [unfilled pause] *a software engineer*);

(6) filled pauses: total number of nonlexical pauses such as uh and um (e.g., *In the future I'd like to work in uh* [one filled pause] *corporate finance*), with both filled and unfilled pause measures normalized by dividing pause frequency by the total duration of the recording (yielding pause frequency per second of speaking time);

(7) speech rate: total number of syllables produced (including pauses and dysfluencies) over the total duration of the recording (syllables per second);

(8) repetitions/self-corrections: all immediately repeated and self-corrected words (e.g., *I I* [repeated] *worked in China for for* [repeated] *about seven years as a softwa ware* [corrected] *engineer*);

(9) grammar errors: number of words with at least one error in sentence structure, morphology, or syntax (e.g., *The first time I touched the computer is in my primary school* spoken with a definite article before *computer* and the wrong tense of the verb *to be*) divided by the total number of words;

(10) lexical errors: number of incorrectly used or inappropriate lexical expressions (e.g., *desired job* instead of *dream job*) over the total number of words;

(11) token frequency or the total number of words spoken;

(12) type frequency or the total number of unique words produced, with both token and type frequency corrected for differences in sample length by dividing the raw counts by the total sample duration (yielding token and type rates per second of speaking time).

All measures were first coded by a trained coder, then recoded by another trained coder. Although all coding decisions involve a certain degree of subjectivity and may not reflect the variability found in spoken language, only 33 (9%) of all coded data cells involved disagreement, which was resolved through discussion. Table 1 summarizes the 12 coded linguistic variables from the L2 speakers' recordings in the reading and interview tasks.

Table 1

Summary of Coded Linguistic Variables in L2 Speakers' Speech in Reading and Interview Tasks

Variable	Reading task		Interview task	
	M (SD)	*Range*	*M (SD)*	*Range*
Segmental errors	.20 (.13)	.08-.48	.11 (.08)	.00-.29
Syllable structure errors	.07 (.06)	.00-.23	.06 (.06)	.00-.23
Word stress errors	.16 (.13)	.00-.40	.20 (.23)	.00-.23
Sample duration	9.78 (1.30)	8.33-12.17	9.61 (2.36)	6.52-14.48
Unfilled pauses	.11 (.04)	.00-.19	.20 (.13)	.00-.40
Filled pauses	.01 (.03)	.00-.10	.21 (.20)	.00-.72
Speech rate	4.30 (.55)	3.32-4.98	3.49 (.87)	2.07-5.53
Repetitions/self-corrections	.02 (.02)	.00-.07	.05 (.07)	.00-.20
Grammar errors	.01 (.01)	.00-.04	.05 (.01)	.00-.22
Lexical errors	.00 (.00)	.00-.00	.02 (.03)	.00-.09
Token frequency	2.64 (.30)	2.06-3.00	2.54 (.51)	1.64-3.53
Type frequency	2.42 (.29)	1.89-2.76	1.95 (.35)	1.35-2.66

Analysis

The data from the similarity rating task were analyzed using multidimensional scaling (MDS), an exploratory procedure which uses similarity or dissimilarity matrices to generate a representation of stimuli in geometric (Euclidian) space, with each stimulus (e.g., an item or a person) plotted as a point and inter-stimulus distances showing similarity or "psychological distances" between them. The spatial map yielded by MDS represents a visual depiction of underlying dimensions governing a stimulus set, and a researcher's challenge is to identify and interpret these dimensions (Borg & Groenen, 1997). In this study, all MDS outputs were generated through SPSS 21.0 using the PROXSCAL algorithm (Busing, Commandeur, & Heiser, 1997), with 100 random iterations, and all similarity data (based on 9-point Likert scales) treated as ordinal. Because L2 listeners' ratings of their own speech, relative to the speech of their peers, may have impacted their judgments (e.g., with own speech rated more favourably, compared to the speech of others), the final similarity matrices for L2 listeners excluded the 14 datapoints involving listeners' own speech. The final similarity matrices were thus based on a total of 91 pairwise comparisons for each L2 listener and 105 pairwise comparisons for each native listener.

Results

Reading Task

Scree plots, which depict stress (a measure of goodness of fit between estimated inter-stimulus distances and the original listener-based similarity matrices), were inspected first to determine the optimal dimensionality of MDS outputs for L2 listener and native listener data in the reading task. For both outputs, a two-dimensional solution was chosen because adding subsequent dimensions failed to substantially increase fit. The final two-dimensional models featured low stress functions ($S_{\text{L2 listener}} = .10$; $S_{\text{native listener}} = .10$), which were considered excellent (Jaworska & Chupetlovska-Anastasova, 2009), and high dispersion indexes ($\text{DAF}_{\text{L2 listener}} = .97$; $\text{DAF}_{\text{native listener}} = .96$), which exceeded the minimum acceptable value of .60 (Meyer, Heath, Eaves, & Chakravarti, 2005), with each model accounting for over 96% of the variance in the input data. Therefore, both MDS outputs were plotted in two-dimensional Euclidian space, using MDS dimensional coordinates, with the first dimension plotted along the *x*-axis and the second dimension along the *y*-axis (Figure 1).

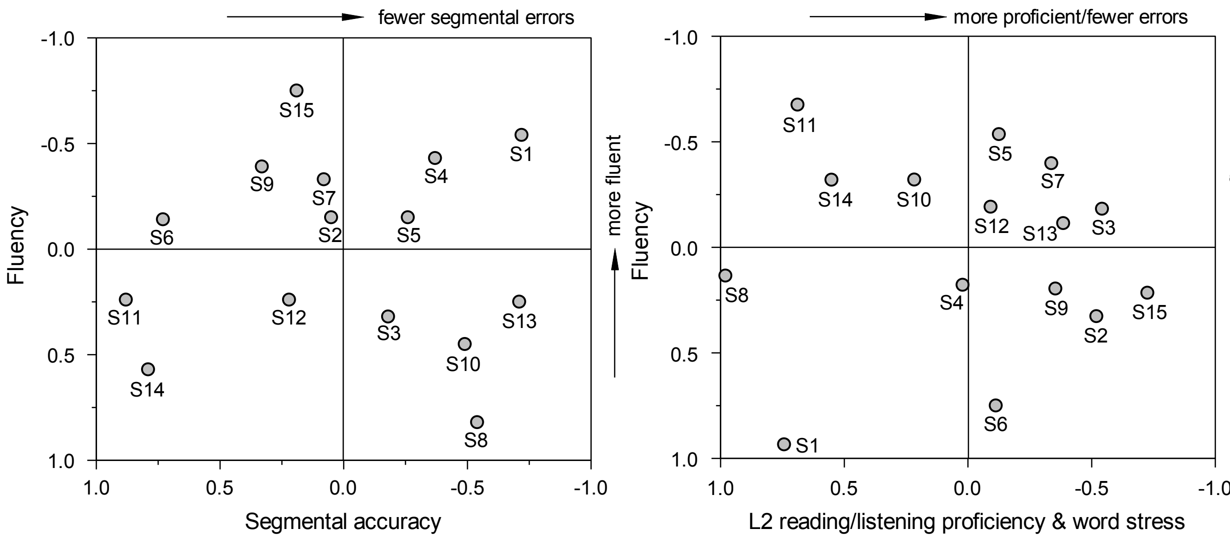


Figure 1. Two-dimensional plots representing MDS outputs for L2 listeners (left) and native listeners (right) in the reading task. The plots depict two dimensions best explaining listeners' dissimilarity ratings for 15 L2 speakers (S1-S15), with each speaker compared against all other speakers.

To interpret MDS output dimensions for both the L2 listener and native listener outputs, two-tailed Spearman correlations were computed between the dimensional coordinates from each MDS output and all relevant L2 speaker background characteristics (e.g., L1 group, TOEFL score) and speech variables (e.g., speech rate, lexical errors). The results of these analyses are summarized in Table 2. For L2 listeners, Dimension 1 could best be interpreted in terms of a combination of speakers' L1 background and segmental errors in their speech, reflecting a common observation that segmental errors are specific to speaker background. Dimension 2 was associated with speech rate and with token and type production ratios, expressed as the number of word tokens and types uttered per second of speaking time. Based on these data, the dimensions underlying L2 listener perception of non-native speech in the reading task could be labeled as SEGMENTALS (Dimension 1) and FLUENCY (Dimension 2). For native listeners, Dimension 1 strongly patterned with the speakers' overall TOEFL iBT performance, especially reading and listening subscores, which likely reflected orthography-mediated links between reading and listening, as well as with speakers' word stress errors. Dimension 2 was uniquely linked to rate of unfilled pausing. Thus, the dimensions underlying native listener perception of L2 speech in the reading task could be labeled as L2 READING/LISTENING PROFICIENCY & WORD

STRESS (Dimension 1) and FLUENCY (Dimension 2).

Table 2

Spearman Correlation Coefficients (Two-Tailed) Between MDS Dimensional Coordinates and Various Speaker Background and Speech Characteristics from the Reading Task

Variable	L2 listeners		Native listeners	
	Dimension 1	Dimension 2	Dimension 1	Dimension 2
L1 background	.78***	−.08	−.02	−.36
TOEFL score	−.14	−.15	−.86***	−.31
TOEFL reading subscore	.46	−.19	−.63*	−.23
TOEFL listening subscore	−.02	−.17	−.62*	.20
TOEFL speaking subscore	−.40	.31	−.11	−.51
TOEFL writing subscore	.20	−.11	−.06	−.30
Segmental errors	−.66**	.08	−.15	−.23
Syllable structure errors	.18	.17	.29	−.31
Word stress errors	.36	−.38	−.55*	.30
Sample duration	.47	.56	.03	−.19
Unfilled pauses	−.24	−.25	.14	.59*
Filled pauses	.10	.28	.36	−.47
Speech rate	−.45	−.60*	−.09	.23
Repetitions/self-corrections	.02	.28	−.19	−.29
Grammar errors	−.14	.23	−.14	−.09
Token frequency	−.40	−.61*	−.16	.20
Type frequency	.41	−.61*	−.13	.19

Note. *p < .05, **p < .01, ***p < .001. Directionality of associations is uninformative because the MDS solution was rotated to achieve the best L2 speaker clustering in two-dimensional space (see Figure 1).

To quantify possible differences between the L2 listener and native listener MDS outputs from the reading task, inter-speaker distances from the two MDS outputs were correlated using a Spearman correlation test. The assumption was that perceptual distances between each speaker in L2 listeners' two-dimensional space should match closely the corresponding distances in the two-dimensional space generated by native listeners if both groups approached the task in a similar way (Hout, Papesh, & Goldinger, 2013). This analysis yielded a weak correlation, $r(103) = .36$, $p < .0001$, suggesting that only about 13% of variance in inter-speaker distance was common between the MDS outputs for the two listener groups. Put differently, the speaker pairs perceived as being similar by L2 listeners were often not the same pairs perceived as similar by native listeners. In essence, while both MDS outputs included a similar number of dimensions, the two groups approached L2 speech rating in the reading task in different ways.

Interview Task

As in the previous analyses, scree plots were consulted first to determine most optimal solutions for MDS outputs from the interview task. For both outputs, two-dimensional solutions were deemed most appropriate, with excellent stress functions ($S_{\text{L2 listener}} = .12$; $S_{\text{native listener}} = .10$) and high dispersion indexes ($DAF_{\text{L2 listener}} = .95$; $DAF_{\text{native listener}} = .96$) explaining over 95% of the variance in the input data. Therefore, as with the data from the reading task, both MDS outputs from the interview task were plotted in two-dimensional Euclidian space (Figure 2).

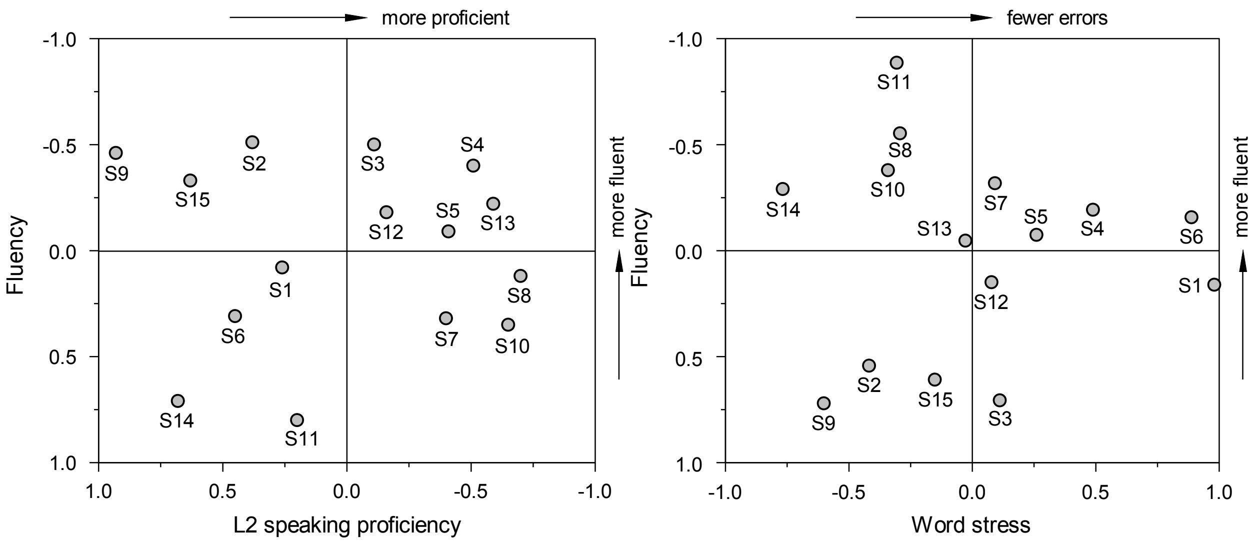


Figure 2. Two-dimensional plots representing MDS outputs for L2 listeners (left) and native listeners (right) in the interview task. The plots depict two dimensions best explaining listeners' dissimilarity ratings for 15 L2 speakers (S1-S15), with each speaker compared against all other speakers.

To interpret MDS output dimensions, similar Spearman correlational analyses were conducted for both the L2 listener and native listener outputs, relating MDS dimensional coordinates to several background and speech variables (see Table 3). For L2 listeners, Dimension 1 patterned singly with the speakers' TOEFL iBT speaking subscore, while Dimension 2 was linked to the total duration of the speech sample, rate of unfilled pausing, and token frequency, expressed as word tokens spoken per second of speaking time. Thus, the MDS dimensions underlying L2 listener perception of non-native speech in the interview task could be labeled as L2 SPEAKING PROFICIENCY (Dimension 1) and FLUENCY (Dimension 2). For native listeners, Dimension 1 was not uniquely associated with any single variable investigated here, but the strongest association involved speakers' word stress errors ($r = -.50$, $p = .06$). Dimension 2 was uniquely linked to the total duration of the speech sample, likely reflecting speakers' verbal fluency within total amount of speaking time. Therefore, the MDS dimensions underlying native listener perception of L2 speech in the interview task could be tentatively referred to as WORD STRESS (Dimension 1) and FLUENCY (Dimension 2).

Table 3

Spearman Correlation Coefficients (Two-Tailed) Between MDS Dimensional Coordinates and Various Speaker Background and Speech Characteristics from the Interview Task

Variable	L2 listener MDS output		Native listener MDS output	
	Dimension 1	Dimension 2	Dimension 1	Dimension 2
L1 background	.47	.38	−.30	−.22
TOEFL score	−.04	−.32	.03	.22
TOEFL reading subscore	.28	−.11	−.21	.13
TOEFL listening subscore	−.17	−.41	−.05	.07
TOEFL speaking subscore	−.67*	.00	.31	−.23
TOEFL writing subscore	.32	.01	.18	.24
Segmental errors	.13	.29	−.09	−.27
Syllable structure errors	.49	−.23	−.28	.43
Word stress errors	.15	.12	−.50	−.07
Sample duration	−.23	.66*	−.14	−.72*
Unfilled pauses	−.17	.55*	.47	−.51
Filled pauses	−.14	−.38	−.18	.36
Speech rate	.36	−.47	−.25	.44
Repetitions/self-corrections	.23	.03	−.08	.23
Grammar errors	.22	−.32	−.16	.41
Lexical errors	−.42	−.04	−.20	−.32
Token frequency	.08	−.54*	−.11	.42
Type frequency	−.13	−.49	−.13	.29

Note. *$p < .05$. Directionality of associations is uninformative because the MDS solution was rotated to achieve the best L2 speaker clustering in two-dimensional space (see Figure 2).

Again, to quantify differences between the two MDS outputs in the interview task, a Spearman correlation was carried out to compare inter-speaker distances from the two outputs. This analysis revealed a moderate correlation, $r(103) = .50$, $p < .0001$, indicating that the two outputs shared about 25% of variance in terms of how closely the same L2 speakers were positioned in the perceptual spaces generated by the two listener groups. Thus, while there was some correspondence in how both sets of listeners approached the task of rating L2 speech in the interview task, the ultimate perceptual spaces generated by the two groups were largely different.

Discussion

The research question asked which dimensions underlie native and L2 listeners' perception of L2 speech in controlled reading and extemporaneous speaking tasks, in the absence of directions for listeners to attend to any specific speech elements. For L2 listeners, segmental accuracy and fluency appeared to underlie listener perception of L2 speech in the reading task, while L2 speaking proficiency and fluency reflected their judgments in the interview task. For native listeners, word stress accuracy, along with L2 reading/listening proficiency, and fluency characterized listener ratings of L2 speech in the reading task, while word stress accuracy and fluency were the two dimensions relevant to listener judgments in the interview task. The most consistent finding was fluency as a common component underlying the perception of L2 speech. Despite these similarities, there was only moderate agreement across the two groups in the interview task and virtually no agreement in the reading task, suggesting that the two listener groups approached the rating of L2 speech in different ways.

Differences Between Listeners

The first of the two dimensions underlying listener perception of L2 speech varied across listeners and tasks. L2 listeners' perception was related to speakers' segmental accuracy in the reading task and to speakers' L2 speaking ability (as measured by TOEFL speaking subscores) in the interview task. The reading task is a formal, controlled speaking activity with identical lexical content across speakers and orthographic support guiding oral production, so it was expected that the speakers would make few grammatical and lexical errors. Therefore, it is unsurprising that segmental substitutions, which are typically specific to speakers' language background, emerged as a dimension underlying L2 listeners' judgments. For instance, segmental errors contribute to both trained and inexperienced raters' judgments of L2 accent (Kennedy & Trofimovich, 2008). Listeners' use of segmental errors to distinguish other L2 speakers from one another is also consistent with a typical instructional emphasis on segmentals in L2 classrooms (Foote, Trofimovich, Collins, & Soler Urzúa, 2013) and with learners' beliefs that segmental errors constitute the greatest challenge to their pronunciation (Derwing, 2003).

Compared to the reading task, the interview task is an extemporaneous speaking activity, which allows speakers more linguistic freedom to express themselves using a vernacular speaking style. In this task, the first of the two dimensions that patterned with L2 listener judgments was speakers' global L2 speaking ability, as measured through TOEFL speaking subscores. The TOEFL speaking section includes six tasks, of which two require test-takers to speak on familiar topics while the remaining four involve either listening to or both reading and listening to relevant information before integrating it into the response. The speaking subscore appears to reflect test-takers' speaking ability, with integrated components contributing minimally to the reading and listening constructs (Sawaki, Stricker, & Oranje, 2009). The speaking subscore also seems to be distinct, compared to other modalities such as listening or writing, providing an added value to the total test score (Sawaki & Sinharay, 2013). It appears, then, that L2 listeners were sensitive to more than segmental errors or features of accent as they tried to distinguish one L2 user from another in the interview task. Yet because only global aspects of L2 proficiency, rather than any of the specific linguistic properties of L2 speech (from among those targeted here), patterned with listener ratings, there is no evidence in these data of which linguistic aspects of speech were relevant to making up the L2 speaking ability construct.

Unlike L2 listeners, native listeners appeared to rely on the dimension of word stress accuracy in speakers' output in both reading and interview tasks, with an additional contribution of speakers' reading/listening proficiency in the reading task. Because reading aloud is an activity with fixed lexical and grammatical expression, it is reasonable that native listeners relied on speakers' reading and listening proficiency (as measured through TOEFL reading and listening subscores) in judging how similar or different L2 speakers sounded. Indeed, reading aloud involves both reading and listening skills, as it requires speakers to code orthography into phonology, then to articulate the prepared speech plan and to perceptually monitor the speech output for accuracy (e.g., Levelt, 1989). However, the finding that word stress was linked to native listeners' judgments of L2 speech in both reading and interview tasks is noteworthy in light of the importance of prosody, which includes such linguistic categories as intonation, stress, and rhythm, for L2 speech learning and teaching. For example, word stress and rhythm (along with other prosodic features) may account for up to 50% of the variance in accent judgments for L2 speakers from varied linguistic backgrounds (Lea). Word stress also contributes to listener perception of comprehensibility for speakers from multiple L1 groups (Crowther et al., 2014) and to intelligibility for both L1 and L2 listeners (Field, 2005). Similarly, speech training focusing on word stress, along with other prosody and fluency characteristics of speech, can lead to measurable gains in L2 learners' comprehensibility in extemporaneous speaking tasks, as compared to an equivalent amount of instruction targeting only individual sounds (Derwing, Munro, & Wiebe, 1998). In fact, the role of word stress in native listeners' perception of speech might be related to stress being one of the most structural and hierarchical aspects of phonology (e.g., in metrical phonology), representing the core element of native speakers' linguistic competence (de la Mora, Nespor, & Toro, 2103).

Similarities Between Listeners

Fluency emerged as the most pervasive characteristic underlying listener judgments of L2 speech, emerging as a dimension across both tasks and both listener groups (cf. Tables 1 and 2). In the reading task, the dimension of fluency encompassed the rate of word type and token production as well as speech rate (for L2 listeners) or pausing frequency (for native listeners). In the interview task, the dimension of fluency involved total sample duration, pausing frequency, and total number of lexical items produced (for L2 listeners) or total sample duration (for native listeners). Although the precise measures of fluency varied across listeners and tasks, these measures share one characteristic, namely, they ainll reflect temporal dimensions of speech output, such as frequency of pausing, duration of speaking, or lexical fluency expressed as total number of words uttered. This implies that fluency plays a prominent role in the perception of non-native speech by both native and L2 listeners. At least one reason for this might be that temporal dimensions of fluency, as perceived by the listener, may mark L2 speakers as non-native language users. For instance, Munro and Derwing (2001) showed that a 10% increase in speaking speed resulted in L2 utterances being rated as less accented by native listeners. Similarly, overall utterance duration is an indicator of how nativelike L2 speakers sound, contributing to the perception of accent (MacKay & Flege, 2004). Compared to native speakers, L2 users often have a different distribution of pauses in their speech, even though the overall number of pauses might be the same, suggesting that measures of pausing can also act as salient markers of nativelikeness (Bosker, Quené, Sanders, & de Jong, 2014). In essence, temporal elements of speech may act as salient cues distinguishing L2 users from one another for both native and L2 listeners. Yet as a comparison of MDS outputs for native and L2 listeners suggests, these temporal fluency cues differed across the two listener groups and were likely weighed by them in different ways, with the consequence that the L2 speakers considered fluent by native listeners were not the same as those judged fluent by L2 listeners (cf. Figures 1 and 2).

Implications

Following the assumption that L2 learners need to notice language features in some way for input to lead to linguistic development (Schmidt, 2001), it is not at all surprising that learning pronunciation is such a complex task, especially for learners in contexts where the majority of language users are L2 speakers. As shown through MDS analyses, both sets of listeners were sensitive to global aspects of L2 speech, such as speaking proficiency (L2 listeners) and reading/listening proficiency (native listeners), as well as to some of specific characteristics of speech, including segmental errors (L2 listeners), word stress (native listeners), and temporal aspects of fluency, such as speech rate and pausing (both listener groups). However, outside the domain of fluency, the specific linguistic variables underlying listener judgments of L2 speech were limited. For instance, word stress accuracy – one aspect of speech prosody – seemed to underlie native speakers' perception in both controlled and extemporaneous tasks. In contrast, L2 listeners appeared to attend only to segmental errors, and only in the controlled reading task. This implies that while L2 speakers might be aware that their speech is different from an interlocutor, they may not have a clear understanding of exactly how and why it is different.

These findings have implications for L2 speech learning, particularly within interactionist approaches to L2 development. As mentioned previously, underlying these views is the idea that specific aspects of interaction – referred to broadly as negotiation for meaning – ultimately lead L2 users to notice the gap between the target language and their own understanding of it, which in turn facilitates language development (Long, 1996; Mackey & Goo, 2007). Thus, if learners are unable to distinguish differences between their own linguistic performance and the language produced by their interlocutors or between the linguistic output of speakers in their communicative environment, negotiation-driven learning may not be as efficient as it could be or may be focused on aspects of speech which may not be as crucial to communicative success as others (e.g., segmentals vs. prosody). The current findings point to this possibility, particularly in contexts where learners are primarily exposed to non-native input.

Needless to say, the results of this study must be interpreted with caution, considering the small sample size targeted and the experimental, lab-based approach employed. The L2 speech analyzed also involved only male speakers from mixed L1 backgrounds. Future studies could therefore look at differences both within and across different language groups, in terms of speakers and listeners. As university-level users of English, the speakers also represented the range of L2 ability considered sufficient for them to pursue academic studies. It would therefore be interesting to see how differences in listener and speaker proficiency could contribute to L2 perception. Further, as listener perception of L2 speech depended on speaking task, future investigations of L2 speech perception should target different task types.

One positive aspect of these findings concerns their instructional implications. Overall, the native and L2 listeners were more similar in the interview task (with 25% shared variance) than in the reading task (with 13% shared variance). It is promising that there is more similarity with the interview task, which more closely mirrors language production in naturalistic settings. However, for the L2 listeners, the more constrained reading task led to more "focused" perception behaviour, as it involved some sensitivity to segmental errors, suggesting that the controlled content of the reading task may have enabled L2 listeners to attend to specific linguistic elements in speech. In contrast, in the interview task, L2 listeners mainly attended to global aspects of L2 ability, not necessarily the specific linguistic features that distinguish speakers from one another. It may therefore help learners to have access to controlled input enabling them to focus attention on linguistic features rather than content.

In fact, it might be highly difficult for L2 learners interacting with other L2 users to notice and acquire new features of pronunciation incidentally. For instance, in a survey of 100 learners of English in Canada, Derwing (2003) found that many were unable to describe their own pronunciation difficulties, and for those that could, small sets of segmental targets were most commonly cited, notably, those that were unlikely to cause communication problems (such as English 'th'). Arguably, the task of figuring out pronunciation difficulties from input alone will be even more challenging in contexts where most language users are L2 speakers. For this reason, explicit instruction, which can draw learners' attention to features that they do not notice through exposure alone, will be important for helping them improve their pronunciation. Notwithstanding the benefits of genuine communication for language learning, explicit pronunciation instruction is needed to help learners develop their pronunciation.

Connecting Study 1 to Study 2

Study 1 showed that L1 and L2 listeners judged non-native speech differently and that there were differences in which speech variables explained the MDS solutions for L1 and L2 listeners. However, Study 1 did not address whether there may be further differences between listeners based on their language backgrounds; all of the L2 listeners in this study were treated as a single group. Further, Study 1 examined which aspects of speech are perceptible to listeners, but not which contribute the most to comprehensibility. Based on the intelligibility principle discussed in the introduction of this dissertation, the focus of pronunciation instruction should ideally be on making learners easier to understand. However, in order to help learners become more comprehensible, it is first necessary to know which aspects of speech are most relevant to comprehensibility. Study 2 thus addressed issues of comprehensibility, focusing on the language background of both the listeners and the speakers.

In addition to investigating which speech variables contribute to comprehensibility for different listener-speaker combinations, Study 2 examined whether the language background of the listener impacts intelligibility beyond what can be explained by the pronunciation, prosody, and fluency characteristics of L2 speech and whether listeners themselves consider language background to be an important factor when making comprehensibility judgments. This further distinguishes Study 2 from Study 1, which did not ask listeners what they consciously attended to when making speech judgments.

Chapter 3: Study 2

Is it because of my language background? A study of language background influence on comprehensibility judgments

To be submitted to TESOL Quarterly

By Jennifer A. Foote

Abstract

Recent years have seen an increase in research investigating the role of first language (L1) background on how second language (L2) speech is perceived. However, there are still unanswered questions about the contribution of different speech variables to the comprehensibility of speech for listeners from different language backgrounds. It is also unclear whether L1 background impacts judgments of comprehensibility beyond what can be explained by linguistic variables. Further, little is known about whether L2 listeners consider language background an important factor when making comprehensibility judgments. In the current study, English speakers from Mandarin, French, Hindi, and English language backgrounds (10 per group) listened to speech samples from 30 L2 speakers of English from Mandarin, French, and Hindi backgrounds (10 per group). The listeners rated the speech samples for comprehensibility and provided verbal reports about each sample indicating their reasons for the ratings. After completing the initial ratings, the participants rated the speech samples again for four speech measures (segmental and word stress errors, intonation, speech rate). Correlations of the speech measures and comprehensibility ratings for each L2 listener-speaker group revealed that there were differences in which speech variables were associated with comprehensibility ratings depending on the L1 backgrounds of the listeners and speakers. Hierarchical regressions carried out for each L2 listener group revealed that language background accounted for an additional six percent of the variance in comprehensibility rating for the Mandarin listeners but did not significantly contribute to explaining comprehensibility judgments for the other L2 listener groups. Verbal reports, coded for any comments indicating that the listeners considered L1 to be a factor in their ratings, showed that French and Mandarin listeners were the most likely to mention L1 when making judgments about comprehensibility and that listeners more often

considered L1 a benefit when rating speech samples from their own L1 and a detriment when rating speech samples from a different L1.

Introduction

The past 20 years have seen an increased interest in comprehensibility both from pedagogical and research perspectives. In pronunciation instruction, there has been a gradual move away from attempting to help second language (L2) learners sound as much like a native speaker of the target language as possible, towards a more realistic goal of helping learners become easier to understand (Derwing & Munro, 2005; Levis, 2005). However, while comprehensibility is an appropriate and realistic goal for L2 learning, it is also complex. There are many variables that contribute to how well one person understands the speech of another. At different times, L2 speakers of a given language may be using their L2 to communicate with native (L1) speakers of the target language or with other L2 speakers, which reflects a lingua franca approach to pronunciation learning and use, with proponents creating guidelines and materials for teaching pronunciation with L2-L2 interaction in mind (e.g., Jenkins, 2000; Walker, 2010). While some research has explored the role of different L2 backgrounds in the perception of L2 speech, much remains unknown. For instance, little is known about the role that L1 background may play in the relative contributions of specific linguistic dimensions of speech, such as segmentals or suprasegmentals, to comprehensibility judgments. Further, there are still unanswered questions about how the L1 backgrounds of both listeners and speakers impact how difficult listeners judge speech to be and the extent to which listeners attribute ease or difficulty in understanding to the language background of the speaker.

Understanding L2 Comprehensibility

When talking about comprehensibility, it is important to distinguish it from intelligibility. While the terms are used synonymously in vernacular speech, and even in L2 acquisition research broadly to refer to how understandable speech is, they have distinct meanings. Comprehensibility denotes subjective judgments about how easy or difficult speech is to understand, while intelligibility measures actual understanding (Munro & Derwing, 1995a). Measurements of comprehensibility are typically made by having listeners judge samples using rating scales (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995a; Trofimovich & Isaacs, 2012). In contrast, measurements of intelligibility often involve participants transcribing utterances to see if speech is actually understood (e.g., Bent & Bradlow, 2003; Derwing & Munro, 1997; Munro & Derwing, 1995a; Xie & Fowler, 2013). At least for L1 listeners, these constructs can be closely related; however, they are not perfectly correlated, as utterances that are fully intelligible may not be judged as perfectly comprehensible (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995a). When considering the role of understanding in successful communication between interlocutors, then, comprehensibility may be the more important measure, as listeners can get frustrated talking with someone who requires a lot of effort to understand, even if those efforts are ultimately successful. And because intelligibility measures often involve the decontextualized identification of words, it is possible that a listener could identify all of the words used, but have difficulty processing a speaker's actual message.

Role of Language Background in L2 Comprehensibility

There is a general belief that L1 background, particularly when L2 users share an L1, impacts understanding of L2 speech. Studies investigating whether test takers are advantaged when sharing the language background of the speaker in listening tests have revealed mixed results, with some language groups showing an advantage, but with little overall effect in some cases and even a disadvantage in other cases (e.g., Harding, 2012; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Smith & Bisazza, 1983; Tauroza & Luk, 1997). Research also suggests that trained test-raters may rate speakers from their own language background more leniently (e.g., Winke & Gass, 2013; Winke, Gass, & Myford, 2012). Outside of language testing, most research investigating the role of L1 background in the understanding of L2 speech has employed intelligibility measures, focusing on what Bent and Bradlow (2003) coined as the interlanguage speech intelligibility benefit (ISIB). Hayes-Harb, Smith, Bent, and Bradlow (2008) further break this concept down to contrast an ISIB for listeners with an ISIB for talkers. The former refers to an advantage for L2 listeners over L1 listeners when listening to L2 speech, and the latter refers to a benefit for L2 speakers when an utterance is from an L2 speaker rather than an L1 speaker.

There is a wealth of evidence that L1 speakers generally find L2 speech more difficult to understand, compared to L1 speech (e.g., Bent & Bradlow, 2003; Hayes-Harb et al., 2008; Smith & Hayes-Harb, 2009; van Wijngaarden, 2001). However, for L2 speakers listening to L1 and L2 speech, the picture is complex. Several studies have found evidence of an ISIB of some type (e.g., Bent & Bradlow, 2003; Hayes-Harb et al., 2008; Imai, Wally, & Flege, 2005; Munro, Derwing, & Morton; 2006; Smith, Bradlow, & Bent, 2003; van Wijngaarden, 2001; Xie & Fowler, 2013). However, other studies have revealed weak or negative results (e.g., Algethami, Ingram, & Nguyen, 2011; Stibbard & Lee, 2006), and several investigations have shown mixed results, often with proficiency of the listeners and speakers impacting the findings. For example, Hayes-Harb et al. (2008) reported an ISIB for listeners but not speakers, and the effect for listeners only held for low proficiency listeners and speakers. ISIB studies are difficult to summarize due to the variety of findings and approaches; however, it is fair to say that while there can be an ISIB for L2 users, it is uncertain how strong the effect is, and it depends on many factors, such as proficiency, context, and learner background characteristics (Smith & Hayes-Harb, 2011; Xie & Fowler, 2013).

The majority of the studies investigating the role of L1 background in understanding L2 speech have been primarily interested in intelligibility rather than comprehensibility judgments, and most used narrow measures of intelligibility, such as the identification of individual words or transcribing short sentences. There are good reasons for using narrower measures. Much of the research on shared intelligibility is focused on the idea that language users with a shared L1 have similar phonological representations, and narrow measures allow for greater control over specific acoustic properties of speech. However, for real-world communication, it is important to understand mutual *comprehensibility* of speech for L1 and L2 users, in addition to mutual intelligibility of speech in tightly controlled conditions. Unfortunately, there has been less research investigating L2 listeners' ratings of comprehensibility. Studies that have used L2 listeners' comprehensibility judgments of L2 speech have tended to be limited in their scope. For example, Matsuura, Chiba, Mahoney, and Rilling (2014) compared Japanese listeners' comprehensibility ratings of speech read by an English L1 speaker and a Hindi speaker, while Matsuura (2007) used a similar design with Japanese listeners rating an L1 speaker and a speaker from Hong Kong.

One study that has investigated the role of L1 in the understanding of L2 speech using both comprehensibility judgments and intelligibility measures with a range of L2 backgrounds is Munro et al. (2006). They had listeners with L1 Cantonese, Mandarin, Japanese, or English backgrounds listen to speech samples from Cantonese, Japanese, Polish, and Spanish speakers of

English. In terms of intelligibility, Japanese speakers better understood speech from their own L1 but Cantonese speaker did not. For comprehensibility, results were similar but not identical. While the English listeners found all groups equally easy to understand, there were group differences for other listeners. For example, the Cantonese listeners found the speech from their own language background easier to understand than the speech by Japanese, Polish, and Spanish speakers, and the Japanese listeners found their L1 group's speech easier to understand than the speech by Cantonese but not by the other speaker groups. However, the effect sizes were small and overall there was fairly strong agreement between the listener groups in their relative judgments of speech, leading Munro et al. to conclude that the role of the L1 of the listener was overall less important than linguistic characteristics of the speech itself.

Role of Linguistic Variables in L2 Comprehensibility

Researchers have increasingly been interested in understanding how different linguistic variables in L2 speech contribute to intelligibility and comprehensibility, targeting for the most part L1 listeners as raters. For example, some studies have investigated the impact of specific speech characteristics on comprehensibility and intelligibility. Hahn (2004) found that primary (sentence level) stress impacted intelligibility, and Munro and Derwing (2001) showed that speech rate had a curvilinear effect on judgments of accentedness and comprehensibility. Munro and Derwing (2006) investigated the role of functional load in comprehensibility judgments, reporting that low functional load errors (i.e. involving vowels and consonants which distinguish few words in a language) had a minimal impact on comprehensibility judgments while high functional load errors had a large impact. Investigating how multiple speech characteristics contribute to comprehensibility ratings, Kang, Rubin, and Pickering (2010) had 188 L1 English speakers rate the oral proficiency and comprehensibility of 26 L2 speakers of English from four different L1 groups. The researchers found that prosodic aspects of L2 speech accounted for around 50% of the variance in the comprehensibility ratings. Among the speech dimensions contributing to L2 comprehensibility were word stress, pitch height, high-rising tones, and a cluster variable labelled "suprasegmental fluency" which encompassed measures such as articulation rate and syllables per second.

Several recent studies have moved beyond prosody to investigate a wider range of linguistic variables and their relationship to L2 comprehensibility. Isaacs and Trofimovich

(2012) had 60 L1 English listeners rate comprehensibility of 40 L1 French speakers of English. These ratings were then correlated with 19 speech measures related to pronunciation, fluency, grammar, lexis, and discourse, revealing that comprehensibility was linked to all targeted linguistic categories. Other recent research has looked at how linguistic measures relate to comprehensibility as a function of the speakers' L1 background (Crowther, Trofimovich, Saito, & Isaacs, 2014), task type (Crowther, Trofimovich, Isaacs, & Saito, in press), and speaker proficiency level (Saito, Trofimovich, & Isaacs, 2015). Although all of these variables appear to influence which L2 speech characteristics contribute to comprehensibility ratings, a common finding across these studies is that comprehensible L2 speech is linked to several linguistic dimensions of L2 output, including its segmental and suprasegmental content and fluency characteristics (e.g., Crowther et al., 2014; Kang et al., 2010; Saito et al., 2015).

Compared to research focusing on L1 listeners, there has been less work examining which speech characteristics contribute to comprehensibility for L2 listeners. Field (2005) investigated the impact of misplaced word stress on intelligibility for L1 and L2 listeners. O'Brien (2014) explored which speech measures were associated with L2 German listeners' ratings of comprehensibility of L1 and L2 German speech, and Winters and O'Brien (2013) investigated the impact of manipulating intonation and syllable duration in German speech for L1 and L2 German listeners. Matsuura et al. (2014) targeted the role of speech rate on intelligibility, and Jun and Li (2010), looking at how verbal reports about the comprehensibility and accent of L2 speech differed for L1 and L2 speakers of English, showed that there were differences between the two groups. There has also been work within a lingua franca perspective evaluating communication between L2 interlocutors for features of pronunciation which impact intelligibility (e.g., Deterding & Kirkpatrick, 2006; Jenkins, 2002). By way of summary, it appears that L2 listeners, especially in lingua franca contexts, may differ from L1 listeners in which aspects of speech they consider most important for comprehensibility; however, some features (e.g., lexical stress) may have a similar impact on intelligibility for both L1 and L2 listeners (e.g., Field, 2005). Despite increasing research in this area, it is still unknown how the linguistic measures that contribute to L2 comprehensibility judgments may differ for L2 listeners and speakers from different combinations of L1 backgrounds.

The Current Study

In their research targeting mutual understanding of L2 speech for different groups of L2 speakers and listeners, Munro et al. (2006) suggested, but did not directly test, the possibility that variations in comprehensibility ratings might largely reflect the properties of the speech itself. An interesting question, then, is how the language background of L2 listeners may impact the relative importance of various linguistic dimensions of L2 speech, such as its segmental, suprasegmental, and fluency characteristics, to comprehensibility judgments. Put differently, it is unclear to what extent L2 listeners might enjoy a possible comprehensibility benefit of sharing a language background with speakers, and whether this benefit might extend beyond what can be explained by specific linguistic features of L2 speech. Therefore, the chief goal of the current study was to examine whether L2 listeners' language background determines which speech characteristics contribute to their judgments of comprehensibility for L2 speakers from different language backgrounds and whether this benefit is limited to specific linguistic features used by listeners in rating L2 comprehensibility.

Another interesting question is how much listeners themselves attribute L1 background to their understanding of L2 speech when making comprehensibility judgments. Harding (2008) used qualitative data from L2 listeners to understand their views of having different accents on a listening test. He found that listeners viewed a shared L1 as a distraction, not an advantage, and that accent was less salient to low proficiency listeners. Verbal reports allow for a window into what is salient in the mind of the listener making decisions about speech samples, and offer an understanding of rating processes that quantitative ratings alone cannot (Gass & Mackey, 2000). Therefore, a methodological design feature of the current study was the use of verbal reports to obtain L2 listeners' perspective on the degree to which their rating of L2 comprehensibility can be attributed to a shared language background with L2 speakers.

In sum, there has been more interest in the role speaker background for L1 than for L2 listeners, and prior research has focused on narrow measures intelligibility rather than comprehensibility, with very few studies targeting extemporaneous speech (e.g., Munro et al., 2006). Further, previous research has not investigated whether there are differences in which linguistic variables contribute to comprehensibility ratings for listeners and speakers from different L1 backgrounds and whether there is a comprehensibility benefit beyond what can be explained by speech characteristics. Finally, there has been little research using verbal reports to examine whether listeners attribute their ability to understand speech to the L1 background of the speaker. Therefore, the present study addressed these gaps through a statistical comparison of comprehensibility ratings for different groups of L2 speakers and listeners, and through a descriptive analysis of verbal reports. The chief aim was to investigate the relative contribution of four speech measures (segmental errors, word stress errors, intonation, and speech rate) in the speech of L2 speakers to see whether matched or mismatched L1 background of the listeners and speakers determines which linguistic variables best explain comprehensibility ratings, and whether there is a comprehensibility benefit beyond what can be attributed to these speech measures. The following research questions were asked:

1. How do pronunciation, prosody, and fluency characteristics of L2 speech (segmental errors, word stress, intonation, speech rate) contribute to L2 listeners' judgments of comprehensibility for L2 speakers from different L1 backgrounds?
2. Is there a comprehensibility benefit beyond what can be attributed to pronunciation, prosody, and fluency characteristics of L2 speech (segmental errors, word stress, intonation, speech rate) for L2 listeners whose L1 background matches that of L2 speakers?
3. How much do L2 listeners attribute their L2 comprehensibility judgments to the language background of L2 speakers?

The first two questions were addressed through quantitative analyses of speech measures and comprehensibility ratings; the third question was examined through descriptive analyses of listeners' verbal reports.

Method

Participants

Listeners. The listeners were 30 L2 English users recruited from an English-medium university in Montreal (Canada). They came from three L1 backgrounds, with 10 listeners per group: Mandarin, French, and Hindi (7 women per group), referred to as Mandarin-L, French-L, and Hindi-L groups (where "L" stands for "listeners"), respectively. Another 10 native English listeners (6 women) were recruited for comparison purposes (English-L group). Because

Montreal is a bilingual city, it was expected that a large number of listeners would either be learning or already be proficient in French. Four of the Mandarin listeners were learning French, but none spoke it proficiently; five Hindi listeners were learning French, but none were proficient speakers. Among the English listeners, three were learning French, and seven were proficient speakers. As for the other target languages, one of the French listeners was learning Mandarin, but none of the other listeners spoke any Mandarin or Hindi. Aside from three native French speakers raised in Canada, the French listeners had resided in Canada for a mean of 3.3 years (2 months – 13 years). The Mandarin listeners had been in Canada for a mean of 10 months (2 months – 3.1 years), and the Hindi listeners for 8 months (1–14 months). Table 4 shows background characteristics of the four listener groups.

Table 4

Background Characteristics of Listeners

Background variables	Mandarin-L		French-L		Hindi-L		English-L	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Age (years)	19.8	1.3	24.2	5.6	23.9	2.2	31.5	11.8
Percent English use (0-100%)	54.5	23.2	56.0	17.9	74.5	7.3	81.0	11.0
Self-rated comprehensibility (1-9)	4.3	1.4	6.8	1.3	8.0	1.6	–	–
Familiarity with Mandarin (1-9)	8.7	0.5	3.2	2.3	1.0	0.0	4.4	2.8
Familiarity with French (1-9)	2.1	1.8	9.0	0.0	3.1	1.3	7.6	1.4
Familiarity with Hindi (1-9)	1.5	1.6	1.6	0.0	9.0	0.0	3.4	3.2

Note. English use (0% = *none*, 100% = *all the time*). Comprehensibility (1 = *very difficult to understand*, 9 = *very easy to understand*). Familiarity (1 = *not familiar*, 9 = *very familiar*).

Speakers. The 30 target speech stimuli were selected from an unpublished corpus of 143 L2 English speakers completing several speaking tasks (Isaacs & Trofimovich, 2011). The speakers came from three L1 backgrounds: Mandarin (5 women), French (4 women), and Hindi

(1 woman), referred to as Mandarin-S, French-S, and Hindi-S groups (where "S" stands for "speakers"), respectively. All speakers had an English proficiency level high enough to be admitted to credit programs at the university where the study took place. The target samples included the speakers' performance in a TOEFL iBT integrated task which requires speakers to listen to a short academic lecture and read a short passage on a similar topic and then answer a question that relates to both the lecture and the reading. This task was chosen because it requires speakers to use an academic register which language learners encounter, and are expected to produce, when studying in a post-secondary context. Two comparable versions of the task were used, with half of the speakers in each L1 group completing one and half completing the other. The speech samples were edited such that only the first 30 s of each were used for the ratings. Table 5 shows background characteristics of the three speaker groups.

Table 5

Background Characteristics of Speakers

Speaker variables	Mandarin-S		French-S		Hindi-S	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age (years)	23.4	3.1	20.7	2.2	24.0	1.4
Time in Canada (years)	0.7	0.4	0.5	0.2	0.5	.2
English study (years)	10.6	3.4	9.9	3.1	14.6	8.7
Speaking ability (1-10)	5.3	1.2	6.1	1.3	7.5	1.0

Note. Speaking ability (1 = *extremely poor*, 10 = *extremely fluent*).

The ratings given by the comparison group of L1 English listeners to each of the three L2 speaker groups were compared to determine if there were pre-existing differences across the L2 speakers. The targeted ratings included L1 English listeners' ratings of comprehensibility, segmental errors, word stress, intonation, and speech rate given to each of the three L2 speaker groups (see below). Repeated-measures ANOVAs revealed significant *F*-ratios for comprehensibility, $F(2, 27) = 15.5$, $p < .0001$, $\eta_p^2 = .54$, segmental errors, $F(2, 27) = 7.42$, $p =$

.003, $\eta_p^2 = .36$, word stress errors, $F(2, 27) = 11.23$, $p = .0001$, $\eta_p^2 = .46$, intonation, $F(2, 27) = 9.80$, $p = .001$, $\eta_p^2 = .42$, and speech rate, $F(2, 27) = 10.17$, $p = .001$, $\eta_p^2 = .43$. Bonferroni-corrected between-group comparisons further revealed that, for all measures apart from word stress errors, the French and Hindi speakers were not significantly different from each other, but both were rated significantly higher than the Mandarin speakers. For word stress, the French and Mandarin speakers did not differ significantly from each other, but the Hindi speakers were rated significantly higher than either group. Thus, the Mandarin speaker group was perceived by L1 English listeners as being less comprehensible and as having more problems with individual sounds, intonation, and speech rate, compared to the other two L2 speaker groups.

Speech Rating

The listeners were given individual appointments with the researcher. After completing a language background questionnaire, they first rated the 30 target speech samples for comprehensibility. The TOEFL iBT integrated task was explained to the listeners, and they were given a summary of the task to read (see Appendix A). They were then instructed that comprehensibility refers to a judgment of listening effort and were warned that the speech samples would be cut off after 30 s. The ratings were carried out using a computer-based continuous sliding scale, programmed in MATLAB, developed by Saito, Trofimovich, and Isaacs (in press). The scale, which presented the samples in a unique random order for each listener, included two labeled endpoints, corresponding to the rating of 0 as the left endpoint (*hard to understand*) and the rating of 1000 at the right endpoint (*easy to understand*). Consistent with prior research (e.g., Saito et al., 2015), listeners were not allowed to replay each file but could proceed to the next sample at their own pace. Listeners were also instructed to make a verbal report after each rating, explaining their reasons behind the comprehensibility ratings they gave for each speaker; these reports were audio recorded. The listeners were not given any examples of the types of explanations they could give to minimize researcher bias, but were encouraged to speak more if they provided little information during the practice phase.

After the comprehensibility ratings were completed, the listeners proceeded to rate the speech samples again, this time for several speech measures (including degree of foreign accent, which was not analyzed further) using similar 0-1000 continuous sliding scales (see Saito et al., in press, for validation of these rated measures against coded analyses of speech):

1. Segmental errors, which refer to vowel and consonant errors (e.g., substituting /d/ for /ð/ in “that”), with 0 corresponding to *frequent* and 1000 to *infrequent or absent.*
2. Word stress errors, which refer to errors in the placement of stress on words that contain more than one syllable (e.g., saying COMputer instead of comPUter), with 0 corresponding to *frequent* and 1000 to *infrequent or absent*.
3. Intonation, which applies to utterances longer than a single word and refers to the expected pitch contours associated with spoken utterances (e.g., using falling pitch to indicate a complete utterance) with 0 corresponding to *unnatural* and 1000 to *natural*.
4. Speech rate, which refers to how quickly or slowly someone speaks, with 0 corresponding to *too slow or too fast* and 1000 to *optimal*.

The listeners received a handout with explanations and examples of each speech measure (see Appendix B), and the researcher clarified any remaining questions about them. The rating procedure was similar to that used for the judgments of comprehensibility, with the exception that there were several scales on the screen, and the listeners could hear each speech sample as many times as needed to be confident in their judgments. Prior to starting the ratings, the listeners were given four practice files (two from each version of the task) using speakers from L1 backgrounds not used in this study. The listeners rated the speech samples using high quality headphones, and the researcher remained in the room while they completed the task, but sat at a computer facing away from the listeners in order to minimize their discomfort.

Analysis

The overall agreement among the L2 listeners, calculated using two-way random intraclass correlations, was high for all five rated measures: comprehensibility (α = .97), segmental errors (α = .90), word stress errors (α = .89), intonation (α = .94), and speech rate (α = .95). Therefore, the 10 listeners’ ratings within each listener group were averaged to derive a single mean rating per speaker. Pearson correlations and hierarchical regressions were used to investigate the relationship between the comprehensibility scores and the four speech measures for the L2 listener and speaker groups. A summary of all rated measures appears in Appendix C.

The verbal reports were transcribed and coded using 40 categories related to different aspects of comprehensibility. The current analysis centred only on the comments related to listeners’ own language background (L1-Own) or other language backgrounds (L1-Other).

Anytime listeners mentioned the L1 background of the speaker, the comment was coded as L1-Own if listeners were commenting about a speaker from their own language background and as L1-Other if they were commenting about a speaker from a different language group. In some cases, comments were coded even if an explicit reference to the L1 was not made, provided there was a clear reference to the rating being related to a specific L1. For example, the comment "it's easier to understand 'cause I I've hear this accent before" (Hindi-L 23) was coded as L1-Other even though this listener did not explicitly mention that the speaker was French.

Once the comments were extracted, they were further coded for whether the listener found the L1 of the speaker to be positive, neutral, or negative in terms of its relevance to comprehensibility. A comment was coded as positive even if the listener found the speaker difficult to understand, provided he or she saw the L1 as beneficial, as in "The speaker 26 is pretty hard to understand but since I'm French I get it a bit" (French-L 8). Similarly, a comment was coded as negative, even though the speech sample was considered easy to understand, if the speaker indicated that the L1 was harmful to comprehensibility, as in "I think I can guess what her idea even even though her accent, uh which I'm not familiar with" (Mandarin-L 2). Comments showing no clear positive or negative attitudes towards the L1 were given a neutral label, as in "…I think um she was again he's uh Indian and uh but it's a little different from the previous one, uh because um his vocab and speak some words more clearly" (Mandarin-L 6).

Results

Speech Measures and L2 Comprehensibility

The first analysis targeted the first research question, namely, the extent to which segmental errors, word stress, intonation, and speech rate contribute to L2 listeners' judgments of comprehensibility for L2 speakers from different L1 backgrounds. To address this question, a series of Pearson correlations were run to compare the relationship between each listener group's comprehensibility ratings and the same group's judgments of the four speech measures for each of the three speaker groups (Bonferroni-corrected *a* = .001). As summarized in Table 6, for the Mandarin speakers, the comprehensibility ratings given by each of the listener groups were not associated with these groups' ratings of the four speech measures. When rating the French speakers, the French listeners' comprehensibility rating was linked to two speech measures (segmental errors, speech rate); the Mandarin listeners' comprehensibility rating was associated with three speech measures (word stress errors, intonation, and speech rate); and the Hindi listeners' comprehensibility rating was correlated with three speech measures (segmental errors, word stress errors, and intonation). Finally, when rating the Hindi speakers, the Hindi listeners' comprehensibility rating was correlated with two speech measures (segmental errors, speech rate); the Mandarin listeners' comprehensibility rating was associated with three speech measures (segmental errors, word stress errors, and intonation); and the French listeners' comprehensibility rating was linked to a single speech measure (speech rate). In essence, whenever significant associations were detected, there was a partial (yet far from perfect) overlap between the speech variables associated with L2 comprehensibility for each of the three listener-speaker group combinations.

Table 6

Pearson Correlations between Comprehensibility Scores and Speech Measures for Each Listener/Speaker Group

Speaker group/rating	Mandarin-L	French-L	Hindi-L
Mandarin-S			
Segmental errors	.23	.47	.77
Word stress errors	.15	.06	.80
Intonation	.67	.29	.50
Speech rate	.81	.52	.80
French-S			
Segmental errors	.79	.90*	.90*
Word stress errors	.86*	.72	.89*
Intonation	.95*	.78	.90*
Speech rate	.93*	.84*	.80

Speaker group/rating	Mandarin-L	French-L	Hindi-L
Hindi-S			
Segmental errors	.86*	.82	.83*
Word stress errors	.90*	.82	.82
Intonation	.90*	.74	.80
Speech rate	.77	.86*	.95*

Note. *$p < .001$ (one-tailed). Shaded cells designate a shared L1 background for speakers and listeners.

Shared Background Benefit

The next analysis targeted the second research question, that is, the extent to which shared L1 background contributed to the variance in L2 comprehensibility scores given by L2 listeners to L2 speakers after accounting for the contribution of the four speech measures. For this analysis, three hierarchical multiple regressions were run (one per L2 listener group), with comprehensibility rating as the outcome variable. In each regression, the four speech measures (segmentals, word stress, intonation, speech rate) were entered into the regression model in Step 1, to estimate the amount of variance in comprehensibility accounted for by the characteristics of L2 speech as rated by a given listener group. Then, language background, coded as a dummy categorical variable with the matching listener-speaker background as a reference group, was added in Step 2, to determine the unique contribution of language background to L2 comprehensibility judgments by each listener group.

For the Mandarin listeners, speech rate, $B = .73$, $SE\ B = .25$, *ß* = .50, and word stress errors, $B = .82$, $SE\ B = .41$, *ß* = .34, together accounted for 73% of the total variance in the comprehensibility scores given to the three L2 speaker groups, $F(4, 25) = 20.99$, $p < .0001$. Adding language background in Step 2 produced a 6% improvement in the model fit, $F(2, 23) = 3.99$, $p = .033$. The Mandarin listeners downgraded both the French speakers relative to the Mandarin speakers, $t = –2.65$, $p = .014$, and the Hindi speakers relative to the Mandarin speakers, $t = –2.21$, $p = .037$. Using a 1000-point rating scale, the Mandarin listeners rated the Mandarin speakers' comprehensibility by 124 points higher, compared to the French speakers, and by 136 points higher, compared to the Hindi speakers.

For the French listeners, segmental errors, $B = .73$, $SE\ B = .24$, $ß = .43$, and speech rate, $B = .45$, $SE\ B = .20$, $ß = .37$, together accounted for 82% of the total variance in the comprehensibility scores given to the three L2 speaker groups, $F(4, 25) = 34.14$, $p < .0001$. Adding language background in Step 2 did not produce a significant improvement in model fit, $F(2, 23) = 1.56$, $p = .23$. For the Hindi listeners, speech rate alone, $B = .75$, $SE\ B = .19$, $ß =.63$, accounted for 86% of the total variance in comprehensibility scores given to the three L2 speaker groups, $F(4, 25) = 46.22$, $p < .0001$. Again, adding language background in Step 2 did not produce an overall significant improvement in model fit, $F(2, 23) = 2.69$, $p = .09$. Nevertheless, the Hindi listeners showed a significant tendency to downgrade the Chinese speakers relative to the Hindi speakers, $t = –2.10$, $p = .047$, rating the comprehensibility for the speakers of their own L1 background by 119 points higher.

Verbal Reports about Language Background

The last analysis focused on the final research question, which concerned the degree to which L2 listeners overtly attribute their comprehensibility judgments to the language background of L2 speakers. Analyses of verbal reports indicated that the French listeners overall made the most comments relating to the language backgrounds of the speakers (55), followed by the Mandarin listeners (28), with the Hindi listeners reporting far fewer comments (6). The lower number of comments for some groups, compared to others, is not reflected in the larger dataset. With all 40 coded categories included, the percentage of each group's total comments in the study was roughly similar (Mandarin-L = 24%, French-L = 25%, Hindi-L = 23%, English-L = 29%). And within each listener group, not all listeners commented on language background, with more Mandarin listeners (Own-L1 = 6; Other-L1 = 6) and French listeners (Own-L1 = 2; Other-L1= 6) providing comments, compared to the Hindi listeners (Own-L1 = 1; Other-L1 = 2).

Both the Mandarin and the French listeners made positive comments far more often when talking about their own L1 backgrounds, with 82% of the Mandarin listeners' comments and 77% of the French listeners' comments being positive (see Table 7). For example, a French listener noted, "Uh I have not difficulties to understand him but uh I'm I'm think, I think it's a French so it's easy to me to to understand with his uh with his accent. Uh he has a French pronunciation of the words so I so I understand uh what he said" (French-L 20). There were also comments from three listeners suggesting that a speaker would be hard to understand for those who did not share his or her L1, as in "He's Chinese so I can understand what, what he's talking about. But maybe for some other people who is not familiar with Chinese accent, with Chinese like logic, maybe it's a little bit hard to understand" (Mandarin-L 5). The negative comments about listeners' own background often reflected difficulties the listeners attributed to speakers from their language background. For example, a Mandarin listener commented, "I guess he's a Chinese because his pronunciation is not very very good" (Mandarin-L 17). As there were only two Own-L1 comments in the Hindi-L group, it is difficult to generalize for this group.

Table 7

Frequencies of the L1-Own Category Comments for L2 Listener Groups

Comments	Mandarin-L	French-L	Hindi-L	All L2 listeners
Positive	13	20	1	34
Neutral	1	3	0	4
Negative	2	3	1	6

In terms of talking about other L1 backgrounds (see Table 8), all three L2 listener groups made far more negative than positive comments (Mandarin-L= 75%, French-L = 75%, Hindi-L = 100%). The French listeners' negative comments were divided fairly evenly between the Mandarin (42%) and Hindi speakers (58%). However, both the Mandarin and Hindi listeners made far fewer negative comments about the French speaker group, compared to the other groups (22% and 25%, respectively). Not all of the negative reports indicated that the listeners actually struggled to understand the speaker. For example, one French listener commented, "The speech was understandable even though the Indian accent was present" (French-L 24). However, many of the comments did indicate that the listeners struggled to understand the speech because of the speaker's L1. For example, a Hindi listener reported, "Speaker number 14's uh comprehensibility was hard to understand since he had a Chinese accent for sure" (Hindi-L 16).

Table 8

Frequencies of the L1-Other Category Comments for L2 Listener Groups

	Mandarin-L	French-L	Hindi-L	All L2 listeners
Mandarin-S				
Positive	0	0	0	0
Neutral	0	0	0	0
Negative	0	10	3	13
French-S				
Positive	1	0	0	1
Neutral	0	0	0	0
Negative	2	0	1	3
Hindi-S				
Positive	0	1	0	1
Neutral	2	2	0	4
Negative	7	14	0	21
All L2 speakers				
Positive	1	1	0	2
Neutral	2	2	0	4
Negative	9	24	4	37

Discussion

The current study examined (a) how pronunciation, prosody, and fluency characteristics of L2 speech (rated in terms of segmental errors, word stress errors, intonation, and speech rate) contribute to L2 listeners' judgments of comprehensibility for L2 speakers from different L1 backgrounds, (b) whether there exists a comprehensibility benefit beyond what can be attributed to pronunciation, prosody, and fluency characteristics of L2 speech, and (c) whether L2 listeners overtly attribute their comprehensibility judgments to the language background of L2 speakers. With respect to the relationship between L2 comprehensibility and pronunciation, prosody, and fluency characteristics of L2 speech, correlation analyses revealed that there were differences in which speech measures were significantly related to comprehensibility judgments for different combinations of L2 listeners and speakers. No speech measures were linked to the Mandarin speakers' comprehensibility scores for any of the listener groups, and different clusters of speech measures were associated with the comprehensibility scores given by the listeners to the other speaker groups.

In terms of the effect of shared language background on L2 comprehensibility, regression analyses revealed that after the variance from the four targeted speech measures (segmental errors, word stress errors, intonation, speech rate) was taken into account, language background accounted for an additional 6% of variance in comprehensibility ratings for the Mandarin listeners. That is, the Mandarin listeners downgraded both the French and the Hindi speakers, relative to the Mandarin speakers. For the French and Hindi listeners, language background did not significantly contribute to the comprehensibility scores, although the Hindi listeners showed a tendency to downgrade Mandarin speakers relative to speakers from their own background. Analyses of verbal reports showed that the L2 listeners, at least in some cases, overtly considered L1 background as a factor in rating L2 comprehensibility. This appeared to be the case more for the Mandarin and French listeners, who made more comments related to the L1s of the speakers, than for the Hindi listeners, who made fewer. The L2 listeners were also more likely to make positive comments about L1 background when sharing an L1 with the speaker, and to make negative comments when the L1s did not match.

Speech Variables and Listener-Speaker Language Background

There were clear L1 background effects with respect to the relationship between listeners' judgments of L2 comprehensibility and listener-rated pronunciation, prosody, and fluency characteristics in L2 speakers' output (see Table 6). There was a partial (yet far from complete) overlap in the speech variables associated with ratings of L2 comprehensibility across the three listener groups. This was particularly true for L2 users of French and Hindi, where a mismatch in language background between listeners and speakers was characterized by either a narrower or a wider range of speech measures associated with L2 comprehensibility, relative to the shared listener-speaker L1 background. For instance, for the French listeners evaluating the French speakers, segmental errors and speech rate were the two speech measures associated with L2 comprehensibility. However, for the Mandarin and Hindi listeners, the French speakers' comprehensibility was linked only to one of these two measures, and these listener groups associated the French speakers' comprehensibility with two additional variables, which did not seem to be relevant to the French listeners. If ISIB effects arise due to shared phonological knowledge between listeners and speakers (e.g., Bent & Bradlow, 2003; Hayes-Harb et al., 2008), then shared comprehensibility benefits, as demonstrated here, likely stem from a similar knowledge overlap. With no common language background available, listeners likely resort to the linguistic cues which appear most relevant for comprehension.

The Mandarin speakers were the only group for which no speech measure correlated with comprehensibility ratings given by any listener group. However, compared to the French and the Hindi speakers, the Mandarin speakers were also rated the weakest in four out of five speech ratings they received from the English listeners, and the ratings given to the Mandarin speakers by the other L2 listener groups generally featured smaller values, compared to the ratings given to other speakers (see Appendix C). In fact, the English listeners also showed no significant associations between their judgments of the Mandarin speakers' comprehensibility and their ratings of the four speech measures for these same speakers (Appendix C). Thus, it may have been difficult for both L1 and L2 listeners to detect any salient relationships between speech variables and L2 comprehensibility for the (low proficiency) Mandarin speakers, particularly if listeners overall struggled with understanding these speakers. At a broader level, this finding parallels the results reported by Crowther et al. (2014) for relatively high proficiency Farsi speakers, where none of the 10 targeted speech measures bore strong relationships with their L2 comprehensibility as assessed by L1 English listeners. It could be that for L2 speakers at relatively high and low oral ability levels, whether or not they are evaluated by L1 or L2 listeners, comprehensibility may be based on a range of variables, with no single factor bearing a particularly strong relationship with comprehensibility. This possibility needs to be explored in future research.

Taken together, these results align well with findings from recent research showing that the linguistics variables which contribute to comprehensibility can vary based on several factors, including speakers' L1 (Crowther et al., 2014), speaking task being targeted (Crowther et al., in press), and speakers' proficiency level (Saito et al., 2015). Although the current dataset revealed clear L1 background effects for L2 listeners evaluating L2 speech, such that the pronunciation, prosody, and fluency characteristics of L2 output relevant to comprehensibility may be specific to a particular listener-speaker combination, the current findings also implied some similarities. For example, as shown in Table 6, for the French and Hindi listener-speaker combinations, the two speech measures most strongly correlated with comprehensibility were identical (segmental errors, speech rate), and these same measures featured the largest number of significant associations across all listener groups (four, in each case). Moreover, regression analyses yielded speech rate as the only speech measure to serve as a significant predictor of L2 comprehensibility for all three L2 listener groups. It may well be that, regardless of listener-speaker combinations, some aspects of segmental accuracy and, most prominently, some characteristics of speech fluency, such as speech rate, might be universally relevant to L2 listener perception of L2 comprehensibility. This finding would certainly support lingua franca approaches to L2 pronunciation learning and teaching, with their strong emphasis on segmental aspects of L2 speech (e.g., Jenkins, 2000; Walker, 2010). And this finding is also compatible with L1 and L2 listeners' sensitivity to various fluency dimensions of L2 speech (as shown in Study 1 of this thesis). If some speech dimensions feeding into listener judgments of L2 comprehensibility, such as segmental accuracy and speech rate, are similar across various L2 speaker-listener combinations, then, these dimensions could serve as instructional foci in language classrooms comprised of learners with different language backgrounds who will be using their L2 with different interlocutors.

Benefits of Shared L1 Background

Findings from ISIB research have shown that the benefit of a shared L1 can be moderated by proficiency. For example, Xie and Fowler (2013) found a gradient effect for the proficiency of listeners (operationalized through perceptual accuracy), whereby stronger ISIB effects were associated with listeners of lower L2 proficiency. Similarly, Hayes-Harb et al. (2008) found that low proficiency Mandarin listeners (defined through accent ratings) exposed to Mandarin speakers of low proficiency outperformed English L1 listeners in a word identification task while the high proficiency Mandarin listeners did not. The current finding are consistent with these results, in that the Mandarin listeners enjoyed the greatest benefit from having a shared L1 background with the speakers whose oral ability level was relatively low. Although an independent measure of L2 speaking proficiency for the listeners was not available (e.g., TOEFL or IELTS speaking score), the Mandarin listeners also self-reported the lowest comprehensibility ratings of the three groups (see Table 3). However, neither the listeners nor the speakers were beginners in English, and all were enrolled in an English-medium university. While the Mandarin listeners may have been of lower overall proficiency, compared to the other listeners, they were nevertheless within the proficiency range that would be expected of students accepted to an English post-secondary setting in Canada. These findings thus suggest that listeners who share an L1 background with lower proficiency speakers may perceive that speech as easier to understand than the speech of speakers from other L1 backgrounds, even after the shared knowledge of pronunciation, prosody, and fluency characteristics of L2 output is considered. This has implications for learners from the same L1 background who are studying in a common L2, such that these learners may have an inflated sense of how comprehensible each other's speech will sound to interlocutors who do not share their language background (see Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2015, for evidence of low proficiency L2 listeners overestimating their L2 comprehensibility).

One interesting question is what exactly brought about a unique comprehensibility benefit from shared listener-speaker L1 background, beyond what was explained by shared understanding of pronunciation, prosody, and fluency characteristics of L2 speech. This additional benefit was found for the Chinese listeners and speakers, and a similar tendency was attested for the Hindi listeners and speakers. Because this study targeted extemporaneous speech, as opposed to scripted or repeated language, it is likely that additional contributions to the comprehensibility benefit based on a shared L1 background stemmed from aspects of L2 speech not captured in this study. For example, several studies have found that linguistic dimensions unrelated to pronunciation, such as grammar, vocabulary, and discourse richness, contribute to comprehensibility (e.g., Crowther et al., 2014, in press; Isaacs & Trofimovich, 2012). The following comment from a Mandarin listener suggests that L1-based discourse structure may have played a role in increased understanding, "He's Chinese so I can understand what, what he's talking about. But maybe for some other people who is not familiar with Chinese accent, with Chinese like logic, maybe it's a little bit hard to understand" (Mandarin-L 5).

While ISIB research is typically only interested in specific phonological aspects of a shared interlanguage, such as voicing in word endings, other speech commonalities that are part of a shared language background could be important to comprehensibility in real-world interaction. There is also a wealth of evidence that factors unrelated to speech can play a role in speech judgments, such as bias and listener expectations (e.g., Kang & Rubin, 2009; Lippi-Green, 1997). Kang (2012) found that variance in undergraduate students' proficiency ratings of international teaching assistants could partially be explained by measures of prosody, but that the listeners' native-speaker status and other background variables (e.g., experience with L2 speakers and tutoring experience) also contributed to the ratings, though to a lesser degree. Thus, the additional 6% increase in comprehensibility (not captured through ratings of pronunciation, prosody, and fluency) for the Mandarin listeners evaluating the Mandarin speakers may have stemmed from a variety of linguistic and non-linguistic factors, including listener expectations and bias. Understanding possible benefits of shared L1 backgrounds that reach beyond the properties of L2 speech therefore represents a most interesting avenue for future research into L2-L2 comprehensibility.

Verbal Reports as a Window on Shared L1 Benefits

The verbal report data only partially supported what was found in the regression analyses. The Mandarin listeners made a number of comments related to an improved understanding of L2 speech from their own language background and a decreased understanding of speakers from other backgrounds, which was reflected in the regression model for the Mandarin listeners. However, language background had no predictive power in regression models for the French listeners, yet the French listeners made the most comments about the ease with which they could understand speech from their own L1 background. The Mandarin and French listeners also made negative comments about understanding the Hindi speakers, which was not reflected in the quantitative analysis. Thus, a relatively poor agreement between the results of quantitative analyses of speech ratings and descriptive coding of verbal reports suggests that what listeners report may not be what actually influences their speech ratings. Hayes-Harb and Hacking (2015) recently investigated reasons underlying L1 English listeners' accent judgments of L2 English speakers. They found that the raters frequently strayed from simply analyzing the speech, instead "demonstrating a tendency to imagine and attempt to describe the social-cultural backgrounds of the speakers" (p. 62). It may well be, then, that listeners overtly view L1 as being important, when in actuality it may not dramatically impact the ratings they give to speakers of their own and other L1 backgrounds or may impact their ratings differently from what they ascribe L1 influence to be. At least in the current dataset, it is possible that the characteristics of the speech which made a particular speaker easy or difficult to understand were captured in the speech ratings. As a result, L1 background may not have featured prominently in the listeners' verbal reports, especially because the listeners' task was to report on *anything* that mattered for L2 comprehensibility. Thus, what is interesting to explore in future research is how various linguistic and non-linguistic aspects of L2 speech fare against L1 background, as reflected in listeners' verbal reports about L2 comprehensibility.

Limitations and Conclusion

There are a number of limitations to this research that should be addressed in future studies. First, the sample size in this study was limited, making both the correlational and regression analyses exploratory in nature. With more participants, future research could more easily compare different listener groups. Further, tighter controls over the proficiency of listeners and speakers would allow for a clearer understanding of the role of proficiency and linguistic variables in comprehensibility judgments. It would also be interesting to target different listener and speaker groups, and include a more exhaustive list of speech measures, including those unrelated to pronunciation, such as grammatical accuracy or vocabulary use. Finally, a more explicit set of instructions accompanying verbal reports may have yielded clearer findings. The lack of direction given to the listeners when making verbal reports was intentional, and was done to prevent researcher bias. However, interviews or focus groups may provide more appropriate measures when looking for a specific type of information that listeners attend to in L2 speech.

As English continues to spread around the globe, with an increasing number of English users being L2 speakers communicating with other L2 speakers, it is important that studies investigating comprehensibility and intelligibility reflect this complexity. This study found that different L1 listener-speaker combinations resulted in overlapping yet non-identical linguistic variables contributing to comprehensibility ratings. Further, comprehensibility ratings were uniquely associated with a matching listener-speaker L1 background benefit, at least for the speakers with lower L2 speaking ability. Finally, listeners often attributed the language background of the speakers to the ease or difficulty they experienced when rating speech samples, though this was not always reflected in the actual speech ratings. The entire concept of having speech that is easy or difficult to understand is multifaceted, comprised of qualities of speech and characteristics of the listener, not to mention many variables relating to the context of a given interaction. While it is impossible for any study to capture the full range of these complexities, the closer we come to understanding comprehensibility in a wide range of contexts, the more potential there will be to help language learners communicate successfully in whichever setting they plan to use their L2.

Connecting Study 1 and Study 2 to Study 3

Taken together, the first two studies in this dissertation demonstrated that there are differences in how speech is perceived by different listeners. Study 2 takes this further by demonstrating that there are differences in what matters most to comprehensibility based on the language background of the speaker and the listener. This presents a challenge in terms of pronunciation instruction because – for learners from different language backgrounds – different aspects of pronunciation may require more focus. Study 3 is an attempt to find a partial solution to this problem by testing the effectiveness of a pronunciation practice activity that has the potential to help learners improve their pronunciation even if the areas in which they need the most help may vary. The activity presented in Study 3 – shadowing – is a practice technique in which learners imitate speech models and listen to their own recorded voices to better notice how their own speech differs from that in their target language.

Study 3 features research that has direct implications for pedagogy, as one of my main goals as a researcher is to conduct research that is applied in nature and can help learners improve their pronunciation. In this sense, this study represents pedagogical extension of the ideas targeted in the first two studies in this dissertation.

Chapter 4: Study 3

Using Shadowing with Mobile Technology to Improve L2 Pronunciation

To be submitted to the Journal of Second Language Pronunciation

By Jennifer A. Foote

Abstract

Shadowing, a common pronunciation practice technique, has been demonstrated to improve various aspects of second language learners' pronunciation but few studies have investigated whether these changes in pronunciation impact untrained listeners' perceptions. In the present study, sixteen participants were given iPods to use to practice shadowing short dialogues from television shows for eight weeks. The participants were required to practice at least four times per week and to record themselves while shadowing. Two language tasks (a shadowing task and an extemporaneous speaking task) were administered as pre-, mid-, and post-tests, and were rated by 21 first language speakers of English. The shadowing task was rated for learners' ability to imitate a speech model and the extemporaneous speaking task was rated for comprehensibility, accentedness, and fluency. Interview data were also collected during the study to gauge participants' opinions of the activities. Results indicated that the participants improved significantly on all speaking measures apart from accentedness and were largely positive about the activities.

Introduction

Second language pronunciation has traditionally received less attention from second language acquisition (SLA) researchers than other language skills such as grammar and vocabulary. However, as has been noted in a recent review of studies in this area (Thomson & Derwing, 2014), the past several years have seen an explosion of interest in pronunciation. There are now enough studies to warrant a meta-analysis on the efficacy of instruction on pronunciation and it is clear that pronunciation instruction is often effective (Lee, Jang, & Plonksy, 2014). However, while many studies have demonstrated that pronunciation can be changed, only a small percentage have used what Thomson and Derwing (2014) refer to as the "gold standard" of demonstrating that the changes measured in pronunciation studies impact how comprehensible participants' speech is after treatment (p. 7). A technique may show promise when it can be demonstrated to lead to changes in pronunciation, but if those changes don't make an impact on learners' abilities to successfully communicate in their second language (L2), then it is questionable whether such a technique is actually worth the time it takes to complete it. For example, a study may show that measurements of a learners' accuracy in producing a phoneme such a /r/ improves after a series of training exercises on a computer. However, these changes may not make any noticeable differences to how that person sounds when speaking with an interlocutor. Further, even if those changes can be detected by a listener, they may not significantly improve how easy that learner is to understand.

Even if a pronunciation intervention passes the gold standard test of demonstrating that it can lead to improved comprehensibility, its utility for pronunciation instruction may still be limited if it can only be shown to work in a controlled laboratory setting. In a survey of 75 pronunciation intervention studies, Thomson and Derwing (2014) found 39% used some form of computer assisted pronunciation instruction as an intervention rather than traditional classroom instruction. When only looking at studies published in peer-reviewed journals, this increased to 69%. While some of these studies allowed learners to access training in their own time, outside of a laboratory setting, most did not, leading to a problem of ecological validity among many of the studies using computer-assisted pronunciation instruction. With growing evidence that instruction can improve pronunciation, it is now important for more studies to investigate whether instruction can lead to changes in learners' comprehensibility, and if those changes can occur outside of a controlled laboratory setting.

In order to discuss what it means for a study to show changes in learners' comprehensibility, it is necessary to understand the relationship between accentedness and comprehensibility/intelligibility. Accentedness refers to how much a speaker's phonology differs from that of a first language (L1) speaker of that language. In the literature, intelligibility and comprehensibility are often used interchangeably, as both refer to how understandable speech is. However, in research contexts, they are often operationalized differently (e.g., Munro & Derwing, 1995a). Comprehensibility is usually a subjective measure of how difficult to understand a listener perceives speech to be, while intelligibility typically refers to objective measures of whether speech was actually understood. Accentedness and comprehensibility are generally measured by having raters make subjective scalar judgments, while intelligibility is typically measured by testing whether listeners were actually able to understand an utterance (e.g., by having the listener transcribe what they hear). Accentedness, though partially related to comprehensibility and intelligibility, is also partially independent (Derwing & Munro, 1997; Munro & Derwing, 1995a, b); while a person who is considered to be low in terms of comprehensibility/intelligibility will also be rated as highly accented, the reverse is not necessarily true. A person can have a very noticeable accent but still be considered easy to understand by his or her interlocutors.

A number of studies investigating pronunciation and comprehensibility also investigate some aspect of fluency. Fluency, used here in the narrow temporal sense of "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (Ellis, 2003, p. 342) rather than the general meaning of overall proficiency, also contributes to a listener's comfort level when listening to L2 speech (Derwing, Munro, & Thomson, 2008). Not surprisingly, it is also related to comprehensibility (Derwing, Rossiter, Munro, & Thomson, 2004; Isaacs & Trofimovich, 2012) as temporal measures such as pausing can also be an important aspect of prosody in pronunciation. For this reason, interventions designed primarily to impact pronunciation, may also provide additional benefits to speakers by improving their fluency (e.g., Derwing, Munro, & Wiebe, 1998) though this is not always the case (Derwing, Munro, Foote, Waugh, & Fleming, 2014). Although increased fluency is not the primary focus of most pronunciation studies and teaching methods, a pronunciation technique that also improves learners' fluency provides additional benefits to helping learners communicate successfully in their L2.

While some pronunciation programs and techniques are primarily concerned with eliminating an L2 speaker's accent as much as possible, it is now generally accepted that a goal of increased comprehensibility is both more realistic and more appropriate for pronunciation instruction (see Levis, 2005, for an overview of the history of these two views of pronunciation instruction). This has led to a number of studies investigating which aspects of pronunciation most strongly contribute to making speech more easily understood (e.g., Field, 2005; Hahn, 2004; Isaacs & Trofimovich, 2012; Munro & Derwing, 2006; Zielinski, 2008). However, regardless of which aspects of speech contribute most to comprehensibility in general, the specific difficulties and challenges learners face will vary based both on L1 influence and learners' individual differences. One partial solution to this problem is to find activities and techniques that can help learners notice the gaps between their own pronunciation and that of their target language without limiting learners to focusing on a specific feature. In this way, the same activity could be of equal benefit to learners with different pronunciation difficulties that are affecting their comprehensibility: a situation that is common in multilingual classrooms. It will also help learners who want to study independently, but who are not sure which aspects of their own pronunciation need the most practice, which is the case for many L2 speakers of English (Derwing, 2003). One technique that shows promise in this area is *shadowing*, a technique that has been common in pronunciation teaching manuals and classrooms for many years, but has only recently received much attention from L2 researchers.

Shadowing is an activity where learners imitate a presented speech stimulus "as closely and quickly as possible" (Luo, Shimomura, Minematsu, Yamauchi, & Hirose, 2008, p. 4) though the repetition can be near simultaneous or have a small delay (e.g., Goldinger, 1998; Hiramatsu, 2000; Schweda-Nicholson, 1990). While shadowing has existed for decades, its roots are not in language instruction. Shadowing has been used in cognitive psychology for testing selective attention (Boyee & Stewart, 2009) and as a language therapy to help treat stuttering (Li-Chi, 2009). In Japan, it is a popular, though somewhat controversial, technique for training simultaneous interpreters (Boyee & Stewart, 2009). In fact, searches for recent research on shadowing yield more articles on Japanese interpreter training than L2 language training. This may be the reason why most of studies conducted on using shadowing for language teaching purposes have been done in Japan where it has spread from interpreter training to become a very popular and common classroom activity (Hiramatsu, 2000; Saito, Nagasawa, & Ishikawa, 2011). While shadowing is not as common-place in other countries, it is still fairly prevalent in pronunciation classrooms. It is included in popular pronunciation teaching guides and handbooks (e.g., Avery & Ehlrich, 1992; Celce-Murcia, Brinton, & Goodwin, 2010) and for over thirty years, articles promoting shadowing for pronunciation instruction and offering suggestions on how to use it in the classroom have been published in journals (e.g., Acton, 1984; Rosse, 1999). A search of "pronunciation shadowing" on Google™ will reveal a multitude of instructor- and learner-oriented websites promoting shadowing for language development. With advances in technology, shadowing is an activity that can easily be completed by learners outside of a classroom setting and at minimal cost, making it potentially very useful for learners who do not have access to formal pronunciation instruction.

Despite its presence in language classrooms, the actual research that has been conducted on shadowing as a language-learning tool is limited. This may be partially because it is reminiscent of the much-maligned audiolingual approach to language teaching, leading detractors to argue that it is "…just vocalized repetitions and only results in meaningless parrot-like practice" (Bovee & Stewart, 2009, p. 20). However, the anecdotal experiences of instructors as well as research that has been done on shadowing to date, have shown promise. My own interest in this subject stems from conversations with highly successful language learners who would reveal that repeatedly shadowing dialogues from TV shows or movies had been a large part of their language practice activities at home. Case studies of highly successful language learners have reported similar findings with successful learners often attributing their success in part to substantial amounts of time practicing imitating voices from speech models (Ding, 2007). A study by Martinsen, Alvord, and Tanner (2014) investigated the role of motivation, instruction, cultural sensitivity, and time studying abroad on accentedness ratings of 102 learners of Spanish had a similar discovery. They found that:

> *surprisingly, the highest rated learner in the study did not have extensive experience abroad... However, beginning as much as six months prior to studying abroad, she developed a plan to improve her pronunciation. Three to four days each week, she spent a minimum of 15 minutes listening to Spanish newscasts or other online media and imitating as closely as possible the speaker's words and phrases* (p. 74).

Of course, these individual cases of successful language learning are not sufficient evidence of the efficacy of shadowing. However, the studies that have been done indicate that shadowing is an effective technique for improving various aspects of L2 language development.

Many of the studies that have investigated the efficacy of shadowing, have been primarily interested in the role of shadowing in listening comprehension. Overall, these studies have found shadowing to be effective for this language skill (see Bovee & Stewart, 2009, and Hamada, 2014, for overviews). There has also been some interest in the effects of shadowing on general measures of speaking proficiency (e.g., Li-Chi, 2009). However, there have also been studies that have investigated the efficacy of shadowing for pronunciation improvement (e.g., Bovee & Stewart, 2009; Hsieh, Dong, & Wang, 2013; Mori, 2011; Rongna & Hayashi, 2012). All of these studies found some improvements in the speech measures they used for analysis. Unfortunately, these studies have not tended to follow Thomson and Derwing's (2014) gold standard approach to assessing pronunciation. The speech samples used for analysis came from participants completing either a shadowing task or read aloud tasks. The measurements were most often made by computers or, in one case, a combination of computer-based acoustic analysis and one expert rater (Rongna & Hayashi, 2012). Only one of the studies used multiple human raters to gauge improvement. In this case (Bovee & Stewart, 2009) eight L1 speakers of English rated randomizations of the pre- and post-test shadowing samples. However, the raters were asked to judge the speech samples "on the basis of overall quality (i.e., closeness to native-like pronunciation)" (p. 892). The other studies measured discrete features including pitch accent, intonation, final lengthening, and pronunciation of words, though one study (Hsieh et al., 2013) also included a computer analysis of "overall pronunciation" (p. 892). For this reason, all of these studies add to the evidence that shadowing can improve pronunciation; however, only one has measured whether this improvement is noticeable by non-expert human listeners, and none of them tested whether the improvements led to more comprehensible speech in extemporaneous speaking contexts.

Due to the increasing ubiquity of portable technologies such as smart phones, tablets, and digital music players, shadowing is an activity that is now very easy to implement as a homework activity. Further, for learners who are not enrolled in a pronunciation class, shadowing offers a cost-effective way of practicing pronunciation independently. This is especially important given that pronunciation does not always get much attention in language classes (Foote, Trofimovich, Collins, & Soler-Urzúa, 2013), possibly because teachers feel underprepared to teach pronunciation effectively (e.g., Foote, Holtby, & Derwing, 2011; Henderson et al., 2012). While much of the research on computer-assisted pronunciation instruction has focused on using actual computers, the consumer demand for pronunciation materials on mobile technology platforms has led to a wide range of commercial pronunciation and accent reduction apps appearing on the market in the past few years (Foote & Smith, 2013). However, in order for shadowing to be effective when used independently with mobile devices, learners need to see its value and choose to use it rather than seeing it as the "meaningless parrot-like practice" it has been accused of (Bovee & Stewart, 2009, p. 20).

There is some research to draw on when looking at learners' attitudes towards shadowing. Some of the studies that have examined the impact of shadowing on pronunciation have also included surveys asking for learners' opinions of the activities. Li-Chi's (2009) study had 25 eighth grade students in Taiwan take part in shadowing activities for 15 hours over five weeks. Questionnaires and semi-structured interviews with the participants revealed that, overall, participants felt more confident about their speaking abilities after completing the shadowing project. Some participants also noted that they thought shadowing would be good for studying outside of class, and that the use of recorders enabled them to find and correct their own errors. However, over a quarter of the participants reported that they found shadowing difficult and some noted that they found the activities overly repetitive and boring. Only one study has asked learners about their opinions of shadowing activities conducted outside of class time. Bovee and Stewart (2009) had learners complete shadowing activities as homework assignments completed using computers. They found that 67% of respondents thought their pronunciation of individual words had improved, 73% thought intonation improved, and 80% thought the activity had educational value. The participants were also given the option of writing comments about shadowing, with 64% of comments being classified as positive comments and 34% classified as negative. Of negative comments, 30% related to technical and logistical issues, 20% were about the difficulty of the task, 20% found it too time consuming, and the remaining comments related to the conditions of the computer labs at the university where the study took place. These surveys suggest that students' experiences with shadowing have been generally, but not entirely, positive.

In sum, a number of studies have demonstrated that shadowing is a potentially useful activity for learners who want to improve their pronunciation. With the ubiquity of mobile technology, it is easy and affordable to implement shadowing in or outside of a classroom setting. However, there is little research addressing whether shadowing can lead to changes that can be perceived by non-expert human raters. Further, if learners are able to improve in their ability to shadow enough to be detectable by human raters, will those changes also manifest in in improved ratings over accentedness, comprehensibility, and fluency in extemporaneous speech? In addition, in order for shadowing to be recommended for learning outside of a classroom or lab setting, it is important to understand how learners perceive the activities.

The research questions were as follows:

1) Can regular individual practice using shadowing with mobile technology improve ratings of advanced L2 English speakers' (a) ability to imitate a speech models and (b) comprehensibility, accentedness, and fluency in extemporaneous speech?

2) How do advanced L2 learners feel about doing shadowing activities with mobile technology in their own time to improve their pronunciation and other language skills?

Method

Participants

Twenty-two L2 speakers of English were recruited from an English-medium university in Montreal, Canada. Of the original 22, 16 (male = 7, female = 9) remained for the duration of the study. Their ages ranged from 18 to 38 (*M* = 25.55, *SD* = 7.13), and they came from five different L1 backgrounds (Chinese = 10, French = 3, Arabic = 1, Bengali = 1, Russian = 1). All of the participants had a high enough level of English to gain admittance to credit programs at the university. Their time in Canada ranged from one month to 60 months, with the exception of one Francophone speaker who was born and raised in Montreal but spoke English as second language. Many of the participants were enrolled in academic ESL credit classes at the university. Some of these courses focused primarily on reading, writing, grammar, and vocabulary in academic settings. Others were enrolled in academic oral skills classes. Overall, the participants reported spending 25-60% (*M* = 53.44, *SD* = 13.90) of their time in Montreal using English as opposed to another language. Table 9 summarizes the background information of the L2 speaking participants.

Table 9

The L2 Speakers' Background Characteristics

Participant	L1	Age	Gender	Months in Canada	Enrolled in ESL Class[a]	Use of English (%)
1	French	28	F	11	Oral	60
2	French	34	M	whole life	Oral	25
3	Chinese	20	F	4	Writ	70
4	Chinese	19	M	4	Writ	50
5	Chinese	30	M	12	Oral	50
6	French	22	F	23	Oral	60
7	Bengali	38	M	25	Writ	70
8	Chinese	20	M	4	Both	50
9	Chinese	21	F	1	Writ	40
10	Russian	32	F	17	Oral	50
11	Arabic	38	F	60	None	30
12	Chinese	23	M	40	Writ	75
13	Chinese	19	F	4	Writ	60
14	Chinese	19	F	5	Both	50
15	Chinese	18	F	4	None	65
16	Chinese	20	M	1	Writ	50

Note. [a] *Oral* indicates that the participant was enrolled in an academic oral skills ESL class at the time of the study, while "Writ" indicates that the participant was enrolled in an academic ESL class that did not focus on oral skills.

Twenty-two L1 English listeners (male = 6, female = 16) were recruited from a Canadian university in Alberta, Canada. Their ages ranged from 20 to 41 years of age (*M* = 25.60, *SD* = 6.05). All reported having normal hearing. Of the twenty-two speakers, six reported speaking an L2 fluently (French = 3, Chinese = 1, Japanese = 1, French and Chinese = 1). Two of the speakers had taken linguistics classes. All participants in this study were paid for their participation.

Materials for Shadowing

An iPod with a wall charger and headphones was set up for each L2 speaker in the study. The iPods were loaded with eight audio dialogues, one for each week of the study. The dialogues were taken from popular television sitcoms including *Friends*, *The Big Bang Theory*, *How I Met Your Mother*, *Raising Hope*, and *New Girl*. Each of the dialogues had only two speakers, and all were close to one minute in length. All of the videos were available to watch on YouTube, so that participants would have the option of viewing the scenes as well as listening to them. The iPods had a free app installed on them called *Multi Track*[1] and e-mail accounts set up on the iPods that the participants could use for the duration of the study. This app was originally designed for musicians, but it allowed the participants to easily load dialogues from the iTunes library onto the app, and enabled learners to listen to the recordings using ear buds while simultaneously recording themselves shadowing the dialogue. This gave users the ability to listen to their recordings overtop of the original speakers, or in isolation. The app allowed for easy saving and emailing of recordings.

A booklet was also created for each participant, which included all instructions and information needed to complete the study, including clear instructions on how to use the app. The scripts were included for each dialogue as well as a URL for each of the corresponding YouTube videos. The participants received both paper and electronic versions of the booklet.

[1] A full description of the app can be found at https://itunes.apple.com/ca/app/multi-track-song-recorder/id390599090?mt=8

Testing Instruments

Language assessments

Two different language tests were given at each testing time. One was a picture narrative task, commonly referred to as The Suitcase Narrative (see Appendix D). For this task, the participants were asked to look at a serious of pictures, and when they were comfortable and familiar with the story they were asked to tell the story in the pictures. This task was developed by Derwing, Munro, Thomson, and Rossiter (2009) and is commonly used in pronunciation research (e.g., Derwing et al., 2014; Isaacs & Trofimovich, 2012). The second task was a shadowing dialogue similar to the types of dialogues used for practice throughout the study (see Appendix E). For the shadowing test, the participants were given an iPod with the dialogue loaded onto it and were also given a laptop with the YouTube video loaded and ready to play if they wished to watch it. Each participant was given 10 minutes to practice the dialogue, during which time, the researcher left the room. They were instructed to only shadow one of the speakers in the dialogue. After 10 minutes, the researcher returned and the participants were audio-recorded doing the shadowing dialogue they had just practiced.

Interviews

In-person interviews were conducted at each testing point. The interviews included Likert-scale and open-ended questions. The first interview included questions about the participants' language use and views on pronunciation. The second and third interviews asked for the participants' opinions about the shadowing activities also asked for suggestions for improvements. The third interview was of primary interest, as the participants participated in this interview after completing all of the shadowing activities. The questions used in this interview can be found in Appendix F.

Procedure

L2 speakers

The L2 participants met with the researcher individually three times (at the start and end of the study and during week six). The first sessions lasted for about an hour each, though some were longer if the participants had difficulty with the iPods or gave longer than usual responses during their interviews. At the first session, participants were given training on how to use their iPod and clear instructions about what was expected of them throughout the course of the study. They practiced shadowing and using the app to e-mail recordings of the practice attempts. They then completed their first interview and the two pretests.

Every week, for eight weeks, the participants practiced shadowing using the dialogue assigned for that week. As part of joining the study, they were asked to commit to practicing at least four times per week, for at least 10 minutes each time, though more practice was allowed and encouraged. To ensure that participants practiced at least four times per week, they were asked to submit a sample recording via email each time they practiced. They were also required to email a report each week outlining how long they practiced at each session. Each new practice week started on a Monday and if no audio files had been e-mailed by the following Friday, the participants were e-mailed a reminder. They also received a reminder if their weekly reports were not submitted by Tuesday morning of the following week. In order to be invited to the second and third sessions, the participants had to keep up with sending in their audio files and reports (though small lapses such as missing one audio file in a week, were overlooked). Each Tuesday, the participants would either receive an e-mail thanking them for sending everything in, or an e-mail reminding them of what they were missing. If a technical problem arose that could not be solved via e-mail, the researcher would meet with the participant as soon as possible and either fix the problem or issue the participant a new iPod.

The participants were not given strict rules about how to practice shadowing apart from having shadowing explained and demonstrated for them and being told they must spend at least some of the practice time shadowing, and some of the time listening to their own recordings. This choice to have a less rigid practice structure was made in order to uncover how learners like to practice when they aren't being given strict instructions. This also simulates more accurately how this activity would be used outside of an experiment. Along with the language assessments and interviews, numerous tests of individual differences were also administered both at the second and third sessions, but these are not being reported as part of this study. In the second session, the participants also completed a language background questionnaire.

L1 listeners

The L1 listeners were each given individual appointments with the researcher. These appointments lasted about 90 minutes each. Language background questionnaires were administered. Then suitcase narratives from all three sessions were played for the speakers on a computer using a computer-based rating program, created in MATLAB, developed by Saito, Trofimovich, and Isaacs (2014). Following the conventions of previous studies (e.g., Derwing & Munro, 2013; Derwing, Munro, & Thomson, 2008) the first 20 seconds of the narratives were extracted for the ratings, and initial dysfluencies were removed. The MATLAB program uses a sliding scale for the ratings and the placement of the slider results in a score between 1 and 1000. The raters then listened to each speech sample, and rated it for accentedness, comprehensibility, and fluency. They were given explanations of the meaning of all three constructs. When these ratings were completed, the listeners were asked to rate the shadowing dialogues from all three times with instructions to rate how well the L2 speakers were able to imitate the speech model. To ensure that the raters were very familiar with the speech sample, and to give them a sense of the difficulty of the task, they were trained on the iPods, and given eight minutes to practice the test dialogue before they began rating. Intraclass correlation coefficients were used to calculate the inter-rater reliability of the listeners. A high degree of agreement was found for all four measures: shadowing ($\alpha = .86$), accentedness ($\alpha = .91$), comprehensibility ($\alpha = .89$), and fluency ($\alpha = .93$).

Analysis

Speech Measures

The ratings from the L2 listeners for the pre-, mid-, and post-tests were all analyzed using a one way repeated-measures ANOVA. In cases where the ANOVA was significant, post hoc tests were run.

Interview data

Responses from the interview data were analyzed based on the nature of the questions. Likert-scale questions were asked at the midpoint and again at the end, asking participants to rate on a 9-point scale, how much they enjoyed the shadowing activities and how much they thought the activities helped improve their pronunciation. Ranges and average scores were calculated for these questions. Other questions had short answers, and for these, results were tabulated based on responses. Questions with longer detailed responses were analyzed using emergent coding and illustrative quotes were extracted.

Results

Speech Measures

The first research question asked whether the shadowing activity would improve learners' ability to imitate a speech model, and further, whether it could improve accentedness, comprehensibility, and fluency in extemporaneous speaking tasks. The means and standard deviations of the ratings of the L2 participants can be seen in Table 10.

Table 10

Means and Standard Deviations for the Four Tasks at Time 1, Time 2, and Time 3

	Time 1		Time 2		Time 3	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*
Shadowing	439.47	132.66	494.79	140.20	488.90	132.66
Accentedness	473.75	66.36	454.10	102.29	447.13	102.26
Comprehensibility	653.05	116.71	664.62	133.22	682.69	117.67
Fluency	505.38	122.67	510.55	123.50	548.68	127.24

Note. Scores are based on 1-1000 ratings.

The speakers showed overall improvement on all measures apart from accentedness. In order to see whether these changes were significant, one-way repeated measures ANOVAs were run on the each of the four variables with alpha for significance testing set at .0125 due to the large number of tests run (see Table 11). None of the variables violated Mauchly's test so the results assume sphericity. Overall, the participants showed significant improvement on all measures apart from accentedness, where the scores were slightly lower, albeit non-significantly.

Table 11

One-way Way Repeated-Measures ANOVAs for the Four Tasks

Measure	***df***	***F***	***Sig.***	***partial η2***
Shadowing	2	14.26	.0001*	.501
Accentedness	2	3.30	.05	.136
Comprehensibility	2	5.00	.01*	.192
Fluency	2	8.42	.0001*	.286

Note. *p < .0125

Because the ANOVAs for shadowing, comprehensibility, and fluency were significant, post hoc tests were conducted with a Bonferroni adjustment. For shadowing there was a significant improvement from Time 1 to Time 2 ($p < .0001$) and Time 1 to Time 3 ($p < .002$) but there was not a significant change from Time 2 to Time 3. Post hoc comparisons of the comprehensibility scores showed significant improvement from Time 1 to Time 3 ($p < .002$) but not from Time 1 to Time 2, or Time 2 to Time 3. Fluency scores showed significant improvement from Time 1 to Time 3 ($p < .006$) and Time 2 to Time 3 ($p < .013$) but not for Time 1 to Time 2.

Interviews

Overall Opinions

In order to get an overall sense of how the participants felt about the shadowing activities, at the Time 2 and Time 3 interviews, the participants were asked to rate their overall enjoyment of the activities and their perceptions about the effectiveness of these activities. These rating were completed using 9-point Likert scales, with 1 being the lowest, and 9 being the highest rating. The results can be seen in Table 12.

Table 12

Overall Opinions about Shadowing Activities

Questions	Time 2	Time 3
How much do you like the shadowing activity?	7.50 (4-9)	7.63 (6-9)
How much do you think it is helping your pronunciation?	6.81 (4-9)	7.5 (6-9)

Similar open-ended questions were also asked to get a more nuanced sense of participants' opinions about shadowing. These responses are taken from the third interview, when participants had completed all of the required activities. When asked, "How did you find the shadowing project overall?" all but one of the comments were generally positive. The negative comment indicated that the activity could be tedious.

> *If I could do it with somebody for me it would be more, how do we say it, more interesting... yeah it was not bad. We had the challenge of, you know, doing this four times. Sometimes I don't feel like doing it, I say, okay, I've got to do it.* (Part 2)

There was also one participant who noted that at the start of the project she didn't believe the activities worked, but as time passed she felt they were improving her fluency. One other indicated that although the activities were helpful, he still needed more practice. Participants gave several reasons as to why they felt positive about the shadowing activities. Some participants liked the obligatory nature of the practice activities because it pushed them improve. This can be seen in the comment below:

> *I think it's useful... we have to you know, it's like a job. When you accept a task it's like a job and it force you to practice it like every week so it will keep you on the, uh, it's like accent. You can perfect your accent and sound and like the speed and you can also learn some interesting story from the dialogue so, yes, I find it's good.* (Part 3)

Others noted specific skills that they felt had improved, the relatively small amount of time required to complete the activities, and having access to authentic materials.

The participants were also asked "Do you think this is an effective technique for improving pronunciation?" All participants answered positively and some offered reasons for why they found instruction effective. One participant indicated that he found the activity more useful than traditional classroom instruction.

> *I mean compared to the class, the classical class I'm taking, I mean the university courses you know, always testing the rule… but it's not the right thing to do when you really want to change the student, the real way of speaking you know.* (Part 5)

Others indicated that it was useful in the absence of a speaking partner. For example, Part 10 stated, "*Yes, it's um, especially it useful when there is no one who can shadow you. So the recording device would serve as a partner.*"

When asked about whether they believed their own pronunciation had improved, 10 of the participants gave positive responses. However, one did not believe he had improved:

> *I won't say a lot but I think my French background will always stays. I think the better way for me, I think to be really Anglophone Anglophone is like to leave Quebec, live at New York for one year and be really obliged to speak English and I cannot speak French.* (Part 2)

Five others gave qualified responses. For example, Participant 14 wasn't sure how much she had improved: *"Uh, yes it improved, but I don't know how good I improve."* Another hoped she had improved but wasn't sure, and one thought he had improved a little bit. Two participants indicated that they thought a large improvement would take longer than the eight weeks of the project.

> *Uh, I think so. But I want be better. Yeah sometimes still some Chinese accent so I want like you, like the you know native speaker, but it requires a long time.* (Part 4)

> *Yes, little bit but it's not too long time so if I go on and practice I think I will improve more.* (Part 13)

Among the participants who did feel positive about their improvement, one noted that she felt she could better express her emotions though her pronunciation, and one mentioned that she felt more confident talking with native speakers.

The participants were also asked if there was anything they thought would improve the shadowing activities. The responses to this question are summarized in Table 13.

Table 13

Suggested Improvements for the Shadowing Activity

Improvement	Frequency of comments
Getting feedback	2
Practicing with others	2
More intensive practice	2
Technical improvements	4
Content improvements	4

The most common responses related to technical improvements and content improvements. The technical comments included issues with glitches with e-mail and YouTube videos, as well as suggestions for improvements to the app such as a pause function. The comments about content included requests for more varied types of content (e.g., stories as well as dialogues), for participants to have a range of materials from which to choose, for more difficult dialogues, and for dialogues that slowly increased in difficulty as the study progressed.

Finally, the participants were asked whether they would recommend the shadowing activities to friends who wanted to improve their pronunciation. All of the participants said that they would and two mentioned that they already had. One of the participants said that she had recommended it to her friends in China, and another said that she planned to continue with shadowing when she returned to China and would no longer have access to L1-speaking interlocutors.

Discussion

This study investigated whether using mobile technology for individual practice with shadowing could improve advanced English learners' ability to imitate speech models and also improve their comprehensibility, accentedness, and fluency when completing an extemporaneous speaking task. The results indicated that that the participants' significantly improved in their ability to imitate a speech model, and also improved in terms of comprehensibility and fluency. However, accentedness did not improve. The study also investigated how language learners felt about doing shadowing activities in their own time to improve their pronunciation. Overall, the responses indicated that participants were positive about the shadowing activities, and saw them as an effective way to improve their pronunciation.

Speech Ratings

As was discussed earlier, when judging the utility of pronunciation training techniques, it is important to know whether changes in pronunciation after an intervention are such that they can actually make a difference in how a learner's speech is perceived by other listeners. For this reason, human raters rather than linguistic measures were used to assess whether the participants' ability to imitate the speech model improved. The ratings showed that the changes in the participants' speech were noticeable to untrained listeners. More importantly, this improvement was also born out in the comprehensibility ratings of the extemporaneous speaking task. This indicates that improvements in pronunciation after completing shadowing activities meet the gold standard of making the participants' speech easier to understand. Further, the improvements in the participants' fluency ratings indicate that shadowing may carry a potential additional benefit to learners in helping them improve their fluency. Combined, these results indicate that shadowing shows a great deal of promise for helping learners improve their L2 speech.

The one measure that did not show improvement was accentedness. This finding reinforces the evidence demonstrating that accentedness is only partially related to comprehensibility and echoes similar findings from a pronunciation intervention by Derwing et al. (2014) which found that the comprehensibility ratings of their participants improved after a pronunciation course, but that accentedness scores actually worsened significantly when completing the same suitcase narrative task used in this study. However, in the Derwing et al. study, two different extemporaneous speaking tasks were used and accentedness did improve on the other task. It is well documented that listeners are extremely sensitive to accent, with Flege (1984) finding that participants were able to detect a foreign accent in speech even when hearing only a part of one /t/ segment. It may be that very small shifts in pronunciation during a given utterance can lead to differences that are detectable to listeners, but unimportant to having speech that is easy to understand. Given that the participants did show improvement in comprehensibility, a lack of improvement in accent is not of primary concern unless one subscribes to the *nativeness principle* in pronunciation instruction (Levis, 2005).

Learner Opinions About Shadowing

This study strove to have high ecological validity by asking learners to complete the activities in their own time, and, for the most part, to practice how they liked. However, the participants were held accountable to a minimum amount of practice by their weekly reports and by the four speech samples they e-mailed the researcher each week. Further, the participants were paid for their participation, and self-selected to do the study. For learners studying on their own, the only accountability they will have is to themselves. If learners don't like shadowing, or don't believe it is effective, they are unlikely do it for very long. Further, an instructor considering asking learners to spend a significant amount of time outside of class on shadowing activities would want to know that learners are likely to find these activities reasonably enjoyable and beneficial. Overall, the participants in this study were very positive about the shadowing activities. All of the participants believed that it was an effective technique for improving pronunciation, though not all were confident that they had made large amounts of improvement themselves. While some participants mentioned that they would have found the activities more enjoyable if they were able to work with other people, it was also noted that the activities were an effective way to get speaking practice when speaking partners were not available.

It is also interesting that the Likert scale ratings for how much the participants enjoyed the activity, and for how effective they thought it was, were higher at the end of the study than they were partway through. This suggests that despite Bovee and Stewart's (2009) claim that shadowing is seen as "meaningless parrot-like practice", learners actually tend to appreciate it more when sticking with it over a longer period of time. This finding is echoed in Ding's (2007) findings from interviews with highly successful language learners who, when talking about imitation-based activities, noted that they "had been initially forced to use these methods but gradually came to appreciate them" (p. 272). This is not to say the participants initially disliked the activities, but rather that their appreciation of the activities and their belief in their efficacy increased over time rather than waning.

Implications for Instruction

Shadowing activities with mobile technology may offer a valuable tool for learners looking for ways to get improved pronunciation on their own, or for instructors hoping to provide additional to support to language learners who struggle with pronunciation. However, it should be noted that while the activities used in this study show promise, they should not be considered as a replacement for more traditional classroom-based pronunciation instruction. While shadowing may help learners become more comprehensible and fluent in their L2, there is not sufficient evidence to suggest that shadowing alone can help learners improve all aspects of speech that may impact comprehensibility.

If choosing to use shadowing activities with learners, it is important to give careful consideration to the models used for shadowing. While many participants liked using sitcom dialogues because of their idiomatic language, others would have preferred different content. This study had learners use the same dialogues in order to control content as a variable in the study; however, in practice it may make sense to allow learners to choose their own speech models based on their language learning goals, or to choose speech models that are appropriate for specific learner groups.

Limitations and Suggestions for Future Research

This study had several limitations that could be addressed in future research. First, due to a lack of a control group, the results must be interpreted with caution. Many of the learners were enrolled in ESL classes during the course of the study, and all had exposure to English outside of class. Further, while participants were asked to email in examples of their shadowing practice each week, the amount of time learners spent practicing was based on self-report, and as such it is not possible to be certain that all of the participants practiced as much as they claimed. Future studies could investigate a wider range of learner variables, including proficiency level.

The use of recorders with shadowing would also be a variable that could be explored in greater detail in future studies. Shadowing does not by definition require audio recordings, and in practice, using recorders with shadowing is common, but not universal. Mobile technology makes recording easy, and may help learners notice their own errors, thus facilitating change. Certainly using learner recordings is commonly advocated as a method for pronunciation instruction (e.g., Walker, 2005). However, this study did not compare shadowing with and without recorders so it is impossible to know how much the use of recordings impacted the efficacy of shadowing. Repetition is another aspect of shadowing that warrants further investigation as shadowing does not require learners to repeatedly practice the same material, and it would be interesting to see if repetition plays a facilitating role in improvement.

Conclusion

This study demonstrated that shadowing shows promise as a way to help learners improve their pronunciation and fluency. The use of portable mobile technology means that this technique may be a practical solution for learners who do not have access to pronunciation instruction, and instructors who are looking to help students who need extra help with pronunciation. Due to its repetitive nature, there may be a reluctance to recommend this activity to learners, regardless of its efficacy. However, interviews with participants indicate that learners enjoy shadowing and see it as an effective way to improve their pronunciation. The findings of this study are in line with other studies that have investigated shadowing as a technique for pronunciation improvement; this study extends these findings by demonstrating that these improvements can be detected by untrained listeners, and lead to improved comprehensibility of extemporaneous speech making this activity of potentially high utility for leaners who want to communicate more effectively.

Chapter 5: Conclusion

Introduction

This dissertation represents an attempt to increase our understanding of speech perception and pronunciation instruction given the complexity of language use in the world today. While there is an increasing amount of research being done in these areas of applied linguistics research, it is fair to say that we are far from having a complete understanding of the role of second language (L2) users' first language (L1) background in their perception of L2 speech, particularly when considering the range of potential interactional contexts for L2 speakers communicating in a shared L2 (e.g., English) with other L2 speakers. Further, there is a need for more research on how best to help individuals improve their pronunciation to be able to communicate successfully with a wide range of interlocutors. Each study in this dissertation had its own specific goals, and each was designed to be able to stand alone. However, together, these studies address two primary objectives: (a) to increase our understanding of how L2 users of English perceive English speech, and (b) to understand how L2 learners can be helped to better perceive differences between their own speech and their target language in order to facilitate improvements in pronunciation. While each study featured its own discussion and conclusion, in this chapter, I give a brief summary of the three studies, connecting each one and focusing on the key findings from each. I then draw conclusions from the studies, discussing what the three studies say about speech perception and pronunciation pedagogy when considered together. This is followed by pedagogical implications, limitations of the studies, suggestions for future research, and concluding thoughts.

Overview of Key Findings

Study 1 and Study 2 were primarily concerned with the first objective of this dissertation: understanding how L2 learners perceive L2 English speech. The specific objective of Study 1 was to uncover what underlies judgments of L2 speech, particularly for L2 speakers (though L1 speakers were included for comparison purposes). This study sheds light on which aspects of speech L2 listeners attend to when listening to other L2 speakers, with the idea that aspects of speech which are salient to listeners are more likely to be noticed through interaction and thus potentially more amenable to improvement without instructional intervention. Two groups of listeners – 15 L2 English listeners, and 10 L1 English listeners – judged speech samples from a group of mixed L2 speakers from a variety of L1 backgrounds. The speakers completed a read aloud task and a spontaneous speaking task which were evaluated in two separate rating blocks using a paired comparison method in which listeners judged how dissimilar each sample was from every other speech sample. Multidimensional scaling (MDS) was used in order represent the speakers in equidistant Euclidean space, based on the listeners' judgments. The coordinates of the speakers' positions in the MDS solution were then correlated with a wide range of speech variables to see which aspects of speech could explain the ratings. The key findings from Study 1 were that (a) the L2 listeners were sensitive to global aspects of speech, such as overall speaking proficiency as well as segmental errors and fluency measures, and (b) while L2 and L1 listeners had some similarities in what aspects of speech they attended to, especially in relation to fluency, there was little correspondence in how the two groups judged the speech samples, particularly for the read aloud task. This suggests that the L1 and L2 listeners approached the rating tasks in different ways.

The differences between the L1 and L2 listener groups in Study 1 suggest that language background may be a source of important differences in how listeners perceive L2 speech based on their language backgrounds. However, Study 1 treated all L2 listeners (and all L2 speakers) as a single group, making it impossible to know whether, or how, the specific L1 backgrounds of the L2 listeners and speakers may have impacted their perceptions. Study 2 extended the findings of Study 1 by addressing the issue of whether there are differences in how language background impacts judgments of L2 speech from different language groups. The objectives of Study 2 were to understand whether there were differences in which speech characteristics were associated with comprehensibility judgments for different L1 listener-speaker groups and, further, whether L1 background impacted the judgments of the comprehensibility of L2 speech beyond what could be explained by characteristics of the speech itself. In addition, Study 2 sought to discover whether the listeners themselves believed that L1 was an important factor when making comprehensibility judgments. Forty English listeners with different L1 backgrounds (10 each from French, Mandarin, Hindi, and English groups) rated the comprehensibility of 30 L2 speakers from the same three L2 groups. After making each rating, the listeners gave a verbal report in which they explained the reasons for their ratings. The listeners then completed a second rating task, judging the same speech samples but this time for common aspects of pronunciation (segmental errors, word stress, intonation, and speech rate). Correlations were used to investigate the relationship between the speech measures and comprehensibility ratings for each listener-speaker group. Next, a hierarchical regression was run for each L2 listener group with comprehensibility as the outcome variable. The four speech variables were entered as predictor variables in Step 1 and language background was entered in Step 2. The key findings from Study 2 were that (a) different speech characteristics were associated with the comprehensibility ratings of the speakers for each listener-speaker combination, (b) language background explained the variance of speech ratings above and beyond what could be explained by specific speech characteristics only in the case of the Mandarin listeners, who downgraded the Hindi and French speakers in relation to the speakers who shared their L1 (and who were also the lowest proficiency speakers), and (c) listeners' verbal reports showed only a moderate correspondence to their actual rating behaviour.

Together, Study 1 and Study 2 suggest that L2 listeners' perceptions of L2 speech may differ both from the judgments of L1 listeners and other L2 listeners with different L1 backgrounds. Study 1 also found that L2 listeners may be aware that their speech differs from the speech of other speakers, but not have a clear understanding of how it is different. Study 2 suggests that the aspects of speech most important for L2 learners to develop in order to be comprehensible to their interlocutors may vary depending on the contexts in which the L2 user will be speaking English. This makes it very difficult to find activities that can be of use to learners who come from different L1s and who will be interacting with interlocutors from a range of L1 backgrounds. Study 3 tested the effectiveness of shadowing, a pronunciation training technique that offers a possible solution to this problem The objectives of Study 3 were to investigate whether shadowing using iPods with voice recorders could help L2 learners from a variety of L1 backgrounds improve not only in their ability to imitate a speech model but also in their accentedness, their fluency, and most importantly, their comprehensibility. A final objective was to explore whether learners found shadowing to be an effective way to improve pronunciation. Sixteen L2 speakers from mixed L1 backgrounds completed shadowing activities with iPods for eight weeks. Pre-, mid-, and post-tests were conducted and the speech samples were rated for the four above-mentioned variables by 21 L1 English speakers and analyzed using one-way repeated-measures ANOVAs. Interviews were conducted to uncover learners' opinions about the efficacy of the activities. The key findings from Study 3 were that (a) the learners improved in all measures apart from accentedness and (b) they were generally positive about using shadowing to improve their pronunciation.

Drawing Conclusions from the Three Studies

The L2 Interlocutor in Pronunciation Research

The findings from the first two studies highlight the importance of considering L2 users not only as speakers who need help to become intelligible to listeners, but also as potential target listeners. Both Study 1 and Study 2 found some differences in how listeners perceived English L2 speech based on their language backgrounds. This appears to run somewhat counter to the findings such as those reported by Munro, Derwing, and Morton (2006) who noted that "there is a notable degree of shared experience when listeners from diverse language backgrounds hear L2 speakers' utterances" (p. 127). However, the findings here do not indicate that there is little commonality in how listeners perceive L2 speech. In fact, to think that there can be no shared experience in how two speakers, regardless of background characteristics, hear the same utterance is hard to believe. As with Munro et al.'s study, Study 2 found an overall very high inter-rater reliability for the listener groups. An effect of L1 above and beyond what could be explained by the speech characteristics was only found in the case of the Mandarin listeners evaluating the Mandarin speakers, and this may be explained by their overall (low) proficiency. However, Study 1 and Study 2 do suggest that the ways in which listeners from different language backgrounds perceive L2 speech are not identical. These findings are important to bear in mind when generalizing findings from studies that only use listeners from a specific L1.

The Role of Fluency

One commonality in the results of Study 1 and Study 2 was the prevalence of fluency measures associated with the speech judgments. Fluency was the most common speech characteristic that explained the MDS models for both L1 and L2 listeners across the two tasks in Study 1. Speech rate was also associated with comprehensibility ratings of the French and Hindi speakers in Study 2. Studies investigating comprehensibility with L1 listeners have found comprehensibility and fluency to be related (e.g., Derwing, Rossiter, Munro, & Thomson, 2004; Isaacs & Trofimovich, 2012). Isaacs and Trofimovich's study was a comprehensive mixed-methods investigation of which speech characteristics best distinguish the comprehensibility level of learners. The authors correlated L1 English listeners' comprehensibility ratings of L2 speech samples with a wide range of speech measures and triangulated the results using English instructors' introspective reports of the same speech samples. They found that fluency was one of the five constructs useful for the assessment of L2 comprehensibility. Fluency as a concept is quite broad and there are many speech characteristics that contribute to it. When discussing fluency in terms of the results of Studies 1, 2, and 3, the term is being used very broadly as it was measured differently in different studies (word type and token production, frequency of pausing, and sample duration in Study 1; listener judgments of speaking rate in Study 2; and listener judgments of fluency in Study 3). However, there is an indication that fluency measures used in this broader sense play a consistent role in speech perception for a wide range of listeners.

In light of the importance of fluency measures in the results of Studies 1 and 2, it is encouraging that the fluency ratings of the participants' speech improved as a result of the shadowing activities in Study 3. Because pronunciation is considered to be a part of fluency, broadly defined (Levis, 2006), and comprehensibility and fluency are related, it is not surprising to see improvements in fluency when there are improvements in comprehensibility and vice versa. Further, based on the results of Study 1, fluency is a salient speech characteristic for both L1 and L2 listeners, which likely makes it more amenable to improvement. However, it is interesting to note that Derwing, Munro, Foote, Waugh, and Fleming (2014) and Derwing, Munro, and Foote (2015) found that in a pronunciation intervention study which used the same type of measurements for comprehensibility and fluency as Study 3 (i.e. listener-based ratings with similarly defined constructs), fluency measures did not improve despite significant improvements in comprehensibility, suggesting that not all pronunciation interventions for different learners across various contexts will come hand-in-hand with fluency improvements.

Shadowing Interventions

Study 1 highlights the need for pronunciation interventions that can help learners better perceive their own speech errors; L2 listeners tended to pay attention to global aspects of speech, segmentals, and temporal characteristics of speech. However, other speech properties, including word stress – which L1 listeners attended to in both tasks – did not appear to be as salient to L2 listeners. Study 2 suggested that there are some differences in how listeners perceive speech based on L1 background. These results, taken together, suggest that there is a need for pronunciation interventions that can be used to focus on different pronunciation issues for learners and for techniques that enable learners to better perceive how their own speech differs from that in the their target language. Shadowing, as it was operationalized in this study, shows potential as a technique that can work in a wide range of contexts and with a wide range of learners. Shadowing allows learners the opportunity to focus on their own productions in greater detail, both due to the repetitive nature of the practice, and because of the use of portable voice recorders. Shadowing also allows for the possibility of choosing different speech models for learners from different language backgrounds. In Study 3, all of the target speech models chosen were from L1 speakers. This choice was made largely because it was the first time shadowing was investigated in this way, and keeping constant the L1 of the shadowing dialogues and the listeners who rated the pre-, mid-, and post-tests allowed for a greater chance of measuring the effects of the intervention in a more controlled way. However, there is no reason that learners couldn't choose their own targets for shadowing based on their individual needs and communicative goals; this might include the speech by other L2 users.

Pedagogical Implications

It would be very gratifying indeed if one study (or even three) could tell pronunciation instructors exactly what they should focus on in the classroom and how they should teach it. However, while these studies fall short of that mark, there are some pedagogical implications that these studies can provide for pronunciation teachers. Findings from Study 1 suggest that learners may not always notice exactly how their speech differs from that of their target language, a finding that is not surprising in light of Derwing's (2003) survey of 100 language learners in Canada, which found that many learners were unaware of what their pronunciation problems were, and that of the problems identified, only 11% were suprasegmental aspects of pronunciation. This highlights the need for explicit pronunciation instruction to help learners who struggle with pronunciation. Recent surveys of pronunciation studies indicate that pronunciation instruction can lead to improvements (Lee, Jang, & Plonsky, 2014; Thomson & Derwing, 2014). In fact, despite some instructors being skeptical about the effectiveness of pronunciation instruction (Foote, Holtby, & Derwing, 2012), in a meta-analysis of the efficacy of pronunciation instruction, Lee et al. (2014) found a very high effect size ($d = .89$) for the impact of pronunciation instruction.

It is promising that shadowing helped learners become more comprehensible and fluent in the absence of explicit pronunciation instruction. However, Study 3 did not test improvements in specific aspects of pronunciation so it is unknown in exactly what way the learners improved. Given the findings of Study 1, instructors who would like to use shadowing with their learners may wish to include explicit instruction and feedback as part of shadowing activities. For example, when using shadowing activities as part of a different pronunciation intervention study (see Derwing et al., 2014), I would sometimes ask learners to focus on specific aspects of pronunciation in the target speech samples which the learners struggled with (e.g., word stress or deletion) and would also have learners perform their shadowing dialogues in order to receive explicit feedback on their pronunciation.

Findings from Study 2 also highlight the importance of providing a variety of speech models to learners. In Study 2, the Mandarin listeners downgraded the Hindi and French speakers in comparison to other Mandarin speakers, and there was a tendency for the Hindi speakers to downgrade the Mandarin speakers. Further, the listeners, particularly French and Mandarin listeners, made many comments indicating that speakers from other L1 backgrounds were difficult to understand because of their accents. As well as helping learners be understood, language instructors need to help learners understand their interlocutors. The most obvious way to do this is to include speech from different language backgrounds in classroom activities. For instance, research indicates that increased familiarity with an accent improves comprehension (e.g., Gass & Varonis, 1984); thus, exposing learners to a variety of speech models will enable them to be successful in a wider range of speaking contexts. Increasingly, pronunciation scholars are recognizing the importance of moving beyond L1 speaker models in pronunciation classrooms (e.g., Matsuda & Friedrich, 2011; Matsuura, 2007), even in Inner Circle contexts (Murphy, 2014), arguing that with the role of English in the world today, a focus on purely L1 models has become outmoded.

There is also a wealth of evidence that accent can cause bias in listeners (see Lindemann & Subtirelu, 2013, for a review). For example, research suggests that sharing an L1 with test-takers may be a source of bias in trained language test-raters, and that test-raters are often concerned that they may have biases based on their familiarity with certain accents (e.g., Winke, Gass, & Myford, 2012, 2013). There is reason to believe that increasing learners' belief in their ability to understand speakers from different L2s may actually aid in their comprehension.

Studies investigating *reverse linguistic stereotyping* (RLS) using matched guise techniques (e.g., Rubin, 1992, 2002) have found that when listeners believe that a person's speech will be difficult to understand, they actually understand less, even when the speech is identical. RLS studies usually control for speech and manipulate the image a listener has of who is speaking. For example, Rubin (1992) had students listen to identical lectures spoken by an L1 speaker of English, but had some participants see an image of a Caucasian woman, while others saw an Asian woman. He found that participants understood less when they believed the speaker was Asian. Similar results were found in a study of L2 English speakers in China who heard an L1 English speaker but in some cases believed it was spoken by an L2 speaker (Hu & Su, 2010, as cited in Lindeman & Subtirelu, 2013).

It is reasonable to assume that a similar effect could occur if one recognizes an accent as belonging to a category one considers difficult. Such beliefs could have impacts beyond comprehension, as learners may avoid interaction that they fear will be difficult. In an interview with L2 speakers in Canada, Derwing (2003) found that some learners had experienced this when talking with native speakers. For example one participant noted, "They don't listen to people who have an accent," and another said, "They don't pay attention to you if your English isn't good" (p. 557). For these reasons, learners' perceptions of difficulty may impact both the understanding of speech and the likelihood of engaging in successful interaction with L2 speakers from language backgrounds speakers perceive as being difficult to understand.

Finally, it is important for instructors and programs to think about what types of interactions their learners will be having after completing their courses. A good example of this comes from my first supervisor when I began my teaching career at a high school in Japan. We were discussing the American accented speech models used in the majority of the listening materials for our high school students. He told me about a conversation he had just had with a former student who had become employed at an international company. The student told him that his high school classes had trained him to communicate with Americans but that in his job he was expected to use English all the time, and virtually never with Americans, nor with other L1 English speakers, and he wished he had been given English classes that would have prepared him for the range of L1 English speakers he interacted with regularly. Even in Inner Circle countries, learners will need to be familiar with a range of accents. In some cases, L1 models may be what students want and need the most, but in other cases, there may be good reasons to include a greater variety of speech models. It is important for instructors to be aware of what the goals of their learners are and adjust their approach accordingly.

Limitations and Suggestions for Future Research

Many of the limitations of the three studies have already been discussed in each individual chapter. However, there are some overarching limitations to the studies that should be addressed. First, despite discussing the importance of L2 listeners as raters and using L2 speech models in Study 1 and Study 2, Study 3 had neither. As was mentioned earlier, L1 listeners and speakers were used to allow for tighter control, but this doesn't change the problem that, as with many other studies, L1 speakers were used as an assessment standard in this study. Further, Study 1 treats L2 users as a homogenous group despite findings from Study 2 that suggest that there are some differences in L2 groups' judgments about L2 speech.

The findings from Study 1 and Study 2 suggest that there is a need for more pronunciation research that takes L2 interlocutors into account. When discussing the intelligibility principle in Chapter 1, I discussed Levis's (2005) two matrices – two by two when discussing native speakers versus non-native speakers, and three by three when considering Inner, Outer, and Expanding Circle contexts – for denoting the different speaker/listener interlocutor combinations that can occur in interaction. In Levis (2006), implications of the four by four matrix for intelligibility were discussed in more detail. In particular, Levis noted that Quadrant C of the matrix, which represents an L1 listener with an L2 speaker, "is the traditional domain of intelligibility research" (p. 257). There is also little doubt that much of the research focusing on L2 perceptions of English speech is focused on Quadrant B, that is, L2 listeners/L1 speakers. There is a need for more research that investigates Quadrant D, namely, L2 listeners/L2 speakers. English as a lingua franca research is often exclusively interested in this type of interaction, which is also problematic in that it does not reflect the reality of communication for many learners who frequently interact with L1 speakers. There has been an increased interest in considering the role of L2 listeners in intelligibility and comprehensibility in the past several years, especially within the ELF and World Englishes paradigms, but there is room for much more work in this area.

In terms of comprehensibility research, it would be interesting to see more investigations into how rater characteristics, including language background and proficiency, impact which aspects of speech most strongly relate to comprehensibility judgments. Study 2 indicated that there may be interesting between-group differences in which aspects of speech contribute to comprehensibility which could be explored in future research. The study presented here had a small sample size and minimal controls over the language content, making it largely exploratory in nature. However, given the importance of comprehensibility for L2 learners, more research investigating a wider range of language backgrounds and a larger number of speech variables could help provide instructors with useful information when choosing instructional foci.

Finally, there is a need for more pedagogically-grounded research that can help instructors and learners find practical solutions for improving comprehensibility in a wide range of contexts. One area of research that would be useful in this area, is finding ways to explore different approaches to helping learners better notice the gaps between their own pronunciation and that of their interlocutors. For example, the development a research-based, user-friendly self-evaluation tool for learners to assess their own pronunciation difficulties could help make activities including – but certainly not limited to – shadowing more effective. While it is gratifying that systematic reviews of pronunciation studies have found pronunciation instruction to be effective, there is still much more research needed to better understand not only what to teach, but also how to teach it.

Concluding Remarks

I began teaching English 15 years ago and struggled to help my learners improve their pronunciation. I found myself frustrated when trying to answer two very basic questions about pronunciation instruction: "What should I teach, and how should I teach it?" My quest for answers to these two questions is what propelled me to move out of the classroom and into a doctoral program. While the three studies presented here can only answer these questions in a small way, my hope is that the results of Study 1 and Study 2 demonstrate the need for further research to better understand the complexities of speech perception and intelligibility when making decisions about what to teach in the classroom. Further, it is my hope that Study 3 provides evidence for one practical technique to help pronunciation instructors grappling with the question of how to teach pronunciation.

References

Acton, W. (1984). Changing fossilized pronunciation. *TESOL Quarterly, 18*, 71-86.

Algethami, G. Ingram, J., & Nguyen, T. (2011). The interlanguage speech intelligibility benefit: The case of Arabic accented English. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference, Sept. 2010* (pp. 30-42). Ames, IA: Iowa State University.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42, 529-555.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*(4), 561-613.

Avery P. & Ehrlich, A. (1992). *Teaching American English Pronunciation*. Oxford: Oxford University Press.

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

Best, C. (1995). A direct-realist view of cross-language perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore: York Press.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer.

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

Bovee, N. & Stewart, J. (2009). The utility of shadowing. In A.M. Stoke (Ed.),*JALT 2008 Conference Proceedings*. Tokyo: JALT

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977-985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299-2310.

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation teaching practices in Canada. *TESL Canada Journal*, *19*, 51-61.

Brown, A. (1991). Introduction. In A. Brown (Ed.), *Teaching English pronunciation: A book of readings* (pp. 1-10). London: Routledge, 1-10.

Burns, A. (2006). Integrating research and professional development on pronunciation teaching in a national adult ESL program. *TESL Reporter, 39*, 34-41.

Busing, F. M. T. A., Commandeur, J. J., & Heiser, W. J. (1997). *PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints*. In W. Bandilla & F. Faulbaun (Eds.), *Softstat 97: Advances in statistical software* (pp. 237-258). Stuttgart: Lucius.

Buss, L. (2015). Beliefs and practices of Brazilian EFL teachers regarding pronunciation. *Language Teaching Research*. Advance online publication. doi: 10.1177/1362168815574145

Celce-Murcia, M., Brinton, D.M., & Goodwin, J.M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge, UK: Cambridge University Press.

Couper, G. (2003). The value of an explicit pronunciation syllabus in ESOL teaching. *Prospect, 18*, 53-70.

Couper, G. (2006). The short- and long-term effects of pronunciation instruction. *Prospect, 21*, 46-66.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (in press) Does a speaking task affect second language comprehensibility? *The Modern Language Journal.*

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2014). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*. Advance online publication. doi: 10.1002/tesq.203

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.

Dauer, R. M. (2005). The Lingua Franca Core: A new model for pronunciation instruction? *TESOL Quarterly*, *39*(3), 543-550.

de la Mora, D. M., Nespor, M., & Toro, J. M. (2103). Do humans and nonhuman animals share the grouping principles of the iambic-trochaic law? *Attention, Perception, and Psychophysics*, *75*, 92-100.

Deng, J., Holtby, A., Howden-Weaver, L., Nessim, L., Nicholas, B., Nickle, K., Pannekoek, C., Stephan, S., & Sun, M. (2009). English pronunciation research: The neglected orphan of second language acquisition studies? (WP 05-09). Edmonton, AB: Prairie Metropolis Centre.

Derwing, T. M. (2003). What do ESL students say about their accents?.*Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *59*(4), 547-567.

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

Derwing, T., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics, 22*(3), 324-337.

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). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *66*(2), 181-202.

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*, 163–185.

Derwing, T.M. & Munro, M.J. (in press). *Pronunciation Fundamentals: Evidence-based Perspectives for L2 Teaching*. John Benjamin's.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*, 526-548.

Derwing, T.M., Munro, M.J., & Foote, J.A. (2015, March). *Long-term benefits of pronunciation instruction in the workplace*. Paper presented at the annual American Association for Applied Linguistics conference, Toronto, Ontario.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*, 359-380.

Derwing, T., Munro, M., Thomson, R., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1997). Pronunciation instruction for fossilized learners. Can it help? *Applied Language Learning*, *8*(2), 217-35.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 393–410.

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

Deterding, D., & Kirkpatrick, A. (2006). Emerging South-East Asian Englishes and intelligibility. *World Englishes, 25*(3-4), 391-409.

Ding, Y. (2007). Text memorization and imitation: The practices of successful Chinese learners of English. *System*, *35*(2), 271-280.

Ellis, R. (2001). Introduction: Investigating form-focussed instruction. *Language Learning, 51* (S1), 1-46.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Ellis, R. (2005). Principles of instructed language learning. *System, 33*(2), 209-224.

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

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America, 76*, 692-707.

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). Timonium, MD: York Press.

Flege, J., & Frieda, E. (1995). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25*, 169–186.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125-3134.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2012). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada journal, 29*(1), 1-22.

Foote, J. A., & Smith, G. (2013, September). *Is there an App for that?* Paper presented at the Pronunciation in Second Language Learning and Teaching conference, Ames, Iowa.

Foote, J. A., Trofimovich, P., Collins, L., & Soler-Urzúa, F. (2013). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal*. Advance online publication. doi: 10.1080/09571736.2013.784345

Fraser, H. (2006). Helping teachers help students with pronunciation, prospect: *A Journal of Australian TESOL, 21*, 80-94.

Fraser, H. (2007). Categories and concepts in phonology: Theory and practice. In A. Schalley & D. Khlentzos (Eds.), *Mental States Vol. 2: Language and Cognitive Structure (Papers from the International Language and Cognition Conference Sept 2004)* (pp. 311-330). Amsterdam: Benjamins.

Fraser, H. (2010). Cognitive theory as a tool for teaching second language pronunciation. In S. De Knop, F. Boers, & A. De Rycker (Eds.), *Fostering language teaching efficiency through cognitive linguistics* (pp. 357-379). Berlin, Mouton de Gruyter.

Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA review, 19*(1), 3-17.

Gass, S.M. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S.M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An Introduction* (pp. 175-199). New York, NY: Routledge.

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

Gilbert, J. (1994). Intonation: A navigation guide for the listener. In J. Morley (Ed.), *Pronunciation pedagogy and theory: New ideas, new directions* (pp. 36-48). Alexandria, VA: TESOL.

Gilbert, J. (2005). *Clear speech: Pronunciation and listening comprehension in American English* (3rd ed.). New York: Cambridge University Press.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Grant, L. (2001). *Well Said: Pronunciation for clear communication* (2nd ed.). Boston: Heinle & Heinle.

Grant, L. (2010). *Well said* (3rd ed.). Boston, MA: Heinle & Heinle.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly*, *38*(2), 201-223.

Hamada, H. (2014). The effectiveness of pre- and post-shadowing in improving listening comprehension skills. *The Language Teacher, 38*, 3-10.

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing, 13*(1), 1-33.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, 29, 163-180.

Hayes-Harb, R. & Hacking, J. (2015). What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation, 1*, 43-64.

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.

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A. & Curnick, L. (2012). The English pronunciation teaching in Europe survey: selected results. *Research in Language, 10*(1), 5-27.

Hewings, M. (2004). *Pronunciation practice activities*. Cambridge: Cambridge University Press.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL*, *15*(1), 3-20.

Hiramatsu, S. (2000). A differentiated/integrated approach to shadowing and repeating. Retrieved August 2013 from Google Scholar.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 93-103.

Hsieh, K-T., Dong, D-A., & Wang, L-Y. (2013) A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics, 11*, 43-66.

Imai, S., Walley, A.C., & Flege, J.E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *Journal of the Acoustical Society of America, 117*, 896–907.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(4), 555-580.

Isaacs, T. (2009). Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal, 27*, 1-12.

Isaacs, T., & Trofimovich, P. (2011). International students at Canadian universities: Validating a pedagogically-oriented pronunciation scale. Unpublished corpus of second language speech.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34*(3), 475-505.

Jakeman, V. & McDowell. (2008). *New insight into IELTS: Student's book with answers*. Cambridge: Cambridge University Press.

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, *5*, 1-10.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals*. Oxford: 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. (2006). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly*, *40*(1), 157-181.

Jenkins, J. (2012). English as a Lingua Franca from the classroom to the classroom. *ELT Journal*, *66*(4), 486-494.

Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. *Language Teaching*, *44*(3), 281-315.

Jun, H. G. & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University, Sept. 2009 (pp. 53-66). Ames, IA: Iowa State University.

Kachru, B. B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics*, *17*, 66-87.

Kachru, B.B. (1982). Meaning in deviation. In B. B. Kachru (Ed.), *The other tongue: English across cultures (2nd ed.)* (pp. 301-326). Urbana, IL: University of Illinois Press

Kachru, B.B. (1992). The second diaspora of English. In T. W. Machan and C. T. Scott (Eds.), *English in its social contexts: Essays in historical sociolinguistics* (pp. 230-252). New York: Oxford University Press.

Kachru, B.B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics*, *17*, 66-87.

Kachru, B., Kachru, Y., & Nelson, C. (Eds.). (2006). *The handbook of world Englishes*. Malden, MA: Blackwell.

Kachru, Y., & Smith, L. E. (2008). *Cultures, contexts, and world Englishes*. New York, NY: Routledge.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*(3), 249-269.

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*, 441-456

Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554-566.

Kashiwagi, A., & Snyder, M. (2010). Speech characteristics of Japanese speakers affecting American and Japanese listener evaluations. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 10*(1), 1-14.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *64*(3), 459-489.

Kim, T. (2008). Accentedness, comprehensibility, intelligibility, and interpretability of NNESTs. *CATESOL Journal, 20*(1), 7–26.

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

Langacker, R.W. (2008). Cognitive grammar as a basis for language instruction. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 66-88). New York and London: Routledge.

Lee, J., Jang, J., & Plonsky, L. (2014). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amu040

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

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Boston: MIT Press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377.

Levis, J.M. (2006). Assessing speaking. In H. Hughes (Ed.), *Spoken English, TESOL, and applied linguistics* (pp. 245-270). New York, NY: Palgrave Macmillan

Levis, J. M., & Grant, L. (2003). Integrating pronunciation into ESL/EFL classrooms. *TESOL Journal, 12*(2), 13-19.

Lin, L. C. (2009). A study of using "shadowing" as a task in Junior High School EFL program in Taiwan. *Unpublished master's thesis, National Taiwan University of Science and Technology*, Taipei, Taiwan.

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning, 63*(3), 567-594.

Lippi-Green. R. (1997). *English with an accent: Language, ideology, and discrimination in the United States.* London: Routledge.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego: Academic Press.

Low, E.L. (2015). P*ronunciation for English as an international language: From research to practice.* New York: Routledge.

Luo, D., Shimomura, N., Minematsu, N., Yamauchi, Y., & Hirose, K. (2008). Automatic pronunciation evaluation of language learners' utterances generated through shadowing. *Interspeech 2008*, 2807-2810.

MacDonald, S. (2002). Pronunciation—Views and practices of reluctant teachers. *Prospect, 17*, 3-18.

MacIntyre, P. D. (2007). Willingness to communicate in the second language: Understanding the decision to speak as a volitional process. *The Modern Language Journal, 91*, 564–576.

MacIntyre, P. D., Dörnyei, Z., Clément, R. & Noels, K. (1998). Conceptualizing willingness to communicate in an L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal, 82*, 545-562.

MacKay, I., & Flege, J. (2004). Effects of the age of second-language (L2) learning on the duration of L1 and L2 sentences: The role of suppression. *Applied Psycholinguistics, 25*, 373-396.

Mackey, A., & Goo, J. M. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Input, interaction and corrective feedback in L2 learning* (pp. 379-452). Oxford: Oxford University Press.

Major, R. (1986). The Ontogeny model: evidence from L2 acquisition of Spanish r. *Language Learning, 36*(4). 453–504.

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*, 173-190.

Martinsen, R.A., Alvord, S.M., & Tanner, J. (2014). Perceived foreign accent: Extended stays abroad, level of instruction, and motivation. *Foreign Language Annals, 47*, 66-78.

Matsuda, A., & Friedrich, P. (2011). English as an international language: A curriculum blueprint. *World Englishes, 30*(3), 332-344

Matsuura, H. (2007). Intelligibility and individual learner differences in the EIL context. *System, 35*(3), 293-304.

Matsuura, H., Chiba, R., Mahoney, S., & Rilling, S. (2014). Accent and speech rate effects in English as a lingua franca. *System, 46*, 143-150.

Meyer, J. M., Heath, A. C., Eaves, L. J., & Chakravarti, A. (2005). Using multidimensional scaling on data from pairs of relatives to explore the dimensionality of categorical multifactorial traits. *Genetic Epidemiology*, *9*, 87-107.

Mori, Y. (2011). Shadowing with Oral Reading: Effects of Combined Training on the Improvement of Japanese EFL Learners' Prosody. *Language Education & Technology, 48*, 1-22.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation and instruction. *Studies in Second Language Acquisition, 21*, 81–108.

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*(4), 451-468.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520-531.

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*(7), 626-637.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111-131.

Murphy, J. M. (2014). Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching. *System, 42*, 258-269.

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

O'Brien, M. G. (2014). L2 Learners' Assessments of Accentedness, Fluency, and Comprehensibility of Native and Nonnative German Speech. *Language Learning, 64*(4), 715-748.

Rau, D., Chang, H. H. A., & Tarone, E. E. (2009). Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative. *Language Learning*, *59*, 581-621.

Riney, T. J., Takagi, N., & Inutsuka, K. (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly, 39*(3), 441-466.

Rongna, A., & Hayahi, R. (2012). Accuracy of Japanese pitch accent rises during and after shadowing training: *Proceedings from 6th International Conference on Speech Prosody*, (n.p.). Retrieved May 2012 from http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_98.pdf

Rosse, M. (1999). Tracking - A method for teaching prosody to ESL learners. *Prospect, 14*, 53-61.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review/La revue canadienne des langues vivantes*, 65(3), 395-412.

Rubin, D. L. (2002). Help! My professor (or doctor or boss) doesn't talk English! In J. Martin, T. Nakayama, & L. Flores (Eds.), *Readings in intercultural communication: Experiences and contexts* (pp. 127-137). Boston: McGraw-Hill.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Saito, K. & Lyster, R. (2012). Effects of form-focused instruction on L2 pronunciation development of /ɹ/ by Japanese Learners of English. *Language Learning, 62*, 2, 595–633.

Saito, Y., Nagasawa, Y., & Ishikawa, S. (2011). Effective instruction of shadowing using a movie. In A. Stewart (Ed.), *JALT2010 Conference Proceedings* (pp. 139-148). Tokyo: JALT.

Saito, K., Trofimovich, P., & Isaacs, T. (2014). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*. Advance online publication. http://applij.oxfordjournals.org/

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 1-24. Advance online publication. doi: 10.1017/S0142716414000502

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT® Test*. *TOEFL iBT-21*. Princeton: Educational Testing Service.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*, 005-30.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002) *E-Prime user's guide*. Pittsburgh: Psychology Software Tools.

Schweda-Nicholson, N. (1990). The role of shadowing in interpreter training. *The Interpreters' Newsletter, 3*.Retrieved August 2012 from http://www.openstarts.units.it/dspace/bitstream/10077/2152/1/07.pdf.

Smith, B.L., Bradlow, A.R., & Bent,T. (2003). Production and perception of temporal contrasts in foreign-accented English. In M.J. Sole, D. Recasens, &J. Romero (Eds.), *Proceedings of the XVth International Congress of Phonetic Sciences, Barcelona, Spain* (pp. 519–522). Causal Productions.

Smith, B. L., & Hayes-Harb, R. (2009). Individual differences in the perception of final consonant voicing among native and non-native speakers of English. *The Journal of the Acoustical Society of America, 125*(4), 2761-2761.

Smith, B. L., & Hayes-Harb, R. (2011). Individual differences in the perception of final consonant voicing among native and non-native speakers of English. *Journal of Phonetics*, *39*(1), 115-120.

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.

Stibbard, R. M., & Lee, J. I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *The Journal of the Acoustical Society of America*, *120*(1), 433-442.

Tajima, K., Port. R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign accented English. *Journal of Phonetics, 25*, 1-24.

Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal, 28*(1), 54-71.

Thomson, R. I. (2011). Computer Assisted Pronunciation Training: Targeting second language vowel perception improves pronunciation. *CALICO Journal, 28*, 744-765

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amu076

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(04), 905-916.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2015). Flawed self-assessment: Investigating self-and other-perception of second language speech. Bilingualism: *Language and Cognition*. Advance online publication. doi: 10.1017/S1366728914000832

van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication, 35*, 103–113.

Walker, R. (2005). Using student-produced recordings with monolingual groups to provide effective, individualized pronunciation practice. *TESOL Quarterly, 39*(3), 550-558.

Walker, R. (2010). *Teaching the pronunciation of English as a lingua franca*. Oxford: Oxford University Press.

Wilkerson, M. E. (2013). The sound of German: Descriptions of accent by native and non-native listeners. *Die Unterrichtspraxis/Teaching German, 46*(1), 106-118.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762-789.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231-252.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231-252.

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication, 55*(3), 486-507.

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*, *41*(5), 369-378.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 6

Appendices

Appendix A: Information for Raters for Study 2

(Adapted from Materials used by Crowther et al., 2014)

TOEFL Integrated task

Part 1

Speakers were given 45 seconds to read through an assigned passage.

Part 2

Speakers listened to an 80-90 second lecture related to the previous passage.

Part 3

Speakers answered a question that required them to use information from both the passage and lecture they were previously exposed to.

Theme of Version 1

Psychology

Question: Explain how the two examples discussed by the professor illustrate differences in the ways people explain behavior.

Theme of Version 2

Social Interaction

Question: Explain how the examples of tying shoes and learning to type demonstrate the principle of audience effects.

Appendix B: Explanation of Speech Measures for Listeners for Study 2

Speech Measures

<table>
<tr><td>**Accentedness**</td><td>○ **This refers to how much the speech differs from the local variety of English spoken**</td></tr>
<tr><td colspan="2">This measure will be rated using this scale:
1 ○---○ 1000
“heavily accented” “no accent at all”</td></tr>
<tr><td>**Vowel and consonant errors**</td><td>○ **This measure applies to individual sounds.**
○ **Refers to errors in the pronunciation of individual sounds within a word. These errors may affect both consonants and vowels:**
○ **Speaker says “that” but you hear “*dat*”**
○ **Speaker says “pen” but you hear “*pin*”**
Such errors also include the removal and additions of sounds:
○ **Speaker says “house” but you hear “*ouse*”**
○ **Speaker says “spray” but you hear “*supray*”**</td></tr>
<tr><td colspan="2">This measure will be rated using this scale:
1 ○---○ 1000
“frequent” “infrequent or absent”</td></tr>
<tr><td>**Word stress errors**</td><td>○ **This measure applies to individual words that are longer than one syllable.**
○ **Refers to errors in the placement of stress in words with more than one syllable. These errors include misplaced stress:**
○ **“*comPUter*” is pronounced as “*compuTER*”**
○ **“*FUture*” is pronounced as “*fuTURE*”**
These errors also include absent stress, such that all syllables sound the same:
○ **“*comPUter*” is pronounced as “computer”**
○ **“*FUture*” is pronounced as “future”**</td></tr>
<tr><td colspan="2">This measure will be rated using this scale:
1 ○---○ 1000
“frequent” “infrequent or absent”</td></tr>
</table>

<table>
<tr><td><b>Intonation</b></td><td>○ <b>This measure applies to utterances longer than a single word.</b>

○ <b>Can be described as the melody of speech. It refers to natural movements of pitch as we produce utterances.</b>
  ○ <b>Pitch goes up in “<i>Will you be home tomorrow?</i>”</b> ↗
  ○ <b>Pitch goes down in “<i>Yeah, I’ll stay at home</i>.”</b> ↘
  ○ <b>Pitch goes down and then up in “<i>Yeah, I’ll stay at home… but only until 3</i>.”</b> ↘↗
  ○ <b>Intonation should come across as natural and unforced.</b></td></tr>
<tr><td colspan="2">This measure will be rated using this scale:
1 ○---○ 1000
“unnatural” “natural”</td></tr>
<tr><td><b>Speech rate</b></td><td>○ <b>This measure applies to utterances.</b>

○ <b>Describes how slowly or quickly someone speaks.</b>
  ○ <b>Speaker can speak slowly with many pauses and hesitations.</b>
  ○ <b>Speaker can speak very fast.</b>
  ○ <b>Speakers can speak at a natural rate and can be comfortable to listen to.</b></td></tr>
<tr><td colspan="2">This measure will be rated using this scale:
1 ○---○ 1000
“too slow or too fast” “optimal”</td></tr>
</table>

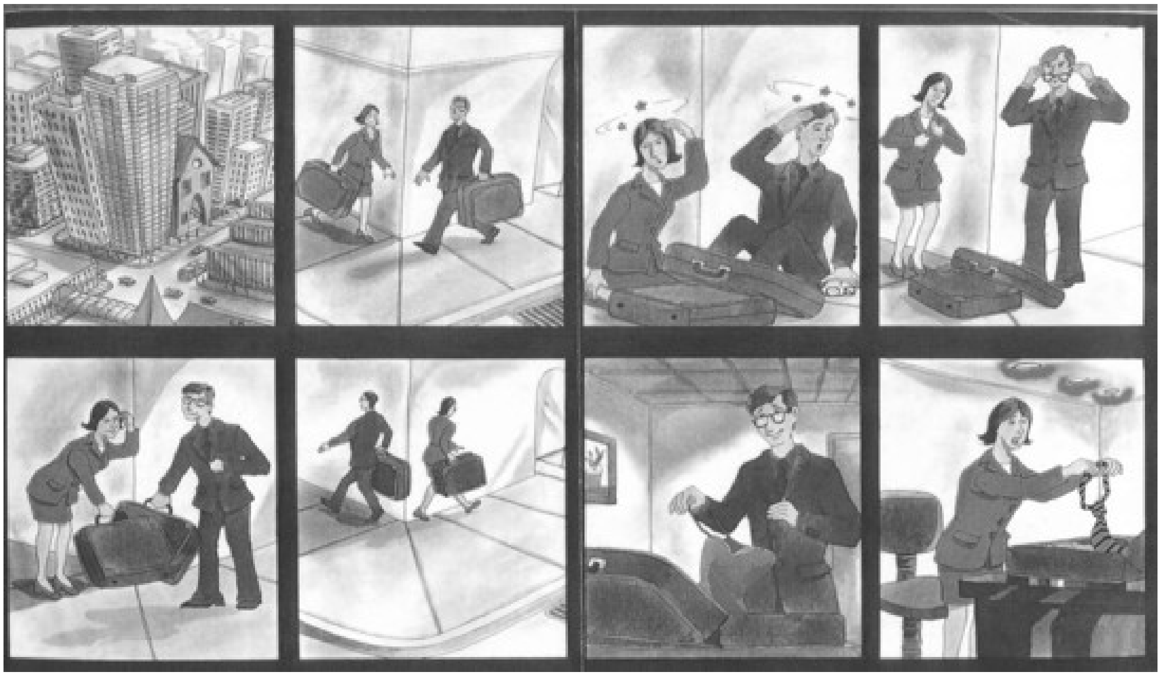
Appendix C: Means and Standard Deviations for Comprehensibility Scores and Speech Measures for Each Listener/Speaker Group for Study 2

Speaker group/rating	Mandarin-L	French-L	Hindi-L	English-L
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Mandarin-S				
Comprehensibility	559 (152)	281 (106)	309 (140)	339 (92)
Segmental errors	494 (79)	298 (56)	405 (102)	287 (58)
Word stress errors	522 (60)	347 (70)	442 (99)	445 (52)
Intonation	428 (85)	291 (56)	414 (63)	393 (54)
Speech rate	438 (165)	281 (126)	281 (119)	374 (138)
French-S				
Comprehensibility	593 (230)	650 (166)	568 (236)	636 (199)
Segmental errors	551 (98)	491 (134)	516 (138)	418 (103)
Word stress errors	576 (100)	455 (103)	519 (155)	503 (74)
Intonation	525 (125)	518 (137)	529 (164)	545 (76)
Speech rate	525 (116)	527 (124)	475 (194)	545 (124)
Hindi-S				
Comprehensibility	561 (251)	592 (215)	733 (173)	677 (133)
Segmental errors	462 (117)	420 (138)	611 (174)	443 (120)
Word stress errors	555 (89)	520 (111)	600 (138)	591 (80)
Intonation	550 (95)	512 (165)	601 (136)	529 (112)
Speech rate	537 (130)	858 (155)	632 (157)	621 (112)

Note. Scores range from 1 (low rating) to 1000 (high rating).

Appendix D: Suitcase Narrative Prompt for Study 3

The pictures below tell a story. Please look at the pictures, then using your own words, tell the story in the pictures.



(Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J., 2009).

Available at http://www.iris-database.org/

Appendix E: Script for Shadowing Test for Study 3

Title on YouTube Clip: Penny giving response to Buffy, she even wants to watch another. *Cute*

URL: http://www.youtube.com/watch?v=eyeXAK_sbV0 (0:00-0:53)

Leonard: So did you love it? Of course you love it. How could you not love it? Tell me how much you loved it.

Penny: It was cute.

Leonard: Aw, don't say cute. That's the worst.

Penny: What's wrong with cute?

Leonard: It-It just makes things seem small. It diminishes them.

Penny: So do you want me to stop calling your little tushy cute?

Leonard: You can try but nobody's gonna believe you. I just, I don't understand how you could watch a show that great and not be excited by it.

Penny: I liked it! I'm excited.

Leonard: Well then, tell your face.

Penny: What do you want from me?

Leonard: You know what? Never mind, we gave it a shot. Let's just see what else is on.

Penny: Oh come on! Don't be like that! I'm sorry I called it cute. Let's watch another one.

Leonard: Really?

Penny: Yeah, it was fun. Kinda reminded me of my high school.

Appendix F: Third Interview For Study 3

1) How did you find the shadowing project overall?

2) Did the way you used the dialogue and practiced change as the experiment went on?

3) Do you feel like being able to listen to your own recorded speech was helpful or was it unnecessary?

4) Do you find the dialogues difficult or easy? Is one week too much, too little, or about right for practicing?

5) What did you think of the length of the study? Do you feel like you are getting tired of the shadowing or would you like to keep going if you could?

6) Is there anything that would improve the shadowing for you?

7) Do you think this is an effect technique for improving pronunciation?

8) Do you feel your pronunciation is improving as a results of the shadowing activities?

9) Do you feel your listening skills are improving?

10) Are there any other skills you think this type of practice is helping you with?

11) Would you recommend this technique to friends?

12) Do you feel that aside from the money, you benefitted from taking part in this project?

13) Can you think of anything else that you would like to say about this project or about the shadowing?

Rate out of 9

How much you like the shadowing activity

I don’t like it I love it

1 2 3 4 5 6 7 8 9

How much do you think it is helping your pronunciation?

Not at all A lot

1 2 3 4 5 6 7 8 9

How much is it helping you listening skills

Not at all A lot

1 2 3 4 5 6 7 8 9