12-1-2009

ESL speakers' production of English lexical stress: The effect of variation in acoustic correlates on perceived intelligibility and nativeness

Paul Edmunds

Paul S. Edmunds
Candidate

Linguistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Caroline L Smith, Chairperson

[signature]

Jill P. Morford

[signature]

ESL speakers' production of English lexical stress: The effect of variation in acoustic correlates on perceived intelligibility and nativeness

BY

Paul Edmunds

B.S., English and Spanish, University of Minnesota, 1996
M.A., Linguistics, University of New Mexico, 2003

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Linguistics

The University of New Mexico
Albuquerque, New Mexico

December, 2009

DEDICATION

Para mis tres Rosas

ACKNOWLEDGMENTS

Most of what I have done for this dissertation I learned to do during the course of this project. The number of people who have lent their support during the long process of developing, working on, and actually completing this dissertation is innumerable. As such, many thanks are in order.

I would first like to thank the members of my dissertation committee: Dr. Caroline Smith, my committee chair, an incredibly patient professor who not only helped me conceptualize the project but also took great care in teaching me how do much of the phonetic analysis, beginning with word segmentation, so I could actually do it; Dr. Amy Neel who got me excited about speech and hearing sciences, in particular studying speech intelligibility, and whose suggestions were always invaluable; Dr. Jill Morford and Dr. Julia Scherba de Valenzuela who gave me excellent comments and suggestions throughout this project and who were incredibly encouraging.

Thanks also to Dr. Philip Dale for his helpful comments, expertise with statistics, and encouragement, and to Dr. Sherman Wilcox for his advice during my graduate career.

This work was supported in part by a UNM Graduate Dean's Dissertation Fellowship and an RPT grant from the Office of Graduate Studies at UNM. In addition to several departments who hired me to teach their courses over the years, I thank the Center for English Language and American Culture (CELAC) at UNM for whom I have worked during the past nine years. Thanks especially to Dr. Larry Smith, my first boss at

CELAC, as well as Dr. Thomas Bogenschild, Director of the Office of International Programs and Studies, and friends who have offered their moral support.

My greatest thanks go to my wife, Rosa María, an amazing person who has encouraged me through the highs and lows. It goes against all logic to believe that I could have done this without her support and love. And to Rosa Paola and Angela Rosa, my two beautiful daughters: everything I do is for you.

ESL speakers' production of English lexical stress: The effect of variation in acoustic correlates on perceived intelligibility and nativeness

BY

Paul Edmunds

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Linguistics

The University of New Mexico
Albuquerque, New Mexico

December, 2009

ESL speakers' production of English lexical stress: The effect of variation in acoustic correlates on perceived intelligibility and nativeness

BY

Paul Edmunds

B.A., English and Spanish, University of Minnesota, 1996
M.A., Linguistics, University of New Mexico, 2003
Ph.D., Linguistics, University of New Mexico, 2009

ABSTRACT

Non-native speakers of English often experience problems in pronunciation as they are learning English, many such problems persisting even when the speaker has achieved a high degree of fluency. Research has shown that for a non-native speaker to sound most natural and intelligible in his or her second language, the speaker must acquire proper prosody, such as native-like speech rhythms (Tajima et al., 1997; Wenk, 1985; Wennerstrom, 2001). This dissertation investigates how native English and Spanish ESL (English as a Second Language) speakers compare in their production of three acoustic correlates of lexical stress in English, namely the relative durations of stressed and unstressed vowels and their relative intensities and fundamental frequency (F0) values. A set of three-syllable words, including cognates and non-cognates, was analyzed. The results from the production study were used to design a listening task that

investigated how the ESL speakers' varying productions of these acoustic cues affected native English listeners' perception of their speech intelligibility and nativeness.

The ESL speakers produced a wider range of values than did the native speakers for all three acoustic correlates. The ESL and native speakers differed statistically in their productions of Spanish/English cognates with different stress patterns in each language and often differed on non-cognates. The ESL group produced the most native-like patterns in cognates with the same stress pattern in each language.

The stimuli for the listening task were words recorded by native English participants and subsequently modified to emulate the production of the acoustic correlates of lexical stress by the ESL speakers. Listeners' ratings of speech intelligibility were statistically higher for words in which intensity was increased on the vowel that is expected to receive lexical stress compared to the adjacent unstressed vowel. Increasing vowel duration on the unstressed vowel led to statistically lower ratings of both intelligibility and nativeness.

This dissertation contributes to a small body of research regarding the production of prosody in second language speech. The results suggest that a speaker's prosody alone can influence a native listener's perception of speech intelligibility and nativeness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.0 Overview

Non-native speakers of English often experience problems in pronunciation as they are learning English, many such problems persisting even when the speaker has achieved a high degree of fluency. Difficulties with pronunciation involve segmental errors (i.e. incorrect production of target consonants and vowels) and prosodic (or suprasegmental) errors (e.g. rhythm and stress). Much previous research has analyzed differences in the production of segmental information by non-native speakers compared to native speakers (e.g. Flege, 1984, 1993; Rogers & Dalby, 2005; Tarone, 2005). Studies regarding the perception of segmental structure depending on a speaker's native language and cross-linguistically have also received a great amount of attention (cf. Grabe et al., 2003). Relatively little, however, is known about non-native speakers' prosody when speaking in their second language. The present study investigates the production of one aspect of prosody – lexical stress – in the speech of Spanish ESL (English as a Second Language) speakers.

Lexical stress refers to the assignment of stress to a specific syllable within a word. Lexical stress can be marked in various ways. Segmentally, lexical stress in English is associated with long or unreduced vowels (Ladefoged, 1993). Prosodically, lexical stress in English is cued by various acoustic dimensions including fundamental frequency (F0), intensity, and duration (Lehiste, 1996). Lexical stress in English can be used to distinguish between lexical items in some noun/verb pairs. For example, placing stress on the first syllable of the word ***re****cord* indicates that the word being pronounced is a noun, whereas placing stress on the second syllable (*re****cord***) indicates that the word is a verb. Although noun/verb pairs such as those in English do not exist in Spanish, lexical stress disambiguates the meaning of words in a similar way in Spanish among conjugated verb forms. For example, the word ***ha**blo* means "I speak" while *ha**bló*** means "he/she spoke."

The correct placement of stress on a particular syllable is crucial in conveying the intended meaning of a word in languages that use stress distinctively such as English and Spanish. Consequently, this prosodic feature of speech can affect the perception of intelligibility, "the degree to which a speaker's utterance is understood by a listener" (Tarone, 2005, p. 493). It can also impact the perception of nativeness, the degree to which a speaker sounds like a native speaker of a particular language. In this dissertation, the production of lexical stress by Spanish ESL speakers is compared to that of native English speakers and the degree to which variation in the acoustic correlates of lexical stress affects speech intelligibility and perceived nativeness is investigated.

This dissertation is composed of six chapters. Chapter 1 gives a brief overview of lexical stress and the importance of producing it in a native-like way in order to ensure efficient communication between the speaker and his or her listener. This chapter also provides motivation for the study, the research questions and hypotheses, and a brief overview of second language teaching pedagogy with a focus on pronunciation teaching to frame this work in a broader, educational context. A synopsis of the following chapters is given below in Section 1.5.

1.1 Background and Motivation for the Study

My interest for doing this study centered on English pronunciation comes from over ten years as an ESL instructor. I have witnessed the importance of pronunciation in all core ESL classes (reading, writing, grammar, and speaking) at almost every level (i.e. from beginning to highly advanced students). In speaking classes, in particular, pronunciation instruction is an important component of the coursework, though pronunciation issues also come into play in reading, writing and grammar classes. Through casual observation, I came to notice that correct production of stress could be particularly difficult for Spanish ESL students and got into the habit of writing down words the students mispronounced during class. I would often transcribe these words phonetically (particularly if the pronunciation problem was segmental in nature), make a note when one syllable was substituted for another, or when patterns of stress placement did not sound as expected. Though not a systematic or scientific study, one thing became clear to me: mispronunciations due to incorrect placement of lexical stress were common, and these differing pronunciation patterns were very noticeable.

In addition to noticing that misplacing lexical stress was a very salient error, I noticed that these errors often occurred on Spanish/English cognates such as the word *catholic*. Native Spanish speakers often pronounce the word *catholic* with stress on the second syllable ("ca-**tho**-lic") [kæ.ˈϴa.lɪk] instead of with stress on the first syllable ("**ca**-tho-lic") [ˈkæ-ϴə-lɪk]. It seemed that word familiarity was not always related to these mispronunciations. For example, the ESL speakers seemed almost as likely to misplace stress on a very common word such as *catholic* as they were to do on a less common word like *lunatic*. These rather casual observations motivated me to think more deeply about how the cognate status of a word could affect a language learner's pronunciation patterns. It also made me wonder just how much these divergent patterns could affect a listener's perception of the speaker's speech intelligibility or if they could affect the perception of nativeness.

Certainly, producing words that are easy to understand (that is, highly intelligible) is beneficial to enabling communication. For a non-native speaker, being intelligible is often the number one goal. Frustration ensues, for both the speaker and the listener, when one's words are difficult (if not impossible) to understand. As a second language learner of Spanish, and a person who lived in a Spanish speaking country on different occasions, I know first-hand the feeling of not being understood, something my students here feel as non-native speakers of English in their host country. Besides feeling frustration myself, I know that my listeners were sometimes frustrated as they could not easily understand me. I always remember a friend who would tell me, "Habla mejor. No te entiendo" (Speak better. I don't understand you). This request stayed with me as an ongoing motivation to produce speech that was as *intelligible* as possible.

But being intelligible is only part of the picture. While living in Venezuela, my host family would call me a "loro" (parrot) as I made a habit out of repeating their words out loud – imitating their speech as closely as possible in order to practice my pronunciation. This exercise made me realize how important it can be (at least for some language learners) to sound as *native* as possible. These personal experiences as a language learner and language teacher have come together in this dissertation. I feel privileged to be able to work on a project such as this in which I was able to get specific insight into a question that has intrigued me about language learning, specifically how native and non-native speakers of English differ in terms of pronunciation and how those differences can affect communication in terms of a listener's perception of speech intelligibility and nativeness.

1.2 Need for the Study

The present study first assesses production of lexical stress in English by native Spanish speakers and later investigates how differences in lexical stress production, compared to that of a native English speaker, can affect a speaker's perceived intelligibility and degree of nativeness. The Spanish ESL speakers involved in this study are university students from South America. Studying the language of Spanish ESL speakers is important because they represent a growing sector of the population of the United States. According to the U.S. Census of 2000, there are more than 35.3 million Hispanic or Latinos residing in the United States, totaling 12.5% of the U.S. population. Of course, not all Hispanic people are ESL speakers. The U.S. Census defines *Hispanic* as "Persons of Hispanic origin, in particular, those who indicated that their origin was Mexican, Puerto Rican, Cuban, Central or South American, or some other Hispanic origin" ("United States Census 2000").

The number of Hispanic people living in the U.S. is increasing, with a projected 120.6 million Hispanics (approximately 24.4% of the total U.S. population) to reside in the U.S. by the year 2050. Close to home, the 2000 U.S. Census states that 42.1% of New Mexico's population is Hispanic and some reports have stated that 50% of the residents of New Mexico speak Spanish. Assuming many Hispanic people have learned or are in the process of learning English as their second language, this dissertation is directly relevant not only for our community in New Mexico but for a substantial portion of the U.S. population.

As Wennerstrom (2001, p. 230) states, “For those nonnative speakers whose goal is to participate in English-language speech communities, an understanding of the English-specific aspects of prosody will be an enhancement.” This statement reflects a reality in the ESL classroom where pronunciation is a major element of instruction. For a non-native speaker to sound more natural in his or her second language, the speaker must acquire proper prosody, such as native-like speech rhythms, in order to be most intelligible (Tajima et al., 1997; Wenk, 1985; Wennerstrom, 2001). However, until recently instruction in pronunciation has largely been limited to practice in segmental aspects of speech. The focus has been on how to produce consonants and vowels correctly, often using drills involving minimal pairs. Although this type of segmental instruction continues to be both popular and important, there is a growing trend for instruction to focus on prosodic aspects of speech in order to help the student be more intelligible and sound more natural. Wennerstrom (2001) gives some evidence for this shift to working on prosodic aspects of pronunciation over segmental aspects when she explains how many of the most current text books, for example *Well Said* (Grant, 2000), have relegated specific instruction on segmental aspects to the appendices. Indeed, *Well Said* begins right away with practice in speech rhythms. As Wenk (1985, p. 170) puts it, “The development of a proper sense of the nature and operation of speech rhythms is possibly the single most important area of pronunciation study.”

The necessity for a complete understanding of the speech of second language speakers extends beyond the ESL classroom. According to a recent newsletter from the National Student Speech Language Hearing Association (NSSLHA), “In national surveys, SLPs [Speech-Language Pathologists] have identified assessment of culturally and linguistically diverse children as a top priority need for enhanced personnel preparation efforts and continuing education” ("¿Qué hay de nuevo? Learn what's new in serving those of diverse backgrounds", 2006, p. 3). Treatment for bilingual clients at speech clinics is in high demand. Thus, in addition to serving those who work in second language teaching, the present study hopes to be of service to those who work with non-native speakers of English in the field of communication science and disorders.

The main desire is that the results of this dissertation further the understanding of prosody as it relates to second language speakers and language learning. It is hoped that the work here will help further our understanding of how lexical stress is realized by both native English and Spanish ESL speakers, and how variation in the acoustic cues related to lexical stress can improve or reduce a speaker’s level of intelligibility or the degree to which he or she sounds like a native speaker of English.

1.3 Research Questions and Hypotheses

The two main goals of this dissertation were to 1) Compare the production of lexical stress in English by Spanish ESL speakers to that of native English speakers and 2) investigate the relation between native and non-native lexical stress production and listeners’ evaluations of speech intelligibility and nativeness.

To investigate these two questions, the research was divided into three parts: a production study, modification and resynthesis of various target words based on observed patterns of production, and a listening study using the words generated with these modifications to be presented as stimuli to native English listeners who rated the words for intelligibility and nativeness. The research questions for the production and perception studies and their hypotheses, respectively, are explained below.

1.3.1 Research Questions for the Production study

The first research question is: How does the production of lexical stress by Spanish ESL speakers compare to that of native English speakers? That is, is there a difference in the acoustic characteristics of vowel duration, intensity and fundamental frequency between stressed and unstressed syllables produced by native and non-native English speakers? Although other acoustic correlates of lexical stress in English have been documented, such as the contribution of spectral tilt and noise in the region of F3 (Okobi, 2006), the three acoustic correlates of lexical stress in English at the focus of this study (duration, intensity and fundamental frequency) are perhaps the best understood and most studied with regard to their production by native English speakers (e.g. Fry, 1955, 1958). For this reason, it is beneficial to select these three acoustic correlates of stress to compare the speech of non-native speakers of English to that of native English speakers.

Pertinent to the first research question of how non-native speakers produce lexical stress in English is the issue of language transfer from a speaker's first language (L1) to his or her second language (L2). Studies in bilingualism have shown that language transfer from L1 to L2 is common and that second language learners often experience varying degrees of transfer (Wenk, 1985). Transfer is a possible reason for mispronunciations in a speaker's second language, and it is hypothesized below that cognates may be especially susceptible to effects of language transfer.

This leads to the second research question of the production study: Does the cognate status of a word affect a Spanish ESL speaker's production of lexical stress? For example, do Spanish ESL speakers pronounce cognates with differing stress patterns in the two languages differently when speaking in English? Furthermore, do pronunciation errors occur more frequently in cognates with differing stress patterns in each language than cognates with the same stress patterns? And how do non-cognates fit into this pattern?

To address these research questions, the production of three acoustic correlates of lexical stress were measured in the vowel that is expected to receive lexical stress and compared in an adjacent unstressed vowel. These acoustic measures were then compared between native and non-native speaker groups. The results of the production study and answers to these research questions are found in Chapter 4.

1.3.2 Hypotheses for the Production study

First, ESL speakers, who have less knowledge of the English phonological system, may rely heavily on increasing fundamental frequency to mark stress. As increasing F0 is perhaps the most important of the three acoustic correlates in creating the perception of stress in Spanish (cf. Llisterri et al., 2003), language transfer of F0 patterns from Spanish to English may result.

Secondly, language transfer may cause higher degrees of variability in English/Spanish cognates with lexical stress on different syllables in the two languages (DSCs) (e.g. ***lu***-*na-tic / lu-***ná***-tico*). It is hypothesized that the productions of the DSCs will be the most divergent from the native speakers' productions because negative transfer from the Spanish ESL speakers' L1 will influence his or her production of the word English. Same-stress cognates (SSCs) (e.g. English ***ra***-*di-o /* Spanish ***ra***-*di-o*) should be produced more like the productions of the native English speakers as a positive transfer of patterns is expected. The production of non-cognates is hypothesized to be more similar to that of the native speakers because their pronunciation may have been learned by rote, though more variation is predicted than in the SSCs.

1.3.3 Research questions for Listening study

In the second part of this study, the consequences of variation in the production of lexical stress by the non-native speakers are examined. Stress patterns from non-native speakers' productions were modeled by modifying and resynthesizing original recordings of native speakers in order to imitate prosodic characteristics of the non-native speakers. Modifying the speech of native English speakers ensured good segmental quality in the resynthesized tokens in order to avoid the effect of potential segmental errors by the ESL speakers. This allowed the focus of the investigation to remain strictly on prosodic aspects of the speech signal. This, in part, addresses a claim by Munro (2008, p. 203) that "volume, loudness of speech, voice quality, and clarity can affect judgments [but] may be very difficult or even impossible to assess … [though] they are clearly relevant to communication." Manipulating the various acoustic correlates of stress individually makes it possible to investigate the extent that single acoustic dimensions affect the perception of intelligibility and nativeness of a particular speaker.

The first research question of the listening study attempts to determine if changes to individual acoustic cues for lexical stress have an effect on a native English listener's ratings of speech intelligibility and nativeness. For example, does producing the stressed vowel with longer duration than the unstressed vowel (as native speakers typically do) make a speaker sound more intelligible than if the unstressed vowel has increased duration? Similarly, does raising fundamental frequency on the stressed vowel make a speaker sound more intelligible than raising F0 on the unstressed vowel? And for these patterns of production, do native English listeners rate the stimuli as sounding more native-like when vowel duration is lengthened or F0 is increased on the stressed vowel rather than on the unstressed vowel?

The second main research question for the listening study is: Do ratings of speech intelligibility for the manipulated stimuli correlate positively with ratings for nativeness? In other words, does a speaker who sounds highly intelligible also sound highly native? The results of the listening study and answers to these research questions are given in Chapter 5.

1.3.4 Hypotheses for the Listening study

It is hypothesized that increasing vowel duration, intensity and F0 on the stressed vowel will lead to higher ratings of intelligibility and nativeness than tokens having the same values for vowel duration, intensity and F0 on both the stressed and unstressed vowels, or tokens with less vowel duration, intensity and F0 on the stressed vowel than on the unstressed vowel. This hypothesis is based on the literature that has shown that increases in these variables are trademarks for marking prominence and lexical stress in English (Fry, 1955, 1958; Lehiste, 1996). An absence of correlation between speech intelligibility and nativeness is hypothesized as a speaker may sound highly intelligible though not necessarily native.

The next section provides a brief overview of language teaching, in particular with respect to the teaching of pronunciation, in order to put the goals of this dissertation in a broader, educational context.

1.4 A brief overview of second language teaching

According to Richards and Rodgers (2001, p. 3) an estimated 60% of the world's population is multilingual, making "from both a contemporary and a historical perspective, bilingualism or multilingualism…the norm rather than the exception." It has also been estimated that there are at least 750 million English language speakers in the world (Morley, 1991) and that more than half of all English speakers are non-native (Jenkins, 1998; Morley, 1991). These figures help explain why there has been a "population explosion…in adult and near-adult learner groups" of English as a second language (Morley, 1991, p. 482).

The demand for English language instruction over the years has motivated a wide variety of teaching pedagogies. One of the first widely adopted language teaching approaches was called the Grammar-Translation method, an approach that dominated language teaching for a century, from the 1840s to the 1940s (Richards & Rodgers, 2001). The goal of this method was to teach students how to recognize grammatical patterns in order to be able to read and translate texts. The focus of this method was on reading and writing with very little emphasis on speaking or listening. The focus of language instruction began to change in the 1860s when "linguists emphasized that speech, rather than the written word, was the primary form of language" (Richards & Rodgers, 2001, p. 9). It was also around this time, in 1886, that the International Phonetic Alphabet was developed. One of the motivations behind the development of this system of transcribing sound was to "advocate the improvement of language teaching" (Richards & Rodgers, 2001, p. 9). At about this time, the Direct Method for teaching language emerged, a method in which "learners are exposed to the language, use it, and gradually absorb its grammatical patterns" (Richards & Rodgers, 2001, p. 9). This method emphasized that all instruction be in the target language and that oral communication skills be the focus of learning. However, the Direct Method lost popularity during the 1920s as it lacked a strong methodological basis (Richards & Rodgers, 2001).

Following the decline of the Direct Method, the Audiolingual Method came into popularity, particularly from the 1940s to the 1960s. Pronunciation practice was a very important component of this curriculum, if not the most important aspect. The Audiolingual Method was derived from the Army Specialized Training Program that came into existence before World War II, whose goal was to prepare personnel to quickly learn enough of a language in order to communicate with people in other countries. According to Richards and Rodgers (2001), the short-term goal of the Audiolingual Method was to improve listening comprehension and acquire accurate pronunciation, while the long-term goal was to speak like a native speaker of the target language. To accomplish this goal, dialogues were pronounced by teachers and students repeated the dialogues, sometimes adding small pieces of information to them, in order to practice "correct pronunciation, stress, rhythm, [and] intonation" (Richards & Rodgers, 2001, p. 59). The instruction also "featured articulatory explanations, imitation, and memorization of patterns through drills and dialogues, with extensive attention to correction" (Morley, 1991, p. 485). A guidebook to teaching language via the Audiolingual Method by Hocket (1959) (cf. Richards & Rodgers, 2001, p. 52) stated that classroom practice required "drill, drill, and more drill, and only enough vocabulary to make such drills possible." The narrow focus of the Audiolingual Method, which largely disregarded situational contexts in language use, attributed to the decline in its popularity. This method was found to transfer poorly to daily speaking situations in which any variety of topics or contexts could arise. However, although this approach has largely been replaced by the Communicative Approach (discussed below), pronunciation drills from the Audiolingual method are still utilized in many pronunciation classes today.

As Richards and Rodgers (2001) point out, changes in instruction reflect learners' needs. Major shifts in approaches have moved from reading and writing in order to translate texts (e.g. the Grammar-Translation Method) to speaking fluently in very pre-determined situations (e.g. the Audiolingual Method) to being able to gain proficiency, fluency, and "communicative competence" (Hymes, 1972), that is, the ability to communicate effectively in a language community in activities that are necessary for the speaker and those around him or her.

In the 1970s, the currently most popular language teaching pedagogy was born. The Communicative Approach (or Communicative Language Teaching) emphasizes the need to focus language teaching on proficiency in communication rather than mastering particular grammatical or pronunciation patterns. Today, the majority of ESL institutes and intensive English programs in the United States adopt the Communicative Approach to language teaching. This method is comprehensive as it strives to "develop procedures for the teaching of the four language skills [reading, writing, grammar, and speaking] that acknowledge the interdependence of language and communication" (Richards & Rodgers, 2001, p. 155). Of these four language skills, speaking is "usually viewed as the most complex and difficult to master" (Tarone, 2005, p. 485)

When the Communicative Approach was first in use, instruction in pronunciation was limited as scholars debated the importance of teaching pronunciation and whether, in fact, pronunciation could be taught explicitly (Morley, 1991). According to Morley (1991, p. 485), many ESL programs began to drop the pronunciation component in instruction altogether and little new pronunciation material was published in favor of focusing almost entirely on more “authentic…learning activities and materials.” Benrabah (1997, p. 157) states that the switch away from practicing pronunciation could have been a reaction to the belief that this piece of ESL pedagogy was linked only to meaningless drills and repetition “at the expense of fluency and meaningful interaction.”

However, as the communicative approach came into its own, pronunciation practice began to regain popularity, particularly beginning in the mid-1980s. Pronunciation practice became an essential component in the communicative model (Morley, 1991). At this time, the focus began to change from narrower linguistic to broader communicative competencies with an emphasis on the teaching of suprasegmentals (Hahn, 2004). This was perhaps due to the belief that “suprasegmental (or prosodic) features are often the primary factors affecting judgments of the quality of L2 pronunciation” (Tarone, 2005, p. 492) and because speech intelligibility, in particular, is considered an essential, if not primordial, component of communicative competence (Chun et al., 2008; Derwing & Munro, 2005; Morley, 1991; Munro, 2008; Tarone, 2005). As Derwing and Munro (2005, p. 384) state, in “communicatively oriented ESL settings, improved intelligibility is generally identified by pedagogical specialists as the most important outcome of pronunciation instruction.” Before this shift, pronunciation instruction had traditionally focused on the acquisition of a native-like accent (Scales et al., 2006).

Morley (1991, p. 489) suggests that attention to pronunciation is growing stronger because teachers feel that it is “imperative that students’ educational, occupational, and personal/social language needs, including reasonably intelligible pronunciation, be served with instruction that will give them communicative empowerment – effective language use that will help them not just to survive, but to succeed.” And it is not only teachers who feel this way. Students, experiencing that breakdowns in communication due to pronunciation difficulties can be frustrating (Hewings, 2004), often want to practice pronunciation and request that this element of instruction be included in weekly lesson plans.

In intensive English programs in the United States, where students study on average 20 hours per week, class time is typically divided into daily hour-long blocks of reading, writing, grammar and speaking. It is primarily in the speaking class that attention is given to specific pronunciation instruction in an integrated (often spontaneous) approach, though some programs teach pronunciation as a separate component. Following strategies of the task-based Communicative Approach, students are given a situation in which they have to perform a particular function, for example requesting a transaction at a bank. During this activity, the student will likely have to read a form, write in personal information, speak to a bank employee, and understand the spoken response. The focus is on completing the task in a competent manner and producing fluent speech while doing so. Instruction in prosody is frequently integrated into such an activity (Chun et al., 2008) by explaining and practicing the use of lexical stress to add narrow (or sentence) focus to an utterance with an otherwise neutral stance (for example, “I want to deposit one-**hundred** dollars” vs. “I want to **deposit** one-hundred dollars”).

Implementation of prosody in real world situations benefits speakers and listeners alike by facilitating mutual understanding.

1.5 Organization of the Dissertation

This first chapter has given an overview of this dissertation and has provided a practical context for the work as it relates to second language teaching. It is hoped that the results from this study can be implemented in language classrooms or other mediums for language support such as speech clinics. Chapter 2 reviews the literature pertinent to this study, focusing on issues such as lexical stress and previous studies of speech production and perception in the areas of speech intelligibility, accentedness, and nativeness. Chapter 3 begins by providing the methodology for the production study, giving details about the participants, recording procedure, stimuli, and the analyses used on the production data. It continues with an explanation of how the stimuli for the listening study were created and how the listening study was subsequently carried out. Chapter 4 provides the results of the production study as well as a discussion of how the results address the research questions for that study. Chapter 5 then details the results of the listening study and provides a discussion of these results in terms of its research questions. The dissertation closes with Chapter 6 which summarizes some of the main findings and discusses how the findings of the production study relate to the outcome of the listening study. This final chapter also acknowledges strengths and weaknesses of the research design and points to directions of future research that are inspired by the work in this dissertation.

1.6 Summary

In this chapter, a brief overview of lexical stress in English was given, emphasizing the importance of placing lexical stress on the correct syllable in order to facilitate communication with native language listeners who expect a particular stress pattern for a word, or who may rely on stress cues to disambiguate a word's meaning. The usefulness of this dissertation was argued as few studies have addressed issues of prosody in second language learners, a population that is growing daily in the United States. The research questions and their hypotheses were developed, and a brief overview of ESL instruction, in particular as pronunciation teaching is concerned, was given to frame the study in a broader, educational context. In the next chapter, academic research and literature pertinent to this dissertation and the questions posed herein is reviewed.

Chapter 2: Literature Review

2.0 Overview of Prosody

‘Prosody’ is a term used to refer to suprasegmental properties of the speech signal. Whereas segmental features describe the consonants and vowels which make up syllables and words, “suprasegmental features are properties of speech sounds or their sequences that are simultaneously present [with the segmental features]…[and] do not change the distinctive phonetic quality of the speech sounds, but do modify the sounds in a way that may change the meaning of the utterance” (Lehiste, 1996, p. 227). According to Lehiste (1996), these suprasegmentals are “overlaid” on the segmental features. Both the segmental and suprasegmental information delivered to the listener are important for creating understanding of a speaker’s message.

Prosody plays a fundamental role in everyday communication between speakers and their listeners. Acoustic dimensions that reflect prosodic structure include speech rate, rhythm, pauses, duration, intensity and fundamental frequency. These acoustic features can convey different kinds of information, such as structural information. For example, the prosodic organization of an utterance can disambiguate possible meanings via placement of phrase boundaries (indicated in the following example by the pause related to the placement of the comma): *When you learn gradually, you worry more* vs. *When you learn, gradually you worry more* (Shattuck-Hufnagel & Turk, 1996, p. 198). Thus, prosodic structure can contribute to linguistic meaning in a way that may not be found in the lexicogrammatical structure of an utterance alone (Wennerstrom, 2001).

Prosody can also convey structural information such as sentence focus via pitch accents, prosodic features used to mark particularly salient information in discourse

(Wennerstrom, 2001) (see also Section 2.1 below), and placement of boundary tones. Regarding the latter, a sentence-final rising pitch boundary is often used to elicit a backchannel from a listener, or to designate a prototypical yes/no question. Plateau (flat) boundaries are also linked to subsequent discourse and are often employed in hesitations, for example while a speaker pauses to formulate the completion of his or her utterance. Falling pitch boundaries typically mark the end of a declarative sentence, and that a speaker is yielding the floor for another interlocutor. In Spanish, Fernández Planas and Martínez Celdrán (2003) also found that declarative statements tend to have phrase-final descending contours while yes/no interrogatives have phrase-final rising contours. Finally, differing prosodic patterns can convey extralinguistic information, such as emotion (Ladefoged, 1993). Certainly, then, as Wennerstrom states, prosody is “central to the interpretation of spoken language” (2001, p. vii).

Prosodic variables are expressed by their corresponding acoustic and perceptual features. More specifically, the acoustic features of duration, intensity (or amplitude) and fundamental frequency (F0) correspond to the perceptual features of length, loudness and pitch (Lehiste, 1996; Seddoh, 2004). Researchers take into consideration both acoustic values and perceptual judgments when analyzing prosody (Bunton et al., 2000; Shattuck-Hufnagel & Turk, 1996). Analyzing these features may elucidate how control of prosody can lead to effective communication, or when a reduced ability to produce normal prosody, perhaps while learning to speak a second language or resulting from a particular speech or language disorder, may impair communication (Brookshire, 2003; Duffy, 2005; Munro, 2008).

In the following sections, aspects of prosody that are related to this dissertation will be discussed, beginning with a description of pitch accent and how it cues prominence in speech. The chapter continues with a review of literature concerning previous studies of lexical stress in English and Spanish, the issue of language transfer in bilinguals, and previous perception studies of speech intelligibility, accentedness and nativeness. The chapter concludes by providing information based on the literature regarding the application of pronunciation instruction in prosody to the ESL classroom.

2.1 Pitch accent

Pitch accent refers to "the various tones associated with lexical items that a speaker decides are especially salient in the information structure of the discourse" (Wennerstrom, 2001, pp. 18-19). That is, prominent words are marked as prominent in English by a characteristic pitch pattern. The most common type of pitch accent is an H* tone which is produced as a high pitch on new information introduced into the discourse. According to Wennerstrom (2001), "there is a fundamental association between high pitch and new information in English" (p. 34). Wennerstrom also points out that any phrase can have more than one pitch accent (pp 34-5). The assignment of pitch accent to lexical items depends on the content of the discourse and the intentions of a speaker.

In English, pitch accent is typically produced most clearly on the syllable with lexical stress (Wennerstrom, 2001). The presence (or absence) of pitch accent on a given target word is of great importance to the present study as pitch accent tones affect the amount of movement in fundamental frequency, as well as changes in vowel duration and intensity, that a speaker produces on a particular lexical item. For this reason the target words in this experiment were coded for presence or absence of a nuclear pitch accent in addition to the segmentation of the target word as explained further in Chapter 3.

2.2 Previous studies of lexical stress in English

Non-native speakers often find it difficult to learn how to place lexical stress correctly (Hahn, 2004; Hubicka, 1981). However, acquiring accuracy in stress placement is important as it has been found that misplacement of lexical stress can lead to communication breakdowns (Hubicka, 1981) while speech intelligibility will increase if correct word-stress is achieved (Benrabah, 1997). Previous studies of lexical stress in English, the main focus of this dissertation, are reviewed here, in particular studies concerning the acoustic correlates of stress.

2.2.1 Acoustic correlates of stress in English

In an early study, vowel duration and intensity were found to be especially important for cueing the perception of lexical stress by listeners (Fry, 1955). Fry (1955) compared noun/verb pairs (for example, SUBject/subJECT, DIgest/diGEST) using resynthesized speech in order to control each variable independently. He found that duration, and to a lesser extent intensity, were important cues for marking lexical stress. Fry concentrated on measuring the duration and intensity of vowels since he found that the syllable being stressed or unstressed did not cause significant variation in the consonants. He also noted that maximum intensity within the syllable was always found in the stressed vowels.

Fry later included fundamental frequency, again via resynthesized speech, as a variable in his investigations on the perception of stress. He found that syllables with higher F0 on the vowel were judged to be more stressed than syllables with lower F0 on the vowel (Fry, 1958). In this set of experiments, however, Fry did not separate the cues of vowel duration ratio and F0, thus making it difficult to ascertain if one of these cues is more important than the other in signaling stress. Furthermore, although Fry acknowledges that vowel quality is a powerful factor in the perception of English stress, and potentially one of the strongest factors perceived by listeners, he did not control for vowel quality in his experiments.

Other scholars have also found that fundamental frequency is an important indicator of lexical stress in English. According to Beckman (1986), stress in English is "cued by a complex of phonetic correlates including duration and spectral quality as well as fundamental frequency" (p. 62). The combination of duration, spectral tilt, and noise in the region of the third formant (F3) have also been found to correlate with the perception of stress in English (Okobi, 2006).

Considering the interaction of these variables, many researchers have been interested in which acoustic cue might be most important in conveying stress on a particular syllable. Lehiste and Fox (1992), working with Estonian and English listeners, found subjects to be especially sensitive to variations in amplitude while judging prominent syllables. A similar finding is reported by Kochanski et al. (2005) who found amplitude to be especially important as a signal for stress in British and Irish English, while changes in fundamental frequency had much lesser effects. Finally, Nakatani and Aston (1978) (cited in Beckman, 1986) found that even small changes in amplitude have a large effect on perceived stress. These authors also found that changes to fundamental frequency and duration were also important in listeners' judgments of stress, although the weight of each of these two factors varied depending on the position of the target word in the sentence.

Other researchers working on the perception of lexical stress in English have argued that it can be difficult to separate the cues of duration and intensity. Turk and Sawusch (1996) suggested that "length and loudness are processed as a unit" although "extracting length information appears to be easier than extracting loudness information" (p. 3782). These scholars found that changes in loudness alone contribute only negligibly to the perception of stress (Turk & Sawusch, 1996, p. 3782).

2.2.2 ESL speakers' production of English lexical stress

A small body of research has been dedicated to analyzing ESL speakers' production of lexical stress in English. Zhang et al. (2008, p. 4498) studied the acoustic correlates of lexical stress as produced by Mandarin ESL learners. They state that Mandarin learners of English often have problems with English lexical stress – a difficulty which "may result in large part from the influence of native suprasegmental (tonal) categories." Noting that differences in the production of F0, duration, intensity, or vowel quality can "interfere with perception of stress contrast" (p. 4498), Zhang et al. compared the production of lexical stress in English of 10 Mandarin and 10 native English speakers in bisyllabic words in which the non-native speakers had no doubt as to where to place stress. The researchers sought to determine which correlate was most problematic for the Mandarin ESL speakers. They found that their participants produced "significantly less native-like stress patterns, although they did use all four acoustic correlates to distinguish stressed from unstressed syllables" (p. 4498). In particular, Zhang et al. found that both speaker groups used duration and intensity similarly to distinguish between stressed and unstressed syllables, while the Mandarin ESL speakers used greater increases in F0 on stressed syllables than did native English speakers, likely because of an association with lexical tones in Mandarin. The ESL speakers' production of formant patterns was found to be less consistent, as they sometimes produced these like the native English speakers and other times not.

Anani (1989, p. 15) found that Arabic ESL speakers "tend to divide English words into syllables to make them coincide with the … syllable pattern typical of the underlying phonological structure of Arabic." It was found that Arab learners tended to follow phonological rules and placed stress on the longer final syllables of words such as "headlights" and "pegboard" instead of placing stress on first syllable. Anani argues that these production patterns in English indicate that the Arabic ESL speakers are following Arabic stress placement rules.

In a study comparing the placement of stress by Spanish ESL speakers who had learned English at an early age and those who had learned English as adults, Guion et al. (2004), using bisyllabic non-words as stimuli, found that long vowels attracted stress more often than short vowels, that nouns were more likely to get syllable-initial stress, and that non-words tended to receive stress placement based on the stress pattern of phonologically similar real words. Guion et al. found that the group of ESL speakers who learned English at an early age ("early bilinguals", between 2.5 and 6 years old) came very close to placing stress in a manner similar to the native speakers, while the group who learned English at a later age ("late bilinguals", between 15 and 33 years old) had more variation in their stress placement. This finding "indicates that, given early and prolonged exposure to English, learners may be able to acquire knowledge of stress patterns quite similar, but not identical, to that of native speakers" (Guion et al., 2004, p. 35). This dissertation will add to the literature, such as studies like this one by Guion et al., regarding the production of English lexical stress by Spanish ESL speakers by studying real, three-syllable words and examining the acoustic features of second language speakers' productions lexical stress. The present study also examines if phonologically different words (in particular, different-stress cognates) have an effect on the Spanish ESL speakers' productions.

Much work has been done investigating the perception of lexical stress in English, particularly considering its acoustic correlates relative to perceived syllable prominence. This is fortunate as it provides groundwork for studying the production of lexical stress by second language speakers. This background makes it possible to probe how these factors contribute to perceptions of stress, and listeners' impressions of second language speakers, such as their intelligibility or nativeness.

2.3 Lexical stress in Spanish

To be able to consider possible influence of the L1 on the Spanish ESL speakers in this study, it is important to have an understanding of lexical stress as it is realized in Spanish. Stress patterns in Spanish are fairly predictable. Spanish words which end in a vowel typically receive penultimate stress while words ending in a consonant tend to have stress on the final syllable (Hualde, 2005). Furthermore, Spanish does not have reduced vowels as does English, and therefore Spanish ESL speakers "may lack the vowel reduction necessary for English rhythm" (Avery & Ehrlich, 1992, p. 153).

As approximately 73% of Spanish words end in a vowel, "most words of two syllables or more will have penultimate stress" (Guion et al., 2004, p. 211). Given this regularity, Guion et al. hypothesized that Spanish ESL speakers, when presented with bisyllabic non-words, might follow this same phonological stress patterning from their native language, indicating language transfer from L1 to L2. In fact, their results did show this trend, but only for the late bilinguals who “produced and preferred more initial syllable stress” in bisyllabic forms than the native English and early bilingual groups (p. 223). Guion et al. suggest that their results “may be due to the transfer of knowledge of a distributional pattern in which penultimate stress is the most common in the Spanish lexicon” (p. 223).

Lexical stress in Spanish has also been described, as has English, as being marked by changes in duration, intensity and fundamental frequency (Llisterri et al., 2003; Ortega-Llebaría, 2006). Changes in fundamental frequency may be the most important of the three acoustic features in marking stress in Spanish, although F0 cannot mark stress on its own. Fundamental frequency must be accompanied by either changes in duration, changes in intensity, or both, to signal stress for both real and non-words (Llisterri et al., 2003).

As noted in this section, one factor that potentially influences a second language learner’s speech is language transfer. The issue of language transfer is discussed in the next section.

2.4 Language transfer

As Munro (2008, p. 193) states, “The occurrence of foreign accents provides some of the clearest evidence that knowledge of a first language influences the acquisition of a second.” The perception of a foreign accent by native speakers can be explained by various factors. Transfer (referred to earlier as “interference” (Major, 2008)) can affect both segmental and suprasegmental aspects of speech (Major, 2008; Tarone, 2005; Zampini, 2008). Segmental characteristics can contribute to the perception of a foreign accent when a second language learner produces vowels or consonants that do not have the characteristics of those of native speakers. These segmental differences often result when the segmental inventory of a speaker's L1 does not match that of the L2, in which case the L2 speaker may approximate a sound based on the segmental inventory of the L1 (Flege, 1995). The prosody of a second language learner's speech can also cue this perception of a foreign accent. As Munro (2008) states, the intonation and rhythm of a language learner's speech often differ from that of a native speaker, and these qualities are influenced by properties of the L1 sound system.

Historically, language transfer has been a major consideration in the study of L2 speech. In fact, it was once believed that transfer could predict all errors made by second language speakers, though this is no longer believed to be true (Major, 2008; Tarone, 2005). As Zampini (2008, p. 220) states, "virtually all research on L2 speech production assumes that the learner's L1 sound system impacts L2 pronunciation, at least some of the time or in certain stages of L2 acquisition." This belief is echoed by other researchers such as Major (2008), Tarone (2005), and Zhang et al. (2008). This point is important for the present study as differences in the production of the ESL speakers' stress patterns may possibly be traced back to patterns from their native language, Spanish. In particular, the phonological system of Spanish, in which words ending in a vowel typically have penultimate stress and those ending in a consonant typically have stress on the final syllable, may influence the production of the ESL speakers in this study.

A more current view of language transfer is that a learner passes through different stages of acquisition, and the degree of transfer from L1 to L2 will depend on the stage of acquisition. Wenk (1985), for example, suggests that learners in earlier stages of acquisition, having less knowledge of the L2 phonological system, will likely produce speech with more characteristics that can be traced to the L1, while speakers who are highly experienced will demonstrate fewer L1 characteristics in their L2 speech.

Language transfer can have a positive or negative effect on L2 speech. Zampini (2008) notes that language transfer can facilitate production, particularly when the phonological pattern of a speaker's L1 matches that of his or her L2. As related to the present study, words that are cognates with the same stress pattern in both languages are likely to experience this positive transfer. Therefore, the ESL speakers should have less difficulty producing the words with a rhythm similar to native speakers. On the other hand, as noted by Anani (1989) and others, language transfer can have a negative effect on L2 speech, referred to as negative transfer. In the present study, for example, a cognate that has different stress patterns in the two languages may trigger the phonological pattern in the ESL speaker's L1. If this occurs, the L1 stress pattern may transfer onto the English word, resulting in a non-standard stress pattern in English. As Avery and Ehrlich (1992, p. 186) state, "errors in word stress are often a result of transfer from the learner's first language."

An important issue related to language transfer is a speaker's amount of experience with the L2. The participants in the present study are predominantly undergraduate and graduate students at a university who have consequently spent many years studying English and have had significant contact with native English speakers and with English-speaking settings in general. The degree of language transfer should, therefore, be lower for these individuals than it would be for learners with beginning or intermediate levels of proficiency.

In addition to comparing the production of L2 speech to native speech, another central issue of investigation of this dissertation is the perception of speech intelligibility and nativeness by native English listeners. Previous studies on intelligibility and nativeness are discussed in the following sections, as are previous studies of speech accentedness – a phenomenon that is related to nativeness.

2.5 Intelligibility

Perhaps the primordial goal in pronunciation teaching within the framework of communicative competence is attaining pronunciation that facilitates mutual intelligibility. Many studies have examined the production of prosody and its relationship to the intelligibility (often referred to as "comprehensibility") of second language learners, owing to the belief that intelligibility is especially influenced by suprasegmental features of speech (Derwing, 2008). Derwing, Munro and Wiebe (1998), for example, found that language learners who received instruction and practice in English prosody showed a greater improvement in comprehensibility and accentedness when producing spontaneous English narratives than did students who received either segmental instruction or no pronunciation instruction. Results such as this have lead many researchers to believe that prosodic errors are more detrimental to speech intelligibility than are segmental ones (Munro & Derwing, 2001). However, both segmental aspects of speech, such as the production of consonants and consonant clusters, and suprasegmental aspects, such as vowel length and word stress, are elements of speech that need to be produced in a native-like way in order for language learners to best be understood by English listeners (Hewings, 2004). Particularly relevant to this dissertation, Avery and Ehrlich (1992, p. 186) have stated that “research has shown that incorrect stressing of polysyllabic words greatly affects comprehensibility.”

One such study that investigated the prosody of L2 speech focused on speech rate. Analyzing the effect of speech rate on the perception of comprehensibility, Munro and Derwing (2001, p. 456) found that “listeners tended to assign the highest ratings [for comprehensibility] to L2 speech that was somewhat faster than the rates generally used by L2 users … [while] both very fast and very slow speech tended to be less highly rated.” This result confirmed their hypothesis that having a speaking rate similar to that of typical native English speakers will help ESL speakers be more understandable.

Wennerstrom (1998) evaluated the contribution of prosody to the intelligibility of 20 Chinese ESL speakers who were candidates to become International Teaching Assistants (ITAs). She obtained ratings of these speakers giving short lectures, using a rating scale that included aspects of segmental pronunciation, prosody, fluency and grammar. Of the variables studied, Wennerstrom found that prosody was the most significant contributor to higher scores of communicative competence for the ITAs. In particular, proper use of paratones, modulation of F0 as a function of the topic structure of the discourse, was found to have the strongest correlation with intelligibility of the ITAs’ discourse. Based on her findings, Wennerstrom (2001, p. 1) emphasizes that “intonation is not only a stylistic component of accent but also a meaning-bearing grammatical system.”

Other studies have also investigated the intelligibility of Chinese ESL speakers. Tajima, Port and Dalby (1997) modified the syllable duration of sample recordings of a native English speaker and a Chinese speaker of English. The syllable duration of the native English speaker was modified to match the durational patterns of the Chinese speaker, and conversely the durational patterns of the Chinese speaker were adjusted to approximate those of the native English speaker. The modifications of duration were done while retaining the spectral and source characteristics (e.g. intensity and F0) of the Chinese ESL speaker. For the native English speaker, the authors found that intelligibility judgments decreased the more the original durational patterns were modified to resemble those of the Chinese ESL speaker. Conversely, the Chinese ESL speaker gained increasingly better intelligibility ratings when his temporal patterns were modified to become closer to those of the native English speaker. The authors conclude that "intelligibility of foreign-language speakers may be enhanced if explicit training is provided on temporal properties of their speech" (p. 1). In another study of Chinese ESL speakers, Juffs (1990) reported the perception of many speech errors due to incorrect stress placement.

Hahn (1999) also studied stress placement by Chinese ESL speakers who read a short academic lecture. Three versions of the lecture were presented to native English listeners: one with stress productions like those of a native English speaker, one with non-native stress placement, and one with monotone production (and therefore also non-native). Note that these were all actual productions of the ESL speakers, not manipulations of the native speakers. Hahn found that listeners processed the native-like stress patterns more easily, recalled more information, and rated the speech more highly for communicative effectiveness than either of the non-native versions. In a further study on speech perception involving Chinese ESL speakers, Hahn (2004, p. 201) found that when primary stress was placed correctly, “participants recalled significantly more content and evaluated the speaker significantly more favorably than when primary stress was aberrant or missing.”

The studies mentioned here, many of which have concentrated on the speech of Chinese ESL speakers, suggest that appropriate prosody, in particular correct stress placement, is important for intelligibility. The research questions and findings from the aforementioned studies also highlight areas of investigation that can be asked of other language learner groups such as the Spanish ESL speakers that are studied in this dissertation.

Before continuing with literature relevant to studies of nativeness, the phenomenon of accentedness is reviewed here. Studies regarding the perception of a foreign accent have received a great deal of attention and are directly relevant to the concept and study of the perception of nativeness.

2.6 Accentedness

The issue of “accent” in speech, particularly the speech of second language learners, is important as it is a wide-spread phenomenon. Almost all adult L2 learners speak with some degree of an accent (Derwing, 2003; Munro & Derwing, 2001) though speakers who use an L2 more will likely have less accent, and vice-versa (Major, 2008). As Derwing and Munro (2005, p. 383) state, having an accent is a “normal consequence of second language learning.” This consequence can represent both negative and positive effects to the language learner. First, a speaker may experience negative social consequences, such as discrimination, for having a foreign accent (Derwing, 2008). Although many listeners find foreign accents interesting and attractive, others may have negative feelings toward an L2 speaker because of his or her accent. These negative feelings may lead an L2 speaker to experience diminished acceptability in a community (Munro, 2008), low ratings of status, often tied to perceived future occupations (Brennan & Brennan, 1981) and stereotyping due to their accents (Scales et al., 2006). Negative feelings due to accent can also create a social barrier between the L2 and L1 speakers (Schairer, 1992). The proliferation of accent reduction programs in the United States, programs which attempt to train language learners to sound like native speakers, is due in large part to speakers feeling frustration because of their accent, and wanting to change it in some particular way.

Though a foreign accent is often perceived as a negative attribute, accent can also bring positive effects for an L2 speaker. Munro (2008) states that one potential benefit of a perceived accent is that interlocutors may adjust their speech to adapt to an L2 speaker's needs. For example, the native speaker may speak more slowly or enunciate his or her words more clearly if it is believed that the L2 interlocutor has an accent and therefore may have less comprehension. Another positive effect of accented speech is the retention of social or cultural identity. Many speakers prefer to retain at least some trace of their L1 accent in order to be identified by their interlocutors as a speaker from their particular L1 community (Morley, 1991).

Retention of accent to indicate cultural identity is an important social issue, though it is not embraced by all language learners. In a study by Derwing (2003, p. 560), it was found that "contrary to popular opinion, the participants … did not evince an interest in retaining an accent as an indicator of identity." Derwing found that two-thirds of students did not feel discriminated against because of their accent, though the majority felt that they would be respected more if they pronounced English well. Thus, the social issues regarding accent are bound to be different depending on the individual speaker or the community in which the L2 speaker is participating.

Recent literature provides empirical evidence related to the notion of accentedness. A study by Brennan and Brennan (1981) analyzed ratings of the accents of 43 Mexican-Americans by 37 Anglo high school students. The results showed that ratings of "status" were negatively correlated with ratings of accent. In particular, the researchers found that the degree of accentedness in the speech of the Mexican-Americans was related to their perceived future occupations by the Anglo students. Speakers with more accented speech were believed to be bound for occupations of lower status while those with less accented speech were believed to become more successful.

Previous studies have also looked at the relation between accent and intelligibility. Scales et al. (2006) analyzed the ratings of the speech of ESL speakers by 10 American English speakers. The researchers found that the raters preferred accents that they found easiest to understand. These researchers state that the primary goal should be mutual intelligibility in communication rather than the reduction of foreign accent.

Other studies have examined both segmental and suprasegmental aspects of speech. In her dissertation, Shah (2002) compared the speech of Spanish ESL speakers compared to native English speakers. Specifically, she investigated how temporal contributions of overall word duration, voice onset time, intervocalic flap and stop closure duration, unstressed vowel duration, and stressed-unstressed vowel duration ratios lead to differing judgments of accentedness (p. iv). Shah found that each of these variables contributes to ratings of accentedness, and that "segmental durations in multisyllabic words differed for nonnative speakers' productions when compared to those of native speakers" (p. 75). Although Shah suggests that none of the variables alone have a strong correlation with accentedness, she notes that "some combination of these temporal deviations may account, at least in part, for native listeners' judgments of perceived accentedness" (p. 75). Shah concluded that "temporal differences in acoustic segments of native Spanish speakers of English are correlated, though low to moderately, with native listeners' perception of accentedness on multisyllabic words" (p. 82). Shah's study is especially relevant to the present dissertation as her results suggest that acoustic correlates of accent in the speech of Spanish ESL speakers may contribute to explaining perceptions of accentedness.

2.7 Nativeness

Much previous research has analyzed L2 speech and perceptions of intelligibility and accentedness. Far less research has investigated the relationship between L2 speech and the perception of nativeness. As noted in Chapter 1, the term nativeness in this dissertation is defined as how much an ESL speaker sounds like a native speaker of English. In a sense, nativeness is used here as the opposite of accent. Rather than asking listeners to rate how much the speech of an ESL speaker sounds unlike (i.e. accented) what one would expect from a native speaker of English, it is here asked how similar the speech of the ESL speaker is to a native speaker. The experimental technique in this dissertation looks at perceptions of differences in prosody alone by making comparisons of speech material in which the segmental information is held constant. Thus, as far as segmental aspects of speech are concerned, the speech should sound quite “native” throughout. For this reason, it seems more appropriate to test perceptions of “nativeness” than of “accentedness.” Interestingly, some researchers have equated the terms nativeness and accentedness. For example, Schairer (1992) asked listeners to rate the comprehensibility, voice agreeability, and “nativeness of accent” of L2 Spanish speakers, thus collapsing the issues of nativeness and accent into one category for ratings. Schairer’s (1992, p. 318) findings suggest that “emphasis should be on native-like production of both stressed and unstressed vowels to enhance the communicative potential of the learner’s speech.”

Investigating how “native” an L2 speaker sounds is perhaps less common than questioning the speaker’s degree of “accent” as many scholars argue that obtaining native pronunciation in an L2, particularly when learning an L2 as an adult, is very uncommon. Researchers have argued that only a very small percentage of adult L2 learners actually acquire native pronunciation (Derwing, 2008; Flege et al., 1995a, 1995b; Hewings, 2004). Attaining native pronunciation in an L2 can be difficult due to “maturational restrictions,” psychological factors such as “motivational components, learning strategies, [and] habits…[that] prevent… [a learner] from doing so,” and “socio-emotional mechanisms linked to identity development” (Spoelders et al., 1996, p. 117). Obtaining native pronunciation may also depend on elements such as a learner’s first language, age, language aptitude, attitude, personality, exposure to the language, instructional approach, and communicative context (Avery & Ehrlich, 1992; Celce-Murcia et al., 1996; Derwing, 2008; Munro, 2008).

Munro (2008, p. 194) argues that obtaining "native pronunciation in the L2 is not only uncommon but unnecessary." Munro states that the primary goal for language learners is to be highly intelligible, acknowledging that not even native speech is always comprehensible. This point is echoed by Hewings (2004, p. 14) who states that the main goal of the language learner is to be understandable with "unobtrusive features of a non-English accent." Some scholars add that obtaining native pronunciation is not only unnecessary but also undesirable for many language learners as the speaker may fear loss of identity if they were to acquire native pronunciation in their L2 (Hewings, 2004; Morley, 1991; Tarone, 2005).

Jenkins (1998) also argues that the majority of language learners do not consider acquiring a native-like accent as their ultimate objective. However, Jenkins (1998, p. 125) suggests that "it is important to emphasize that we should all guard against political correctness, in the sense of telling our learners what their goals should be: in particular that they should not want to sound like native speakers if they clearly wish to do so." This point is extremely relevant to those who work daily in the ESL classroom where students often request extra pronunciation practice in order to acquire a more native-like pronunciation of English. Following on classroom experience, the studies detailed below surveyed ESL students' attitudes regarding pronunciation.

Derwing (2003) surveyed 100 ESL students regarding their attitudes toward language accent. In addition to 97% of students responding that they felt it was important to pronounce English well, 95% of students agreed or strongly agreed that they wanted to pronounce English like a native speaker (Derwing, 2003). Derwing (2003, p. 555) notes that, "when asked whether they felt their own identities would be jeopardized if they spoke English without an L2 accent, many replied that their identities were tied to their first languages, which were not threatened." Thus, although some students desired to retain some aspects of their accent as an identity marker, many desired to pronounce English like a native speaker, often citing that they felt they would be more respected by native English listeners if they did not speak with an L2 accent. Timmis (2002) obtained a similar finding. In a questionnaire surveying 400 learners of English in 14 countries, Timmis found that 67% of students wanted to sound like a native speaker.

Scales et al. (2006) surveyed 37 ESL learners consisting of 18 students from Asia and the others from Latin America (Venezuela, Peru, Colombia, Mexico). Of the students surveyed, 62% reported that their goal was to sound like a native English speaker while 38% said their main goal was to be highly intelligible. The future goals of the students surveyed were to live in the United States and work in fields such as education and business, and to travel. Most pertinent to the present dissertation, 9 of 11 (82%) of the native Spanish speakers stated that sounding like a native speaker was their primary goal, though they were not always able to clearly articulate why that was so. One student noted employment issues while another spoke of academic purposes.

The studies noted above suggest that although language researchers often do not feel that obtaining native pronunciation in an L2 is either desirable or necessary, L2 students often do. Therefore, it is worth investigating factors that contribute to attaining native-like pronunciation. Some work has begun to address this issue. Derwing, Munro and Wiebe (1998), for example, found that pronunciation practice can help ESL speakers sound like a native speaker. In particular, Munro and Derwing (2001) found that speaking rates tend to become more native-like with increased proficiency. This is fortunate as they found that listeners make perceptual judgments in part by speech rate. Specifically, L2 learners tend to have slower speaking rates than native speakers, at least at first. Having a speaking rate close to that of native speakers will increase the perception of nativeness by native English listeners.

A study by Schairer (1992) investigated native Spanish speakers' reaction to non-native speech. In her study, listeners rated the speech of Spanish L2s for comprehensibility, voice agreeability, and nativeness of accent. She found that vowel quality was the biggest indicator of native-like production, followed by consonant linkage and the production of the trilled [r]. Schairer (1992, p. 318) argues that "emphasis [of instruction] should be on native-like production of both stressed and unstressed vowels to enhance the communicative potential of the [L2 Spanish] learners' speech."

The studies in the last three sections have focused primarily on how L2 speech can affect perceptual judgments of intelligibility, accentedness and nativeness, as well as L2 learners' views on pronunciation. Certainly, there is a call for more empirical research on L2 speech and determining how to best apply the findings of the research to the classroom. These two issues are the focus of the next two sections of this chapter

2.8 Pronunciation research in applied linguistics

Although attaining native-like pronunciation that facilitates mutual intelligibility is considered important for many language learners and teachers alike, there have been few empirical studies of pronunciation in applied linguistics (Derwing & Munro, 2005; Levis, 2005). For example, Derwing and Munro (2005, p. 386) state that "it is widely accepted that suprasegmentals are very important to intelligibility, but as yet few studies support this belief." This claim is supported by other researchers such as Hahn (2004) and Levis (2005) who states that over the past 25 years there has been encouragement to teach suprasegmentals though very little pedagogy has been based on empirical research.

The usefulness of empirical research for developing more effective pronunciation teaching is obvious. As Levis (2005, pp. 370-371) states, "instruction should focus on those features that are most helpful for understanding and should deemphasize those that are relatively unhelpful." Munro (2008, p. 197) echoes this point when stating that "it is important to establish a set of priorities for teaching. If one aspect of pronunciation instruction is more likely to promote intelligibility than some other aspect, it deserves more immediate attention." Of course, we must first know what the most important elements are to ensure optimal instruction and learning outcomes. As Munro (2008, p. 210) argues, "because prosody encompasses a wide range of speech phenomena, further research is needed to pinpoint those aspects of prosody that are most critical."

Hahn (2004, p. 201) agrees that there is little empirical support for claims that teaching suprasegmentals is helpful and that "knowing how the various prosodic features actually affect the way native speakers…process nonnative speech would substantially strengthen the rationale for current pronunciation pedagogy." For that reason, Hahn (2004) reiterates that it is important to identify the phonological features that are most salient for native listeners. Due to the complex relationship between suprasegmentals and intelligibility, Hahn (2004, p. 201) argues that "it is helpful to isolate particular suprasegmental features for analyses." Hahn's argument supports the importance of the research in this dissertation in which the acoustic correlates of English lexical stress are isolated and manipulated individually to identify which are the most pertinent to the perception of speech intelligibility and nativeness.

Levis (2005) states that pronunciation teaching has been a study in extremes in that it was once considered the most important aspect of language learning (when audiolingual methods were favored) and then became very much marginalized in communicative language teaching. Of the research that has been carried out, such as that on intonation patterns, little of it finds its place in pronunciation textbooks (Derwing, 2008; Derwing & Munro, 2005; Levis, 2005; Tarone, 2005). Therefore, there is a need to first fill a gap in empirical research treating aspects of second language pronunciation and then to ensure that these findings are relayed to professionals in the fields of education and applied linguistics so that L2 students can benefit from these findings.

In the next section, issues involving the application of pronunciation practice in the ESL classroom will be discussed.

2.9 Application: Pronunciation in ESL classroom

Language learners learning English in their home country are considered to be EFL (English as a Foreign Language) students (e.g. a Venezuelan student learning English in Venezuela), whereas a student who learns English in a predominantly English speaking community, is categorized as an ESL (English as a Second Language) student. According to Derwing (2008, p. 348), the ESL student often wishes to “integrate into the local society, both socially and through employment.” These differing contexts affect how pronunciation instruction is delivered, as will be described briefly later in this section.

Different approaches to pronunciation teaching have been utilized over the years. Whereas a bottom-up approach to pronunciation teaching (traditionally the most common) begins with instruction on segmentals (cf. Avery & Ehrlich, 1992) and works up to overall prosodic patterns, pronunciation instruction today often begins with attention to prosodic elements of the language (e.g. stress, intonation, rhythm) in what is referred to as the “top-down” approach (Chun et al., 2008). This is especially true of classrooms using a communicative method (Chun et al., 2008). A common early activity relates to the implementation of intonation patterns that can be used for contextualized purposes such as creating focus on a lexical item within a sentence (Chun et al., 2008).

Once a general framework for the delivery of instruction is chosen, a next step in designing a course of any type is to consider the needs and desires of the students and create course objectives and learning outcomes. As stated earlier, ESL students are typically concerned with issues such as intelligibility, accent and nativeness. Students often voice their goals regarding attaining proficiency in these areas and teachers should consider which goals are realistic (Avery & Ehrlich, 1992). To do so, the students’ current abilities must be assessed in order to target strategies that will help achieve these goals.

Assessing students’ abilities is crucial in planning pronunciation teaching. Derwing (2003; Derwing, 2008) stresses that each student should be assessed individually to identify the student’s strengths and weaknesses and determine individual needs in pronunciation. These assessments can be done in a formal or informal way by the teacher and can include self-reports or self-assessments by the students. Self-assessments by students can provide insight into the students’ perceived needs, though these needs may be biased by the students’ previous experience with pronunciation instruction. Derwing (2003, p. 554) found that “of the pronunciation problems identified [by the students], roughly 79% were segmental [in nature], while only 11% were related to prosody." When interpreting this finding, it could be that students are simply more aware of segmental elements than they are of prosodic ones due to more previous instruction on segmental elements.

Once evaluations have been completed, the question becomes how to address the language learner's pronunciation issues. A complication arises at this point because students in ESL classes typically come from very mixed language backgrounds. Even the varying needs of students in EFL classrooms, where all learners are from the same native language background, can be challenging as individual students have individual needs. Therefore, integrating pronunciation lessons into class activities can be challenging in ESL classrooms as a particular speaker (or group of speakers) may have little difficulty with a particular element of pronunciation while others have great difficulty. A well-known example is Japanese speakers' difficulty acquiring /r/ and /l/ (Bradlow, 2008) which does not cause any trouble for Spanish speakers. As Derwing (2003) advises, focusing heavily on segmental instruction in mixed classrooms is inappropriate due to the variety of language backgrounds and, therefore, prosody should be emphasized as it can have greater importance for a larger diversity of students. Derwing (2008) also argues that instruction in prosody transfers better to spontaneous speech than instruction on segmentals.

Many instructors are reluctant to teach pronunciation and often unsure how to go about doing it (Derwing & Munro, 2005; Hewings, 2004) as they feel underprepared or have little support in terms of course materials. Derwing (2003) estimates that only about 30% of pronunciation teachers have formal linguistic training in pronunciation pedagogy. To address this issue, it is important that empirical research on pronunciation be conveyed in a clear manner to language teachers so that they can pass this information along to students.

To be certain, pronunciation should be considered an important element of ESL classroom instruction. It has been noted above that pronunciation is implicated in critical elements of communication such as speech intelligibility, and can also affect perception of nativeness. In addition, accurate pronunciation is critical for students needing to pass standardized English tests such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) for entrance into English-speaking colleges and universities, or when interviewed by entities such as the Foreign Service Institute which assesses not only a person's grammar and vocabulary but also comprehension, fluency and accent in oral interviews (Varonis & Gass, 1982). Pronunciation is also a key element in programs that prepare international teaching assistants to become teachers in American classrooms (Hahn, 2004; Wennerstrom, 1998).

2.10 Summary

In this chapter, literature has been discussed relating to aspects of prosody that are important for ESL learners, and for native listeners' perception of ESL speakers. In addition, research pointing out the need for more empirical work on pronunciation to apply toward pedagogy in ESL pronunciation was discussed, showing the immediate relevance of the study reported in this dissertation. To be sure, pronunciation, in particular with a focus on prosody, has gained a high level of importance in speaking classes. And as more empirical research is carried out and new teaching materials that incorporate this research become available, teachers will not only become more comfortable teaching pronunciation but also do so more effectively. Successful pronunciation instruction, in particular regarding prosody, deserves to retain a central role in the ESL classroom. As Levis (2005) argues, it would be best to find a middle ground when it comes to implementing pronunciation teaching in the classroom, neither making it the pinnacle of instruction nor disregarding it entirely.

In particular, previous research on speech intelligibility and perceived accentedness and nativeness of non-native speakers of English has brought up some points that will be addressed in this dissertation. First, few studies have been done regarding the prosody of second language learners, and there is a call in the literature for more investigation of this topic. More specifically, many previous studies do not examine the separate contributions of individual factors, such as the different acoustic correlates of lexical stress in English. Controlling each factor is crucial to best understanding how much each one contributes to the perception of speech intelligibility and nativeness.

With these points in mind, this dissertation aims to focus on one specific prosodic dimension, lexical stress, by first identifying variation in the production of the acoustic correlates of lexical stress in English by ESL speakers and then testing the effects of this variation while controlling individual factors (vowel duration, change in intensity, change in fundamental frequency) through speech resynthesis. Avoiding other sources of variation (such as vowel quality) makes it possible to test for the effect of specific factors of prosody on perceived intelligibility and nativeness. Having complete control of each variable allows the determination of which factor(s) are most important in the perception of intelligibility or nativeness by native English listeners. A detailed explanation of the methodology of this dissertation is given in the next chapter.

Chapter 3: Methodology

3.0 Overview

A distinct methodology was developed for each of the three main phases of this dissertation: 1) Production and acoustic analyses of read stimuli, 2) Manipulation and resynthesis of select target words and 3) Elicitation of intelligibility and nativeness ratings based on original and manipulated stimuli. A detailed description of each phase is provided below.

3.1 Production Study

3.1.1 Participants

Twelve native speakers of American English (6 males and 6 females) (mean age 34 years) and twelve native Spanish speakers who speak English as a Second Language (Spanish ESL) (6 males and 6 females) (mean age 29 years), all without strong regional dialects as perceived by the researcher, were recruited to make voice recordings of the study stimuli. The Spanish ESL speakers are from a group of neighboring countries in South America (Venezuela, Ecuador, Peru and Colombia) and speak a standard variety of Latin American Spanish. None of the speakers in the study presented unusual voice characteristics (e.g. extremely breathy or hoarse voices), or had a self-reported speech or hearing disorder. All of the participants in the study were either undergraduate or graduate students at the University of New Mexico. As a compensation for their effort, all participants were paid $20 for their involvement in the study. The speaker demographics are given in Appendix A.

All participants completed a language background questionnaire prior to arrival at the Speech and Hearing Sciences lab for recording. The native English speakers completed an abbreviated version (See Appendix B) which primarily requested their demographic information. The Spanish ESL speakers completed a longer version which, in addition to demographic information, included information about their experience with English and how long they had studied English (See Appendix C). Both versions of the questionnaire also asked if the subject had a known speech or hearing disorder. Only one person reported having a hearing disorder and was thus declined an invitation to participate in the study.

3.1.2 Recording procedure

All participants were recorded in a sound treated booth in the Speech and Hearing Sciences Department at UNM. Upon arrival, the researcher, who was present in the sound booth during the recordings, explained the procedure of the task. All instructions were given in English as this was the target language of the recordings. Before beginning, the participants completed an Informed Consent Form that been approved by the university's Institutional Review Board. This consent form gave background information about the study and explained the procedure for the task.

The researcher first explained how long the process would take: approximately 30 minutes for the native speakers and 60 minutes for the ESL speakers. The participants were given a bottle of water and were advised that they could take a break at any time in the recording process in addition to planned breaks that occurred approximately every 10-15 minutes. The participants were informed that they would be asked to read a list of sentences six times (a description of the study stimuli follows in section 3.1.3 below). It was explained that all the lists contained the same sentences but that in each list the sentences were presented in a different, randomized order. The researcher used a sample sentence that was not one of the experimental stimuli in order to demonstrate that it was important to read the sentences as fluently as possible and at a comfortable pace. The participants were told that any long pauses between words or any false starts would necessitate the repetition of the sentence. The participants were also informed that they would read a short passage ("The Rainbow Passage") three times. If they made a mistake in the reading passage, they were instructed to return to the beginning of the sentence in which the mistake was made and not to the beginning of the passage itself.

The order of readings remained constant for all participants. The participants always read two sets of sentences and then the reading passage, followed by another two sets of sentences and a second repetition of the reading passage, and so on. After each reading of two lists of sentences and the passage, the participants were invited to take a short break that lasted approximately 5 minutes but could last longer if the participant so wished. Additional breaks were granted to participants upon request.

For the recordings, the speakers wore a head-mounted microphone (Shure SM10A) to ensure a constant distance from the speaker's mouth. The researcher ensured that the microphone was always positioned 5cm from the speaker's mouth prior to the beginning of each recording. This consistent distance allowed for accurate sound intensity measurements to be taken for each speaker. The researcher also continuously monitored the sound level input into the voice recorder to ensure that clipping did not occur and followed along on a separate print-out of the study materials to ensure that the speaker did not skip any sentences. Following along also permitted the researcher to request that a participant repeat a sentence if a noticeable pause occurred between words or if the sentence was read incorrectly (e.g. the word order was changed or different words were used).

The speech recordings were made with a solid-state digital recorder (Marantz PMD670) that stores the data onto a compact flash card. The recordings were later transferred onto a personal computer. The participants' speech was recorded at a sampling rate of 44.1 KHz with 16-bit resolution and saved as .wav files. The .wav files were later used for segmentation of the target words and their subsequent acoustic analyses in the speech analysis program Praat (Boersma & Weenink, 2006).

3.1.3 Materials

The primary material used for the production study consisted of a set of constructed sentences with target words in early and late sentence positions (See Appendices A and B). Recordings of a standardized reading passage ("The Rainbow Passage") (for all speakers) and sentence stimuli with similar target words in Spanish (for the native Spanish speakers) were also collected, though not analyzed here. Each set of materials is discussed below.

3.1.3.1 Description of Target Words

The participants read a set of 48 sentences, each of which contained a multi-syllabic target word. These target words range from 2 to 5 syllables in length and were in all cases content words (nouns, adjectives, verbs and adverbs). No words such as *record* or *contract*, which can be pronounced as a noun or verb depending on the syllable that receives lexical stress, were included in the study. The target words were chosen almost exclusively from words found in ESL textbooks (e.g. Gilbert, 1993; Hewings & Goldstein, 1999; Olsen & Biley, 2002a, 2002b; Orion, 1997) designed for intermediate to advanced levels of ESL instruction. The selection was done in this manner in an effort to present the ESL participants with words that they were likely to have encountered while learning English. Therefore, many of the target words should have been familiar to the participants and within the range of their English ability. It was desirable that the target words be somewhat familiar to the ESL speakers to avoid them guessing the pronunciations.

The 48 target words included twelve 2-syllable words, twelve 3-syllable words, twelve 4-syllable words, and twelve 5-syllable words. Furthermore, the target words were equally distributed according to the syllable that should receive lexical stress. For example, six of the 2-syllable target words have stress on the first syllable (e.g. *PURpose*) while the other six have stress on the second syllable (e.g. *conTROL*). For three syllable words, one-third have stress on the first syllable (e.g. *MAnagement*), one-third have stress on the second syllable (e.g. *deVElop*), and one-third have stress on the final syllable (e.g. *introDUCE*). The four syllable words are equally divided for primary lexical stress falling on the first, second and third syllables. For five syllable words, primary lexical stress falls on either the second, third, or fourth syllable. Syllabification and position of lexical stress were taken primarily from the American Heritage Dictionary (Costello, 1992).

In addition to controlling for the distribution of lexical stress among syllables, the cognate status of each target word was taken into account. Half of the target words (*n*=24) are not Spanish/English cognates (e.g. *develop*) and the other half of the target words (*n*=24) are Spanish/English cognates (e.g. *October/octubre*) (see Appendix D for the list of non-cognate target words and Appendix E for the list of cognates). Among the

24 Spanish/English cognates, half (*n*=12) are orthographically identical cognates (e.g. *radio*) while the other half are not (e.g. *engineer/ingeniero*).

Finally, for the group of cognates, each target word was coded as to whether the same or a different syllable is stressed in the two languages. For example, the orthographically identical cognate *radio* receives primary stress on the first syllable in both the English and Spanish production ('ɹeI-di-oʊ / 'ra-di-o). On the other hand, the location of primary stress for the cognate *lunatic* (Spanish *lunático*) differs between the English and Spanish productions. In English, the primary stress for *lunatic* falls on the first syllable ('lu-nə-tIk), while in Spanish the primary stress falls on the second syllable (lu-'na-ti-ko). It was decided that segmental content is more relevant in comparing stress position rather than considering if the stress falls on the final syllable, the penultimate syllable, etc. because a Spanish ESL speaker may be more likely to relate segmental information with syllable stress rather than applying particular phonological stress rules for a given language (e.g. Spanish words ending in a vowel typically have stress on the penultimate syllable).

In total, 6,912 tokens of 2-, 3-, 4-, and 5-syllable target words were collected from the native English and ESL speakers (24 speakers x 6 repetitions x 48 words). Of these tokens, only the 3-syllable words (1,720 tokens total) were analyzed in this dissertation (24 speakers x 6 repetitions x 12 words). The 3-syllable words were chosen as a starting point as it was felt that they could potentially witness variation in production by the ESL speakers but be less difficult to produce overall than the 4- and 5-syllable words. Table 3.1 below illustrates the distribution of target words by cognate status and number of syllables. The Spanish equivalent of the example word is listed in parentheses. A complete list of the target words and the accompanying carrying sentences is included in Appendix D (non-cognates) and Appendix E (cognates).

Table 3.1: Distribution of target words by syllable length and cognate status

Type		**2 syllables**	**3 syllables**	**4 syllables**	**5 syllables**	**Example**
Non-cognate		6	6	6	6	management *(gerencia)*
Cognate						
	identical cognate	3	3	3	3	radio *(radio)*
	non-identical cognate	3	3	3	3	lunatic *(lunático)*
Total (48)		**12**	**12**	**12**	**12**	

3.3.1.2 Sentences

The principal stimuli for the experiment are forty-eight short sentences containing the multi-syllabic target words discussed in the previous section. The sentence length for the stimuli was controlled to be equal except for variation in the length of the target words. Each sentence contains 14 syllables plus the target word. The sentences were randomized six times to create six different lists. Each list of sentences was read once, providing a total of six repetitions of each target word per speaker. The first list of sentences included four sentences at the beginning of the list which were not evaluated. The first of these sentences (*Commercial planes can fly through the air at high speeds*) was used by the investigator to model for the speaker that the sentences needed to be read fluently. The four sentences placed at the beginning of the first list also provided a short warm-up for the speaker before the main target sentences were read for the first time. The last list of sentences also contained the four additional sentences from the first list, but presented at the end of the list. These were used to avoid any potential alteration in reading style when the speaker was aware that the experiment was almost finished.

Half of the target words appeared in sentence-early position and the other half in sentence-late position. Target words in sentence-early position were preceded by three syllables and followed by eleven syllables. For example:

*One way to **develop** the ideas is by brainstorming a list.*

Target words in sentence-late position were preceded by eleven syllables and followed by three syllables. For example:

*The owners are currently looking for good **management** for the store.*

The target words in the stimuli read by the participants were not, of course, bolded. Although not an element of the study at this time, the sentence position of the target words was varied in this way to permit the investigation of whether correct stress placement is more beneficial perceptually in one or the other of these two positions.

3.1.3.3 Reading passage

In addition to the list of sentences, the subjects read a standardized reading passage called "The Rainbow Passage" (Fairbanks, 1960). This passage was chosen as it is commonly used in other studies of speech production and intelligibility, especially in speech and hearing sciences (cf. Brookshire, 2003). It is also used by Occupational Safety and Health Administration (OSHA) for fit testing respirators ("United States Department of Labor, Occupational Safety and Health Administration; Appendix to Section 1910.134: Fit Testing Procedures", 2009). The Rainbow Passage was read three times by the subjects. As this passage is narrative in style, it could inform on differences in production between single-sentence production (such as the invented sentences) and a more continuous, read speech. Analysis of the recordings of this passage was not carried out for the present dissertation.

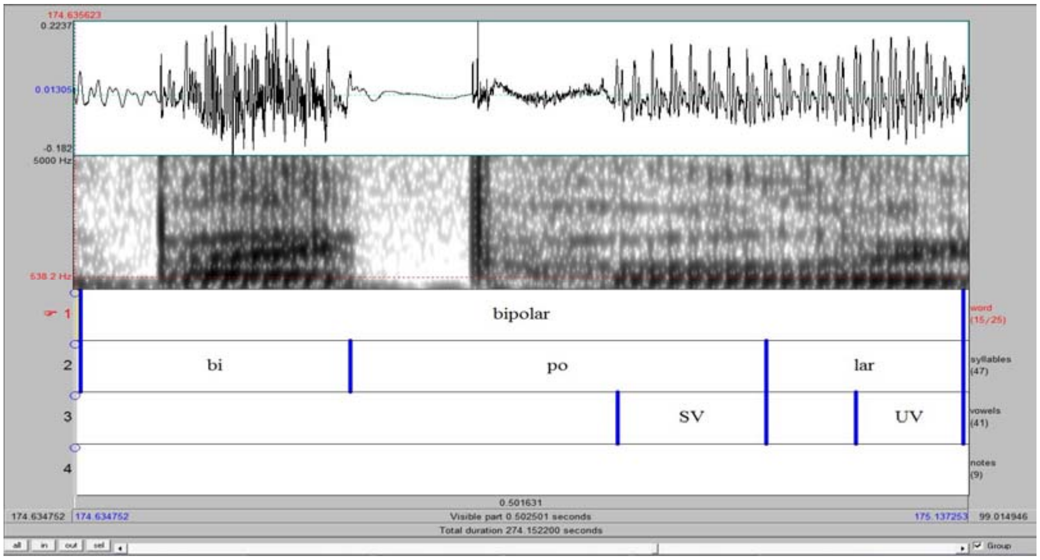
3.2 Analysis of the Recordings

The audio recordings were analyzed with the computer program Praat (Boersma & Weenink, 2006). The multisyllabic target words from the sentence lists were segmented into syllables and vowels following conventional segmentation techniques (Hillenbrand et al., 1995; Kiesling et al., 2006). The segmented words and their orthographic representations were visually displayed along with the waveform and spectrogram using Praat's Annotation tool.

3.2.1 Labeling

A Text Grid annotation for each target word was created in Praat. Each Text Grid included 4 tiers on which were annotated the word segmentation, syllable segmentation, the vowels targeted for measurement, and miscellaneous notes. The vowels targeted for measurement were, in each word, the vowel that was expected to be stressed (the SV) and an unstressed vowel in the same word (the UV). In most words, the vowel labeled as UV was the unstressed vowel immediately following the SV. For example, the word *radio* was segmented as [ɹeI.di.oʊ], with the first syllable labeled SV and the second syllable labeled UV. If the SV was in word-final position, such as in the word *disagree,* the word was segmented as [dIs.ə.gɹi] with the second syllable labeled UV and the third syllable labeled SV. A sample Text Grid is shown below in Figure 3.1.

Figure 3.1: Sample Text Grid displaying the four tiers labeled in the word *bipolar*



3.2.2 Exclusions

Criteria for excluding tokens of target words were: if a long pause on either side of the word was present, if the speaker produced a different number of syllables than expected (e.g. *ma-nage-ment* produced as *mange-ment*), or if a different word was produced in place of the target word (e.g. producing *monument* instead of the target *management*). If the pitch tracker in Praat could not determine F0 values in a target word, that token was examined individually. If this inspection yielded F0 values, the token was included for analysis. Those words in which no F0 value could be determined were excluded from analyses using F0. An additional criterion for exclusion was lack of pitch accent on a target word. This criterion is discussed in greater detail in section 3.4.2 below. Only 7% of the data (125/1720 tokens) was excluded for any of the reasons listed here.

One of the target words, *easily*, was produced as a two-syllable word (*eas-ly*) in 75% of the tokens by male native speakers and in 50% of the tokens by female native speakers. This target word, then, was re-segmented so that the first vowel was labeled as the stressed vowel (SV) and the final vowel in "ly" was labeled as the unstressed vowel (UV). This was done to permit the inclusion of a large number of tokens that would have otherwise had to have been excluded.

3.3 Measures

Acoustic measures were made of individual vowels, rather than syllables, because stress effects in English occur primarily in the vowels, rather than in the consonants (Fry, 1955).

The following basic and derived measures were taken in each of the target words:

1. Basic Measures:
    a. Duration: Three measurements:
        i. The duration of the entire word
        ii. The duration of each syllable within the word
        iii. The duration of the vowel in the syllable expected to receive primary stress (SV) and the adjacent unstressed vowel (UV) identified as described in 3.2.1 above.
    b. Amplitude: Two readings of maximum RMS (root mean square) amplitude of the vowels in the SV and UV were taken, measured in decibels (dB). These were obtained using the Get Intensity command in Praat with a view range of 40 to 140 dB and "mean energy" selected as the averaging technique as parameter settings. The first value was the

maximum within the entire vowel, and the second was the value at the midway point (50%) in the duration of the vowel.

c. Fundamental frequency (F0): Fundamental frequency, measured in Hertz (Hz) was measured in the stressed and unstressed vowels of the target words. These were obtained using the Get Pitch analysis command in Praat with the pitch range set to 60-500 Hz for all speakers. Measurements were taken at 25%, 50%, and 75% of vowel duration. These different measurement points make it possible to determine if there are changes in F0 during the course of the vowel.

2. Derived Measures

a. Vowel duration ratio: This measure is equal to the duration of the SV compared to the duration of the UV. This measurement is displayed as a ratio, i.e. $Duration_{SV}$:$Duration_{UV}$. The ratio of the vowel durations is reported rather than raw durations in order to normalize for variation in speech rate.

b. Change in intensity: The difference between the intensity of the SV and the UV, i.e. $Intensity_{SV}$-$Intensity_{UV}$, reported in dB.

c. Change in fundamental frequency: The difference between the fundamental frequency of the SV and the UV, i.e. $F0_{SV}$-$F0_{UV}$, reported in Hz.

The measurements of the above variables were compared for the native English speakers and the Spanish ESL speakers. The measurements from the native English speakers are considered as a control range to which the Spanish ESL speakers' production can be compared.

3.3.1 Applying the measures to the data

The three derived measures (vowel duration ratio, change in intensity, change in fundamental frequency) were calculated for the three-syllable target words. Two sets of analyses were performed. The first analyses included all 12 3-syllable target words. The second set of analyses included just 6 of the original 12 target words. The six words that were selected for this second analysis were those in which the native speakers consistently increased F0 on the SV compared to the UV.

These six target words are distributed evenly regarding cognate status, i.e. two are same-stress cognates, two are different-stress cognates, and two are non-cognates. The six words are given in Table 3.2 below.

Table 3.2: 6 Target words chosen based on consistently increased F0 on SV by native speakers, by cognate status (English/Spanish)

Same-stress cognate	Different-stress cognate	Non-cognate
radio ['ɹeɪ-di-oʊ / 'ra-di-o]	lunatic ['lu-nə-tɪk / lu-'na-ti-ko]	management ['mæ-nədʒ-mənt]
October [ɑk-ˈtoʊ-bɚ / ok-ˈtu-bre]	bipolar [baɪ-'poʊ-lɚ / bi-po-'lar]	location [loʊ-'keɪ-ʃən]

The motivation for selecting a subset of the 12 target words was to facilitate the comparison of fundamental frequency between the native speakers and the ESL speakers. Both groups of speakers showed considerable variation in their use of F0 over the set of 12 target words. There was also substantial intra-speaker variation. This made it virtually impossible to compare the overall behavior of the two groups. This variation arises from the speakers making variable choices in their overall intonation pattern for the sentences. Such variability is probably inevitable in tasks such as the one used in this study, where sentence structures were varied and speakers were encouraged to read naturally. In 40% of the tokens produced by the native speakers, a change in F0 between the SV and UV of greater than 5Hz was noted, and 26% of the data showed movement of less than -5Hz. For ESL speakers, 23% of the data showed an increase in F0 of greater than +5Hz and 34% of the data showed a decrease in F0 of less than -5Hz. However, in the six target words noted in Table 3.2 the native English speakers tended not to lower F0. By working with a subset of words that were less variable for the native speakers, it was possible to have a baseline for comparison of the ESL speakers' productions. A more in-depth discussion of the distribution of F0 in the target words is given in Chapter 4.

3.3.2 Test Statistics for Production Data

Statistical analyses of the production data were carried out in SPSS Version 17 (SPSS, 2009). Examination of histograms returned from descriptive analyses in SPSS revealed that the majority of the data was normally distributed. This visual examination was verified by results of the Kolmogorov-Smirnov Test of Normality run in SPSS. Therefore, independent sample t-tests were calculated for all data sets to determine if significant differences existed between the productions of the native English and ESL speakers. A non-parametric Mann-Whitney U test was run in addition to the parametric test for cases in which non-significant values in the Kolmogorov-Smirnov Test of Normality were returned. In these cases, the result of the non-parametric test is reported after the parametric statistic.

3.4 Additional Data

The productions of the target words were also evaluated for stress placement and pitch accent as will be explained in the following two sections.

3.4.1 Stress placement ratings

First, the production of each target word by the ESL speakers was evaluated by the researcher and two professional phoneticians to determine if the speaker had stressed the syllable that most commonly receives primary stress in pronunciations by native speakers. For example, if an ESL speaker produced the word *lunatic* as lu-NA-tic (rather than LU-na-tic) it was noted that the speaker had misplaced the stress for that particular production. In this way, productions by the ESL speakers that did not carry the anticipated stress pattern could be compared to those that did.

The native speakers produced two different pronunciations for four of the 3-syllable target words. These words (*disagree, unafraid, engineer,* and *introduce*) were produced in some cases with stress on the last syllable (as listed as the standard pronunciation in the American Heritage Dictionary (Costello, 1992)) and sometimes with stress on the initial syllable. As both productions of these four words sounded perfectly native-like to the experimenter (who is a native English speaker), all productions were included in which the final syllable was perceived as having at least secondary stress. However, if a target word was produced by an ESL speaker in a way that is atypical of a native English speaker's production (for example, pronouncing the word *disagree* with stress on the second syllable), that word was coded as having been produced incorrectly.

Only the ESL speakers' productions were evaluated for correct or incorrect stress placement. The researcher listened to all of the 864 sentences produced by the ESL speakers and decided if the target word was produced in a way that could be expected from a native English speaker. These judgments were coded as correct, incorrect or unsure and entered into an Excel spreadsheet. 88.8% of the tokens were judged to be correct and 11.2% incorrect. None of the tokens were judged as being unclassifiable. The reliability of the ratings was then evaluated by asking two professional phoneticians to rate stress placement in 16.65% of the data. Each phonetician rated one token of every target word produced by each female ESL speaker and each male ESL speaker. In total, these two raters rated 288 of the 864 target words produced by the ESL speakers, for a total of 33.3% of the ESL data. A two-way mixed intra-class correlation coefficient (ICC) was calculated to evaluate inter-rater reliability between the researcher and each professional phonetician. Inter-rater reliability was very high. The first rater agreed with the researcher on 98.6% of male productions (Chronbach's alpha = .956, $p < .0001$) and 97.2% of female productions (Chronbach's alpha = .916, $p < .0001$). The second rater agreed with the researcher on 100% of the tokens for both male and female speakers (Chronbach's alpha = 1, $p < .0001$ for each analysis). The statistical output from SPSS for this calculation is provided in Appendix F. These ratings were used in selecting tokens for manipulation as described below.

3.4.2 Pitch accent ratings

Secondly, the target words were screened to determine if they had been produced with pitch accent. Pitch accent refers to the "various tones associated with lexical items that a speaker decides are especially salient in the information structure of the discourse" (Wennerstrom, 2001, pp. 18-19). A word without pitch accent does not sound any more salient than other words in the discourse. In English, the lexical feature of pitch accent is typically aligned with the syllabic feature of primary stress. As Wennerstrom (2001) states, “if a speaker decides to associate a pitch accent with a lexical item, the pitch accent is usually manifest on the primary stress of that item” (Wennerstrom, 2001, p. 47). Therefore, the acoustic consequence of pitch accent is that a word will tend to be longer, louder, and produced with a more extreme F0 value (either higher or lower) than words without pitch accent. Because the properties that correlate with pitch accent are, to a large extent, the same properties that correlate with lexical stress, it was necessary for this study to control for presence and absence of pitch accent on the target words in order to be able to evaluate the acoustic correlates of stress.

All of the target words were reviewed for pitch accent by the researcher who re-listened to all of the 1,720 sentences in the study (by both native and non-native speakers) to decide if the target word was pitch accented or not. To evaluate the reliability of these judgments, 11% of the data was checked by two professional phoneticians, each of whom reviewed 5.5% of the total data in the study. The results were submitted to a two-way mixed ICC analysis. The pitch accent ratings of both professional phoneticians showed 96% agreement with the researcher’s ratings (Chronbach’s alpha = .844, $p < .0001$). The statistical output from SPSS for this calculation is provided in Appendix G.

Only eight productions of target words were identified as lacking pitch accent. These tokens were therefore excluded from the analyses.

3.4.3 Target Word Familiarity Scale

The Spanish ESL speakers also completed a Target Word Familiarity rating sheet (see Appendix H) after the recording of the sentences and reading passage was complete. This rating sheet asked subjects to estimate how familiar each of the target words was for them by judging how frequently they use each particular word. This method allows for frequency-of-use ratings of words to be generated on an individual-speaker level, rather than relying on frequency counts from other written or oral corpora. The scale included four possible ratings ranging from "rarely/never" use the word to use the word "daily." This information was collected in order to see if there was an association between stress placement errors and the speakers' frequency of use of these words.

Part II: Manipulation and resynthesis of target words

3.5 Overview

In this section, an explanation of how the stimuli were selected, modified and resynthesized for the listening experiment is given. The stimuli created via this methodology were subsequently rated by native English listeners for perceived speech intelligibility and nativeness.

Previous studies of stress have utilized resynthesis of the speech signal to investigate the perception of stress placement or accentedness (e.g. Fry, 1955, 1958; Grabe et al., 2003; Tajima et al., 1997). Modifications and resynthesis of the speech signal were used in this study in order to control variation of duration, intensity and fundamental frequency to see how these properties may affect the perception of speech intelligibility and nativeness. The basic approach used here was to take a representative example of the native English speakers' productions as a baseline. Then, this production was modified and resynthesized using Praat's Manipulation commands. The modifications that were applied were designed to reflect the range of pronunciations that had been produced by the ESL speakers. Importantly, the manipulations allowed for a great level of control over the stimuli for the listening experiment, allowing for precise modification of duration, intensity and F0 on the SV and UV. The methods for selecting which tokens to modify and creating modified versions of them are described in this section.

3.5.1 Selection of words for manipulation

A subset of the target words in the study were selected for manipulation and resynthesis. Some words, such as *radio* and *location*, were produced by the ESL speakers in a manner very similar to that of the native speakers, making it unlikely that they would reflect significant differences in the ESL speakers' productions. In order to investigate such differences, it was necessary to identify a subset of the target words that seemed likely to show differences between the two groups of speakers.

Words were selected for manipulation if they were mispronounced at least once by the ESL speakers, as confirmed by the stress-placement judgments described above in section 3.4.1. The five words that were selected are *lunatic, bipolar, develop, disagree,* and *October.* The first three of these words were mispronounced frequently by the ESL speakers. The last two words were only mispronounced rarely by the ESL speakers but were included to add some variety to the listening experiment.

3.5.2 Choosing the speakers for manipulations

One male and one female native English speaker were chosen as the models on which the manipulated versions would be based. The goal was to choose individuals who represented the general trends for the native speakers. These individuals were chosen by first examining the average values of the various acoustic measures for each individual word across all male or female native speakers. For example, the word *develop* had an average SV:UV vowel duration ratio of 2.1:1 across all male native speakers. Based on these comparisons, the male speaker that had the mean vowel duration ratio closest to 2.1:1 was considered the best baseline for manipulations for this word. Similarly, the female native speaker who produced a mean vowel duration ratio of 1.2 for *bipolar* was selected as this was the mean ratio for all six female native speakers.

The mean vowel duration ratios were compared for all native speakers across the five words to be modeled and the male and female speaker that were closest to the mean vowel duration ratios for all words were chosen as the model speakers. For the male speakers, this was the speaker identified as MNS4 (Male native speaker #4). For the female speakers, the one identified as FNS3 (Female native speaker #3) was chosen. These two speakers were also good choices as they produced clearly articulated the sentences and target words which helped create good stimuli for the listening task stimuli.

3.5.3 Selecting individual tokens to create manipulations

Choosing appropriate tokens from the speakers selected as models was also necessary. The token that was selected for modeling was the one whose duration measurements were closest to the mean across all speakers of the same gender. For example, if the female speakers had an overall mean vowel duration ratio of 2.1:1 for the word *disagree*, the production of FNS3 that was closest to 2.1:1 was used to create the manipulated tokens. When the selected speaker had more than one production with values equal to the mean value, a similar calculation was done using intensity, and the token whose intensity was closest to the mean was chosen. If intensity could not resolve the choice, F0 was used as the final tie-breaker.

Although the procedure outlined in the previous paragraph was followed insofar as possible, it was also considered to be desirable for the selected token to have modal voice quality and a pronunciation close to an expected citation form. If a selected token did not meet these criteria, another token was chosen. For example, the selected production of the word *disagree* by FNS3 had duration measurements closest to the mean production values. However, rather than beginning with a clear stop consonant [d], it began with some frication. For that reason, the production that was the second closest to the mean and began with a clear stop [d] was chosen instead. Similarly, care was taken to choose a production that did not have any creaky voice or unusual voice characteristics as manipulations of such tokens led to stimuli that sounded less natural than hoped.

To summarize, the speakers were chosen for their proximity to the mean of the productions for that gender of speaker across the words and individual tokens for their proximity to the mean production and for their clearness of articulation.

3.5.4 Note on the words chosen for manipulations

Of the five words chosen for manipulation and resynthesis, three were the words most often pronounced incorrectly by the ESL speakers in the production study. These words are *bipolar, develop,* and *lunatic.* The ESL speakers tended to stress the final syllable (rather than the middle syllable) in *bipolar* and *develop* and the second syllable (rather than the first) in *lunatic.* The words *October* and *disagree* were only rarely mispronounced by the ESL speakers. *October* was produced incorrectly only twice by male ESL speakers who placed stress on the first syllable. *Disagree* was pronounced incorrectly by a female ESL speaker who placed stress on the second syllable.

Whereas the bursts after final stops of *develop* and *lunatic* were not included in the duration of the tokens when taking measurements in the production study, they were included in the manipulations for a more complete and natural sounding production.

It is also worth noting that there were other words that were sometimes judged to be stressed incorrectly by the ESL speakers in that they had too much stress on one of the syllables (more than was deemed typical or acceptable). These include *introduce, unafraid* and *engineer*. However, as the last two words were perceived as having too much stress on the correct syllable, it was felt that manipulating and resynthesizing these words would not be in line with the main research question, which addresses what happens when a speaker produces stress on the *incorrect* syllable or does not produce any clear stress pattern at all. The idea of studying what happens when there is a perception of too much stress on the correct syllable is certainly a topic for further investigation.

3.6 Manipulations

Once the speakers and target words for manipulation were selected, the manipulation and resynthesis of each target word was done. Only the speech of native English speakers was used as the basis for the manipulations because only suprasegmental aspects of lexical stress are under investigation. That is, it is critical that the segmental (e.g. consonant and vowel) quality of the target words be pronounced in a native manner and that no segmental errors were present. By removing considerations of segmental quality, one can determine if, or to what extent, the suprasegmental aspects of lexical stress alone affect the perception of intelligibility and nativeness.

The manipulations performed in Praat altered the original values of vowel duration, F0 and RMS intensity and substituted values that mimicked the productions of the ESL speakers. The tokens with the modified values were then resynthesized in order to generate .wav files that could be presented to listeners for evaluation (see section 3.8 below). Vowel duration and F0 of the manipulated tokens were resynthesized using the PSOLA algorithm (Moulines & Charpentier, 1990) in Praat's Manipulation command, and intensity was modified using Praat's Intensity Multiplication command.

3.6.1 Ranges for manipulation

The measurements from the production study were used to determine a range of values to be used in the manipulations that were performed on the stimuli. The values chosen fell within the maximum/minimum values of productions of both male and female speakers. Values toward the extreme ends of variation were chosen to help ensure that some differences would be noticeable by listeners in the perception study. For duration, the values for manipulations ranged from an increase that was slightly greater than the average of the productions of the native English speakers (but within the range of values they produced), and a decrease that was slightly less extreme than what was found commonly in ESL speakers' error productions. For intensity, the value was chosen based on the most negative difference between the SV and UV by the ESL speakers. For F0, the value chosen was close to the overall average increase in F0 on the SV by the native speakers. The values for these three variables and there corresponding conditions are given in Table 3.3 below.

3.6.2 Types of manipulations

The target words were manipulated in three ways, referred to as "increase," "flat," and "decrease" versions. Following the methodology used in labeling the vowels in the production study, the vowel that should carry lexical stress and the unstressed vowel immediately following it were the targets for modification. In the case of a word with stress on the final syllable, e.g. *disagree*, the unstressed vowel for manipulation was the one preceding the final stressed vowel. In the case of *October,* both the first and second vowels were modified because the production of excessively strong stress on the first syllable by one ESL speaker prompted the judges to rate the stress production as incorrect.

In creating the modified versions of the target words, duration ratios, intensity and F0 were manipulated one at a time, with the other two dimensions maintained as originally produced in the token that was being modified until all manipulations on that token were carried out. This resulted in seven different modified versions of each target word. These are shown in Table 3.3 below.

Table 3.3: Naming key for modified versions of one target word

Condition	Description	Manipulation values
Flat	all flat	SV=UV for duration ratio, intensity, F0
IncDur	increase duration ratio	SV:UV 2.5:1
DecDur	decrease duration ratio	SV:UV 1:2.5
IncInt	increase intensity	SV 6dB > UV
DecInt	decrease intensity	SV 6dB < UV
IncF0	increase F0	SV 20Hz > UV
DecF0	decrease F0	SV 20Hz < UV

3.6.3 Particulars of manipulations

Because only one acoustic dimension was varied in each modified version, the other two dimensions retained their original values until the entire series of manipulations had been carried out and the desired result was achieved. However, it was discovered that in Praat, manipulating the values for one dimension could result in unwanted changes to other dimensions. For example, in one trial, doubling the duration of a vowel reduced the intensity of that vowel by .3 dB. In order to correct for such changes, the order in which the manipulations were executed was kept fixed. First, the duration was modified, then pitch, and finally intensity. Intensity had to be modified last because the intensity values of the vowel are easily altered by changes in vowel duration and pitch. Once each manipulation was created, the output sound file was checked with its Text Grid to ensure the desired result was achieved.

3.7 Values for manipulations

The following three sections describe the parameters for the manipulations of the three study variables.

3.7.1 Manipulation of Duration

Duration was modified first. The start and end times for the SV and UV in the target word were noted, as were the durations of these vowels. As described above, duration was either held constant (the “flat,” model), increase, or decrease. The “increase” version had a duration ratio of 2.5:1 for the ratio of the SV to the UV. In other words, the SV was modified to be 2.5 times longer than the UV. This value is consistent with productions by the native English speakers. The “flat” model had a vowel duration ratio of 1:1, meaning that the SV and UV were set to be of the same length. This “flat” version was in most cases set to be equal to the mean of the durations of the SV and UV in the original, unmodified token; the average for these was 83 ms. The “decrease” duration version had a vowel duration ratio of 1:2.5. In other words, the UV was 2.5 times longer than the SV, a value typical of productions in which the ESL speakers produced lexical stress on the vowel that was supposed to be unstressed. Typical values for the versions with increased and decreased duration were 150ms for one vowel and 60ms for the other. Praat scripts were used to generate a new sound file for each modified version of each of the five selected words for each type of manipulation (i.e. duration, intensity, F0). A script is advantageous to doing the manipulations "by hand" as the process is streamlined and the possibility of introducing errors into the manipulations is reduced.

3.7.2 Manipulation of Fundamental Frequency

Fundamental frequency was modified next. A Praat script was used to modify the SV and UV in three different ways following the same pattern as for duration. The pitch range was set to 60-250 Hz for males and 75-300 Hz for females. Although manipulating F0 in Praat can give the impression of a somewhat tinny sound, the end-product of a decreased F0 tended to sound better than an increased F0. Target output for the manipulations of F0 and intensity were mean values of the vowels since the manipulation affects the entire vowel and not just the midpoint.

In the "increase" version, the F0 of the SV averaged 20 Hz higher than the mean F0 of the UV. The value of 20Hz is consistent with average increases in F0 by native English speakers on SVs in the production study. The "flat" version had flat F0, that is, the values for the SV and UV were equal. In the "decrease" version the F0 of the SV was 20 Hz lower than that of the UV. This value was selected to mirror that of the "increase" version, and also because it equaled the average value produced by the ESL speakers when they produced stress on the incorrect syllable. However, speakers produced changes in F0 that ranged from +/- 100Hz for one vowel compared to the other in the same token.

3.7.3 Manipulation of Intensity

As mentioned previously, the manipulation of intensity was necessarily done last because it was found that modifying either duration or F0 would also create changes in the intensity of the vowels, while the opposite was not the case. Modification of intensity was also carried out via a script in Praat. As for duration and F0, the SV and UV were modified to create three versions, the "increase", "flat", and "decrease" versions. The "increase" version had an SV that is 6dB louder than the UV. This value was chosen to be consistent with the measured increases in intensity on stressed vowels produced by native English speakers; another consideration was the findings of Okobi (2006) and Fry (1955) for changes in intensity that are perceptually salient in English stress. The "flat" version had no difference in intensity between the SV and UV, and the SV in the "decrease" version was 6dB quieter than the UV. This value was similar to the measured values for tokens with misplaced stress produced by ESL speakers. Finally, after all the manipulations were completed, it was necessary to scale down the intensity of the entire token, as the intensity manipulation algorithm in Praat tends to increase the overall intensity of the word. Each token was scaled to 65dB, an intensity level that is typical for conversational speech. This was done using the 'Scale Intensity' command in Praat.

Part III: Intelligibility and Nativeness Listening Study

3.8 Overview

In the final phase of the research, listener judgments of the perceived intelligibility and nativeness of the resynthesized tokens were obtained. The stimuli for this part of the study were the resynthesized tokens created as described in the previous section.

3.8.1 Listeners

Each participant completed a background questionnaire before beginning the experiment (see Appendix L). Participants in the listening tasks were native speakers of American English (n=21; mean age 32), mainly humanities undergraduate and graduate students at the University of New Mexico. 10 males and 11 females participated, though one of the female participants had to be excluded from the data analysis as will be explained in Chapter 5. The participants ranged in age from 21 to 53 years old and had lived in New Mexico between 3 months and 29 years. Most participants spoke at least one other language, including Spanish, German, and Russian. The listener demographics are given in Appendix M.

3.8.2 Rating Scale

There are a variety of tasks and scales that can be used to judge speech intelligibility and nativeness. The equally-appearing interval (Likert) scale, and direct magnitude estimation are two such methods that are popular in voice perception research (Kreiman et al., 1993). A variation of the Likert scale is an unmarked visual analog scale that is said to be more robust than the equally-appearing interval scalar method (cf. Schiavetti, 1992). The unmarked visual analog scale was chosen for use in the present research as it allows the rater to choose any spot along a continuum and thus encourages a rater to use a larger proportion of the scale. Kreiman et al. (1993) found that listener performance was quite similar on equally-appearing interval and visual analog tasks though ratings on the visual analog scale tended to drift less during the experiment than did ratings on the equally-appearing interval scale. Therefore, Kreiman et al. state that visual analog scales are more reliable than equally-appearing interval scales. Though the scale used here is unmarked, its range was treated as equal to 700 points in order to be able to analyze the results as if they were on a 7-point scale. A sample of the unmarked scale used in this experiment is provided in Appendix N.

3.8.3 Task

The listeners' task was to listen to the resynthesized tokens and rate them for either intelligibility or nativeness. The stimuli were presented to them by the computer program Alvin (Hillenbrand & Gayvert, 2009) which was also used to collect their rating responses.

Before beginning, the listeners completed an Informed Consent Form, approved by the university's Institutional Review Board, which explained that the research addresses how differing patterns of pronunciation could have an effect on native English listeners' perception of speech intelligibility and nativeness. Next, they completed a practice session so that they would become accustomed to the type of stimuli that would be presented in the actual experiment. In this practice session, eight tokens illustrating the variety of manipulations included in the actual experiment were presented to the listeners. This also enabled them to become accustomed to the types of words they would be hearing and to practice responding on the unmarked scale. The ratings of the tokens used for practice were not included in the data analysis.

Half of the listeners completed the intelligibility task first and the other half did the nativeness task first. The stimuli were presented at a comfortable volume, using Sony Studio Monitor headphones. The listeners were told that they would only hear five different words: *bipolar, develop, disagree, lunatic,* and *October.* They were also told to assume that the word was said in a typical declarative sentence without carrying any particular emphasis or having any stylistic effect. They were told that they could listen to a token a second time by clicking the "Repeat" icon on the screen, and that they should click "Okay" to continue to the next token. If a mistake was made (e.g. the listener clicked "Okay" before the desired selection was made), they could click "Back up" to return to the previous screen and enter the rating. The computer program randomized the tokens before each participant began. Therefore, the stimuli were presented in different orders for the intelligibility and nativeness tasks, and the stimuli were in differing orders for all listeners.

The listeners were instructed to select a point along the scale according to their judgment of the token, from "not intelligible/native at all" to "very intelligible/native." Following the methodology of Kreiman et al. (1993), they were instructed not to base the rating of the current token on any previous one and were encouraged to use as much of the scale as they saw fit. The listeners were also told that once a token had been rated, they were not to go back to make changes to their rating. They were allowed to work through the task at their own pace. The entire experiment took approximately 30 minutes to complete. Listeners were paid $10 for their participation.

Twenty percent of the tokens (n=14) were repeated in each experiment to test for intra-rater reliability. These were equally divided between the voices of the male and female speakers, and included one manipulation of each type for each speaker. The 14 tokens repeated in the intelligibility task were different from the 14 repeated in the nativeness task and were not included in the analysis of the results of the experiment (i.e.

they were only used to test for intra-rater reliability). Thus, a total of 40% of the data was repeated in one of the two experiments.

3.9 Test Statistics for Listening Study

The listeners' responses were imported into SPSS (SPSS, 2009) for analysis. The dependent variables are the intelligibility and nativeness ratings for each token. The independent variables are the seven different types of manipulations to duration, intensity, and F0. The data were analyzed in a one-way repeated measures ANOVA in which the independent variable (the type of manipulation) had 7 different levels. Post-hoc Bonferroni corrected pairwise comparisons were also performed.

Chapter 4: Results and Discussion from the Production Study

4.0 Overview

In this chapter, the realizations of the acoustic correlates of lexical stress by the native English speakers and the ESL speakers are compared. Out of the speech material recorded for this project, as described in Chapter 3, only the analysis of the three-syllable words is reported here. This amounts to a total of 1,728 tokens (12 words x 6 repetitions x 24 speakers) of these words as produced by twenty-four speakers (12 males and 12 females in each language group). The twelve target words that were analyzed are listed in Table 4.1 with their phonetic transcription in English and Spanish. The sentences in which these target words were embedded are given in Appendices A and B.

Table 4.1: 3-syllable target words with stress pattern and cognate status (English / Spanish)

Same-stress cognates (SSC)	**Different-stress cognates (DSC)**	**Non-cognates (NC)**
radio ['ɹeɪ-di-oʊ / 'ra-di-o]	lunatic ['lu-nə-tɪk / lu-'na-ti-ko]	easily ['i-zə-li]
October [ɑk-ˈtoʊ-bɚ / ok-ˈtu-bre]	bipolar [baɪ-'poʊ-lɚ / bi-po-'lar]	management ['mæ-nədʒ-mənt]
	engineer [en-dʒə-'niɹ / in-hen-i-'e-ɾo]	develop [dɪ-'vɛ-ləp]
	introduce [ɪn-tɹə-'dus / in-tro-du-'sir]	location [loʊ-'keɪ-ʃən]
		disagree [dɪs-ə-'gɹi]
		unafraid [ʌn-ə-'fɹeɪd]

As described in Chapter 3, two vowels were labeled in each word. The vowel in the syllable that is supposed to receive lexical stress, according to the American Heritage

Dictionary (Costello, 1992), was labeled SV. The vowel in the following syllable, that is marked as unstressed in the dictionary, was labeled UV. In the case of a word with final-syllable stress, the unstressed syllable preceding the SV was labeled as the UV. Three derived measures were calculated for each token: the ratio of the duration of the SV to the duration of the UV, referred to as the “vowel duration ratio,” the difference in intensity between the SV and the UV, and the difference in F0 between the SV and the UV. An average of these acoustic measurements over the six repetitions of each word was calculated by speaker, and then these values were averaged across the 12 speakers for each speaker group. These averages were used to produce the graphs in this chapter and subsequently to calculate the statistics for each variable. As the primary focus of this chapter lies in the examination of the ESL speakers’ production of English lexical stress, the ESL speakers’ productions are analyzed in light of the values of the native English speakers. In other words, the productions of the native English speakers provide the baseline for this comparison.

The results address two main questions: First, how does the production of lexical stress in English words compare between native Spanish-speaking ESL speakers and native English speakers? This question was investigated by testing whether differences exist in terms of vowel duration ratios, intensity and F0 patterns on the SVs and UVs between these groups of speakers. Secondly, does the cognate status of a word affect a Spanish ESL speaker’s production of lexical stress? This question was addressed by examining the ESL speakers’ production of English words to determine if they produced more variability and differences in cognate words whose stress patterns differ between Spanish and English than in cognates with the same stress pattern in both languages.

4.1 Organization of this chapter

The results are divided into four main sections as follows: 1) Analysis of all productions of the twelve target words (referred to as the "12-word set"). This analysis includes productions in which the ESL speakers at times placed stress on the unexpected syllable; 2) Analysis of a subset of the target words in which the native English speakers consistently increased F0 on the SV (referred to as the "6-word set"). An explanation of why this analysis was carried out is given in section 4.3.3 of this chapter; 3) Analysis of a subset of productions in the 12-word set in which only those productions by ESL speakers that were judged to have primary stress on the syllable that was expected to receive lexical stress (see section 6.1 of Chapter 3) were included; 4) Analysis of a subset of productions in the 6-word set that included only those productions by ESL speakers that were judged to have primary stress on the syllable that was expected to receive lexical stress.

Results for each section (e.g. 12-word set, 6-word set, etc.) are presented in the following order: 1) vowel duration ratio, 2) change in intensity, 3) change in F0. A short synopsis of each result is presented after its graph and statistics. A general discussion highlighting similarities and differences among the analyses follows at the end of the chapter.

4.2 Range of variation in productions

As mentioned above in the Overview, the results reported here are based mainly on average values of productions by the native English and ESL speakers. These average values are useful for comparing the major differences between the two speaker groups. A disadvantage of this technique is that it does not account for variability in the productions. This variability is important in particular because it was used to inform the choice of values for the manipulation of the tokens used in the subsequent listening experiment. Table 4.2 below displays the range of values produced by speakers in each group for each of the three derived variables used in this study.

Table 4.2: Range of variation in productions

	Native speakers		**ESL speakers**	
	low	high	low	high
Vowel Duration Ratio (SV:UV)	0.9	5.1	0.3	5.1
Intensity (SV-UV) (dB)	-3	13.1	-12.4	15.4
F0 (SV-UV) (Hz)	-111	103	-118	131

In interpreting this table, note that a vowel duration ratio greater than 1 means that the SV was longer than the UV, a positive value for intensity means that the SV was louder than the UV, and a positive value for F0 means that the SV had a higher F0 than the UV. This table shows that the ESL speakers produced a wider range of variation than the native speakers in the three study variables. For vowel duration ratio, both speaker groups had the same maximum value (5.1) meaning that, on at least one occurrence for each group, the SV was produced just over five times longer than the UV. The minimum values differed: the lowest vowel duration ratio for a native speaker was .9, meaning that the UV was 1.1 times longer than the SV, whereas the lowest value for an ESL speaker was .3, corresponding to a production of the UV that was 3.3 times longer than the SV. The findings here are similar for each of the study variables in the sense that the ESL speakers produced a wider range of variation between the SV and UV than did the native speakers. For both speaker groups, the production values observed in the stressed and unstressed syllables were not always what one might expect would to occur. The variation in this table will be discussed further throughout this chapter. The next section reports the overall results of the study, beginning with the complete 12-word set.

4.3 Results for 12-word set

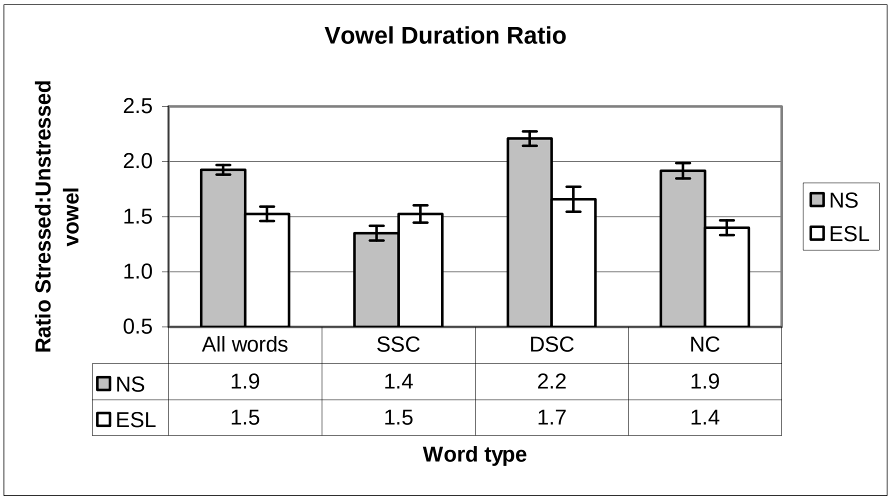
The results of the 12-word set are comprehensive as they include all productions (except for those excluded due to absence of pitch accent, etc., as described in sections 3.2.2 and 3.4.2 of Chapter 3), including productions by the ESL speakers that were judged to have lexical stress on the unexpected syllable. Results for vowel duration ratios, change in intensity and change in F0 are discussed here in turn. The statistical output from SPSS for all calculations in this chapter is provided in Appendix I (parametric statistics) and Appendix J (non-parametric statistics).

4.3.1 Vowel Duration Ratio

Figure 4.1 displays the ratio of the duration of the predicted stressed vowel (SV) to the duration of the predicted unstressed vowel (UV), referred to as SV:UV. For example, a ratio of 2:1 (a value of 2 on the y-axis) means that the vowel in the stressed syllable is twice as long as the vowel in the unstressed syllable. A ratio of 1:1 (a value of 1 on the y-axis) indicates that the predicted SV and predicted UV are of the same duration. A ratio such as 1:2 would indicate that the predicted UV is twice as long as the predicted SV, contrary to expectation. It is important to remember that these ratios are comparisons of the syllable which should receive stress according to the dictionary and a syllable that is expected to be unstressed.

In the following graphs, NS refers to the “native speaker” group and ESL refers to the “English as a Second Language” group. Furthermore, “All words” refers to all target words combined while SSC refers to "same-stress cognates," DSC to "different-stress cognates," and NC to "non-cognates." Standard error bars are also shown.

Figure 4.1: Vowel Duration Ratio



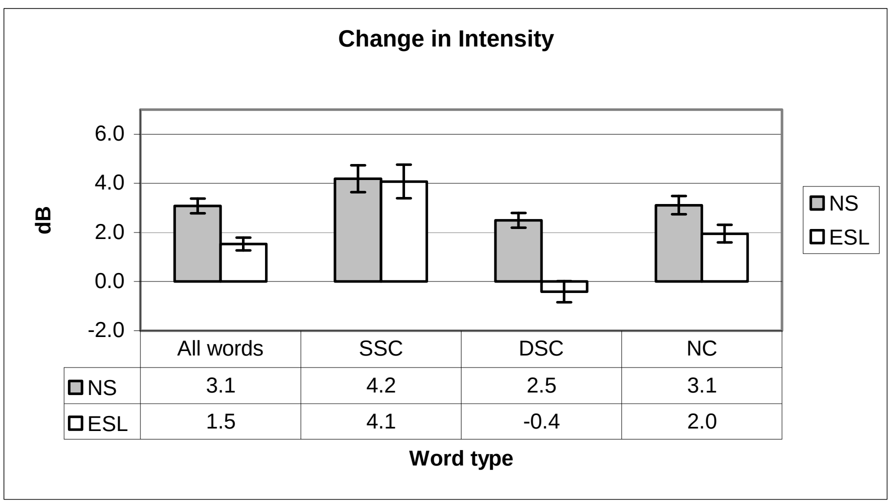
All of the values plotted are greater than 1.0 indicating that, on average, both groups of speakers produced the SV longer than the UV for all word types. Native speakers produced statistically greater vowel duration ratios than the ESL speakers for all categories except same-stress cognates: All words, $t(22) = 5.06$, $p < .0005$, DSC, $t(22) = 4.168$, $p < .0005$, NC, $t(22) = 5.338$, $p < .0005$. In other words, the native speakers differentiated the durations of the SV and the UV more than the ESL speakers for these word types. The difference between vowel duration ratios for same-stress cognates (SSC) was not significant ($p > .05$). As the data in the 'All words' and DSC sets were not normally distributed, they were also submitted to a non-parametric independent samples Mann-Whitney U test. The results of the non-parametric test also show that the differences between speaker groups are significant: All words: U(22) = -3.388, $p <.01$; DSC: U(22) = -2.926, $p < .01$.

These results demonstrate that overall (the ‘All words’ category), native speakers produced vowel duration ratios significantly greater than those produced by the ESL speakers. Regarding cognate status, we see that the same-stress cognates were produced nearly identically by the two groups while the different-stress cognates and non-cognates were not.

4.3.2 Change in Intensity

In the graphs for intensity (measured in dB), a value of 0 on the y-axis represents no difference in intensity between the SV and the UV. A positive value means that the SV was louder than the UV, and negative values mean that the UV was louder than the SV.

Figure 4.2: Change in Intensity



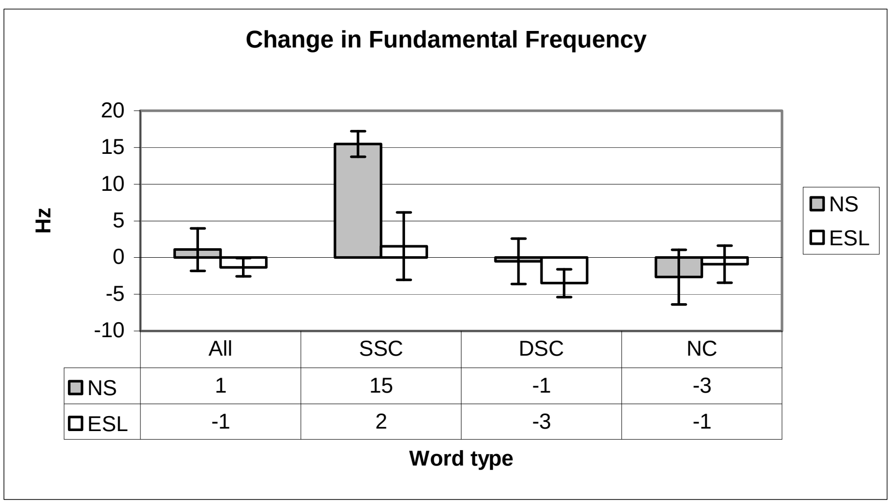
There was a significant difference for: All words, $t(22) = 3.921$, $p < .01$, DSC, $t(22) = 5.526$, $p < .0005$, NC, $t(22) = 2.250$, $p < .05$ with native speakers having a greater change in intensity between SVs and UVs than ESL speakers. The change in intensity between the SV and UV in same-stress cognates (SSC) was not significant ($p > .05$).

Overall, the native speakers increased intensity just over twice as much as the ESL speakers for ‘All words’ (3.1 dB vs. 1.5 dB). Regarding cognate status, DSCs show the greatest divergence between the speaker groups. The negative value for this category for the ESL speakers shows that they often placed more intensity on the UV than the SV. And like vowel duration ratio, NCs were produced with significantly less intensity on the SV by ESL speakers than by the native speakers.

4.3.3 Change in Fundamental Frequency

For fundamental frequency (measured in Hz), a value of 0 on the y-axis of the graphs means that the SV and the UV were produced at the same F0. A positive value means that the SV had higher F0 than the UV, and negative values show that the UV had higher F0 than the SV.

Figure 4.3: Change in Fundamental Frequency



There was a significant difference for same-stress cognates (SSC) $t(22) = 2.834$, $p < .05$ with native speakers producing a greater difference in F0 between the SV and UV than the ESL speakers. The difference in fundamental frequency between the SV and the UV was not significant for any other word types ($p > .05$). As NCs were not normally distributed the Mann-Whitney U statistic was calculated, which also returned a non-significant value ($p > .05$).

To summarize to this point, in this 12-word set the ESL speakers' productions were most similar to the native speakers' in the SSCs, while the NCs and DSC were produced with the most divergent patterns.

Figure 4.3 shows that there was virtually no difference between the speaker groups for the production of F0 for 'All words.' In fact, with the exception of the words *radio* and *October* in the SSC group as produced by the native speakers, neither group of speakers changed F0 much between the SV and UV. However, the overall averages of F0 hide some important information about changes in F0 that were realized by the two speaker groups. Table 4.2 above demonstrated that changes in F0 by both groups of speakers could be quite large. Further analysis revealed that 66% (567/864 tokens) of the data for the native speakers and 57% of data (493/864 tokens) for the ESL speakers had a difference of more than 5 Hz between the SV and the UV. More specifically, for 40% of the native speaker data, the F0 of the SV was more than 5 Hz greater than the F0 of the UV, and for 26% of the data the F0 of the SV was more than 5Hz less than the F0 of the UV. For the ESL speakers, the F0 of the SV was more than 5 Hz greater than the F0 of the UV in 23% of the data, and in 34% of the data the F0 of the SV was more than 5 Hz less than the F0 of the UV. In other words, the speakers did produce large changes in F0 between the SV and UV in many instances but because some of the changes were in one direction and some in the other, their average was close to 0 when combined as shown in the plot of 'All words' in Figure 4.3.

While the native speakers consistently produced the SV as longer (having a greater vowel duration ratio) and louder (having a positive change in intensity), they did not exhibit a consistent pattern for F0. The reason for this inconsistency is likely that different F0 patterns are acceptable in English. A prominent word can be marked as such by either a Low or High pitch accent resulting in a decrease or increase in F0 (Wennerstrom, 2001), and the speakers varied in their choice of pitch accent. This variation makes comparing the productions of the ESL speakers to the native speakers very difficult because of the inconsistency of the native speakers. For this reason, the data were further sorted by whether F0 was higher on the SV compared to the UV, nearly the same, or lower. This analysis revealed that for a subset of the 12 target words, the native speakers almost always produced a higher F0 on the SV than the UV. This subset included two words of each cognate type (SSC, DSC, NC) which were selected to yield a 6-word subset.

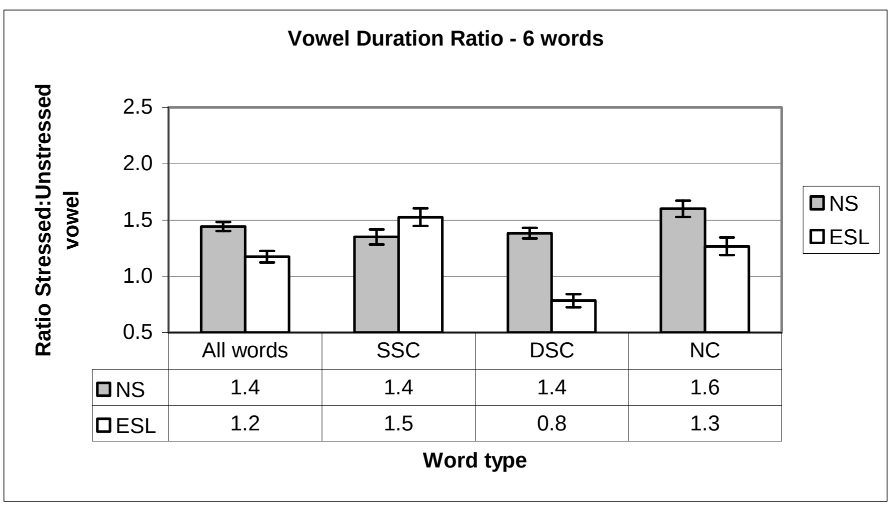
For the following six words, the percentage of tokens produced by native speakers in which the SV was at least 5 Hz higher than the UV was: for SSCs: *radio* (69%), *October* (81%); DSCs: *bipolar* (48%), *lunatic* (79%); NCs: *location* (60%), *management* (55%). When native speakers did not increase F0 by more than 5Hz on these words, they typically produced them so that the F0 of the SV and UV differed by less than 5 Hz. A cut-off point of +/- 5 Hz was chosen as 5 Hz is twice the just noticeable difference for fundamental frequency of complex signals in quiet listening conditions (Moore, 2003). There were very few tokens in which the native English speakers produced these words with the F0 of the UV more than 5 Hz higher than the SV: SSCs: *radio* (6%), *October* (3%); DSC: *bipolar* (8%), *lunatic* (6%); NC: *location* (8%), *management* (3%). Clearly, the native English speakers were more consistent in their productions of these words, making these words good candidates for forming a comparison with the ESL speakers. The distribution of F0 movement by word for the native speakers can be found in Appendix K. Note that the calculations given above represent the total number of tokens per category (e.g. number of tokens that were “at least 5 Hz higher than the UV”) divided by the total number of repetitions of each word minus the exclusions.

The next section reports on the analysis of this 6-word set that is characterized by an increase in F0 on the SV for native speakers. The results are again presented in the order of vowel duration ratio, change in intensity, and change in F0.

4.4 Results for 6-word set with trend for increase in F0 on SV by native speakers

4.4.1 Vowel Duration Ratio

Figure 4.4: Vowel Duration Ratio. 6-word set with a reliable increase in F0 on the SV by native speakers

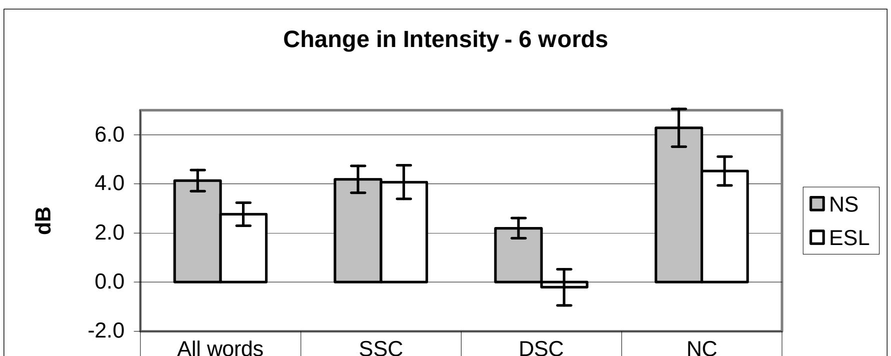


There was a significant difference in vowel duration ratio for: All words, $t(22) = 4.125$, $p < .0005$, DSC, $t(22) = 8.050$, $p < .0005$, NC, $t(22) = 3.120$, $p < .01$ with native speakers having greater vowel duration ratios than the ESL speakers. The difference between vowel duration ratios for SSCs was not significant (p > .05). A Mann-Whitney U test was calculated for the DSCs and NCs and confirmed that there was a significant difference for: DSC: U(22) = -4.184, p < .0005 and NC: U(22) = -3.042, p < .01.

The results for this set of words mirrors the results of the 12-word set. For 'All words,' there is a significant difference in vowel duration ratio between the two speaker groups. Furthermore, the cognate status of the word plays a role in the production of vowel duration by the ESL speakers. For SSCs, the ESL speakers produce vowel durations in a manner statistically identical to the native speakers. However, DSCs and NCs are not produced in such a similar manner.

4.4.2 Change in Intensity

Figure 4.5: Change in Intensity. 6-word set with a reliable increase in F0 on the SV by native speakers



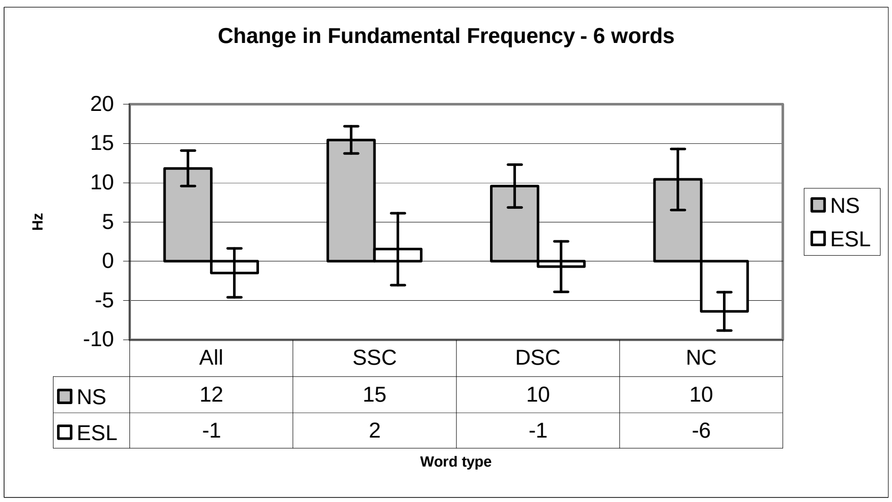
	All words	SSC	DSC	NC
NS	4.1	4.2	2.2	6.3
ESL	2.8	4.1	-0.2	4.5

In this analysis, there was a significant difference for: All words, $t(22) = 2.142$, p < .05 and DSC, $t(22) = 2.850$, p < .01 with native speakers producing greater changes in intensity between the SV and UV than the ESL speakers in these word types. The difference between SSCs and NCs was not significant (p > .05).

As for ‘All words’ and DSCs in the 12-word set, the difference in intensity between the SV and UV in these two word types is statistically significant in this 6-word analysis. It is interesting to note that in addition to being statistically similar to the native speakers on the SSCs, the ESL speakers are also statistically similar in production to the native speakers on the NCs.

4.4.3 Change in Fundamental Frequency

Figure 4.6: Change in Fundamental Frequency. 6-word set with a reliable increase in F0 on the SV by native speakers



A significant difference was returned for all word types in this analysis: All words, $t(22) = 3.466$, $p < .005$, SSC, $t(22) = 2.834$, $p < .05$, DSC, $t(22) = 2.434$, $p < .05$, NC, $t(22) = 3.642$, $p < .01$ with the native speakers producing greater changes in F0 between the SV and UV than the ESL speakers.

The principal purpose of conducting the analysis of this 6-word set was to investigate if the ESL speakers produced a change in F0 similarly to the native speakers when a clear baseline in the production of F0 by the native speakers could be established. Interestingly, this analysis shows that the ESL speakers did not produce a higher value for fundamental frequency on the SV as the native speakers did. In fact, each word type shows a significant difference between the productions of the two groups and that the ESL speakers produced little to no change in F0 between the SV and UV on these words.

To summarize the results of this section, the SSCs were produced most similarly between the speaker groups for vowel duration ratio and intensity. And with the exception of intensity for NCs, the NCs and DSCs were again produced with the greatest amount of divergence.

Significant differences were found between the production of vowel duration ratios, change in intensity and change in fundamental frequency of the two speaker groups in the previous two sections (i.e. the 12- and 6-word sets). Thus, the question could logically be asked if the differences simply exist because the ESL speakers at times mispronounced the words, and the differences found are due to placing lexical stress on an unexpected syllable. In other words, would the values of the acoustic correlates of lexical stress of these highly proficient ESL speakers be different from those of the native speakers if the ESL speakers' productions with lexical stress on the unexpected syllable were excluded from the analyses? This question is the focus of the next two sections.

4.5 Comparison of productions with correct stress placement by the ESL speakers

In order to select the data to be used for analyzing the ESL speakers' productions with "correct" stress placement, the researcher listened to all 864 productions of the target words by the ESL speakers and judged if the word was produced with the correct stress pattern (for more complete detail, see Chapter 3, section 3.4.1). Two professional phoneticians checked subsets of the data and inter-rater reliability was high. The judgments of the researcher agreed with those of one professional phonetician 97% of the time and with the other professional 100% of the time. The tokens that were judged to have incorrect stress patterns are summarized in Table 4.3 below.

Table 4.3: Incorrect productions of target words by the ESL speakers

word	**#incorrect**	**% total data**	**category**
bipolar	17	2.0	DSC
develop	19	2.2	NC
disagree	1	0.1	NC
introduce	2	0.2	DSC
lunatic	57	6.6	DSC
October	1	0.1	SSC
total	**97**	**11.2**	

Pronunciations with lexical stress on the incorrect syllable by the ESL speakers represent 11.2% (97/864 tokens) of the total data. Overall, only 6 of the 12 target words were ever pronounced with lexical stress on the unexpected syllable by the ESL speakers. Furthermore, the ESL speakers only had significant difficulty with three of these words, all of which were different-stress cognates. The word *lunatic* (representing 57 of 97 incorrect productions, or 59%) was the word most frequently mispronounced, followed by *develop* (20%) and *bipolar* (18%). To investigate the cause of these mispronunciations, reported usage of these words by the ESL speakers was analyzed as explained below.

As explained in Chapter 3, the ESL speakers rated each of the target words for frequency of usage on a scale of 1-4 at the completion of the production task. These ratings were requested to examine if correct productions correlated with those words used most frequently by the ESL speakers, or conversely, if words most often mispronounced correlated with those reported as being used least frequently by the ESL speakers. The four-point scale consisted of: 1: rarely or never, 2: once in a while (e.g. 2 or 3 times a month), 3: weekly (e.g. 2 or 3 times a week), and 4: daily. A summary of the ESL speakers' averaged ratings of the target words is given in Table 4.4 below. The complete rating sheet is attached in Appendix H.

Table 4.4: Frequency of usage of target words by the ESL speakers (out of 4 points)

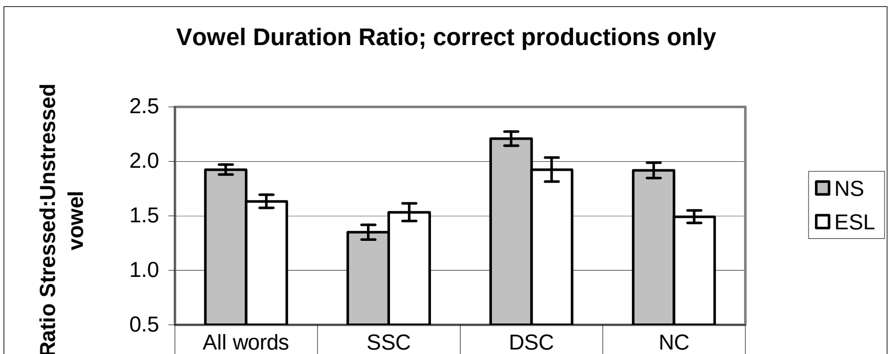
word	**frequency of usage**	**word type**
easily	3.6	NC
disagree	3.5	NC
radio	3.5	SSC
develop	3.3	NC
introduce	3.3	DSC
October	3.3	SSC
location	3.2	NC
engineer	2.8	DSC
management	2.6	NC
lunatic	1.4	DSC
bipolar	1.3	DSC
unafraid	1.3	NC

Table 4.4 suggests that frequency of usage of a particular target word explains mispronunciations of target words to an extent. For example, the word *lunatic*, the most frequently mispronounced word accounting for 59% (57/97 tokens) of the mispronunciations was one of the least commonly used words. The ESL speakers rated this word in terms of frequency of usage as a 1.4 out of 4. The same is true for the word *bipolar*, which accounted for 18% of the incorrect productions and was rated as a 1.3 for usage. Note that both *lunatic* and *bipolar* are DSCs. However, the word *develop*, a NC which accounted for 20% of the incorrect productions, was one of the words the ESL group overall said they used most frequently in their speaking. It was rated as a 3.3 on the 4-point scale and reported as being used less frequently than only the words *easily, disagree,* and *radio*. These results indicate that some of the ESL speakers produced words with stress on an unexpected syllable even though they use the word frequently.

The following analyses were carried out after excluding all incorrect productions from the data set and thus compare only the correct productions of the ESL speakers to the productions of the native English speakers. These analyses include the same data from the native speakers as the previous analyses as the native speakers did not place lexical stress on an unexpected syllable in the production study.

4.5.1 Vowel Duration Ratio

Figure 4.7: Vowel Duration Ratio, ESL speakers' correct productions only



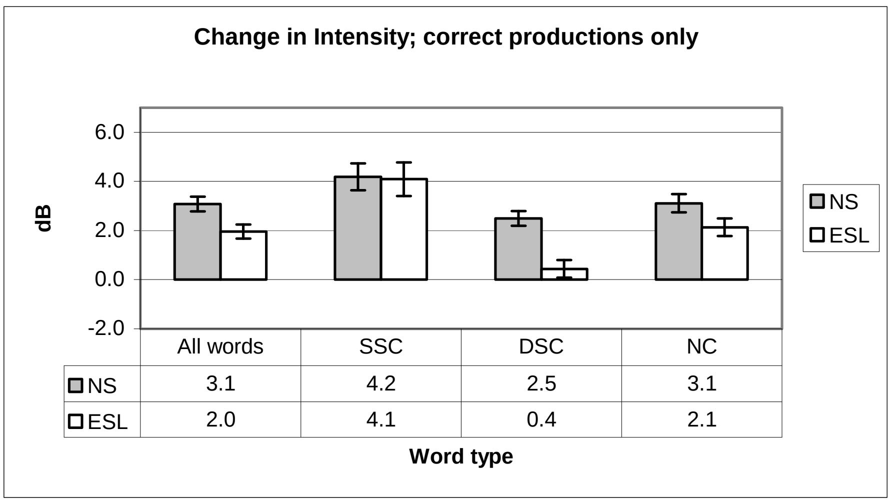
	All words	SSC	DSC	NC
NS	1.9	1.4	2.2	1.9
ESL	1.6	1.5	1.9	1.5

A significant difference in the production of vowel duration ratios was returned in this analysis for: All words, $t(22) = 3.924$, $p < .01$, DSC, $t(22) = 2.219$, $p < .05$, NC, $t(22) = 4.686$, $p < .0005$ with native speakers having greater SV:UV vowel duration ratios than the ESL speakers. The difference between vowel duration ratios SSCs was not significant ($p > .05$). As the 'All words' set was not normally distributed, a Mann-Whitney U test was performed. The non-parametric statistic confirmed that the difference in the 'All word' category was significant: U(22) = -2.961, $p < .01$.

This result mirrors the one in the 12-word set in which all productions (correct and incorrect) were included. Overall, the ESL speakers' production of vowel duration ratios for ‘All words’ was not statistically similar to that of the native speakers. Considering cognate status, the SSCs were again produced by the ESL speakers in a manner very similar to the native speakers. The DSC and NC categories for the ESL speakers again showed statistical differences between the two speaker groups, indicating that cognates influenced the ESL speakers’ production.

4.5.2 Change in Intensity

Figure 4.8: Change in Intensity, ESL speakers’ correct productions only

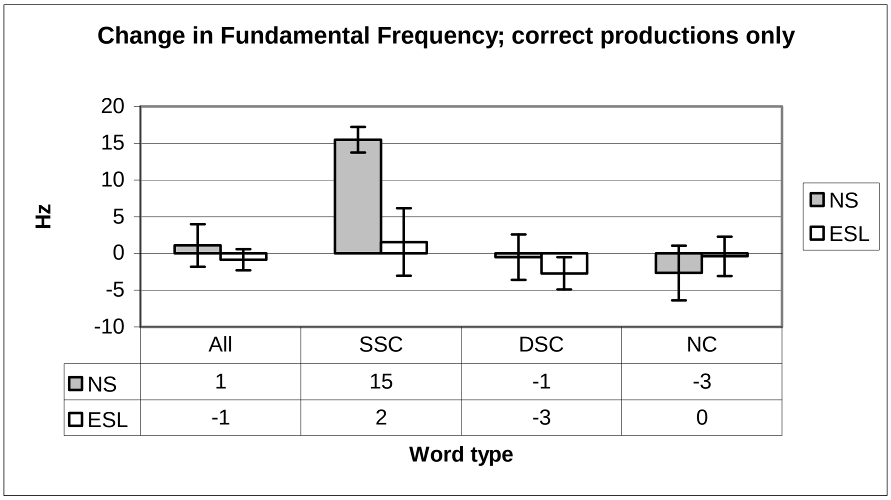


Considering only the correct productions in the data, there was a significant difference in change in intensity for: All words, $t(22) = 2.681$, $p < .05$ and DSC, $t(22) = 4.377$, $p < .0005$ with native speakers demonstrating greater changes in intensity between the SV and UV than the ESL speakers. The difference in intensity for SSCs and NCs was not significant ($p > .05$).

Only the DSCs were produced in a way that differed significantly between the two speaker groups, indicating that the DSCs accounted for a sufficiently large proportion of the data to make them statistically different from the DSCs as produced by the native speakers.

4.5.3 Change in Fundamental Frequency

Figure 4.9: Change in Fundamental Frequency, ESL speakers' correct productions only



There was a significant difference in change in F0 for: SSC, $t(22) = 2.834$, $p < .05$ with native speakers having a greater change in F0 between the SV and UV than the ESL speakers. Differences in F0 for the other word types were not significant ($p > .05$). A non-parametric Mann-Whitney U statistic was calculated for the DSC and NC sets as they were not normally distributed. The result of this test also was not significant ($p > .05$).

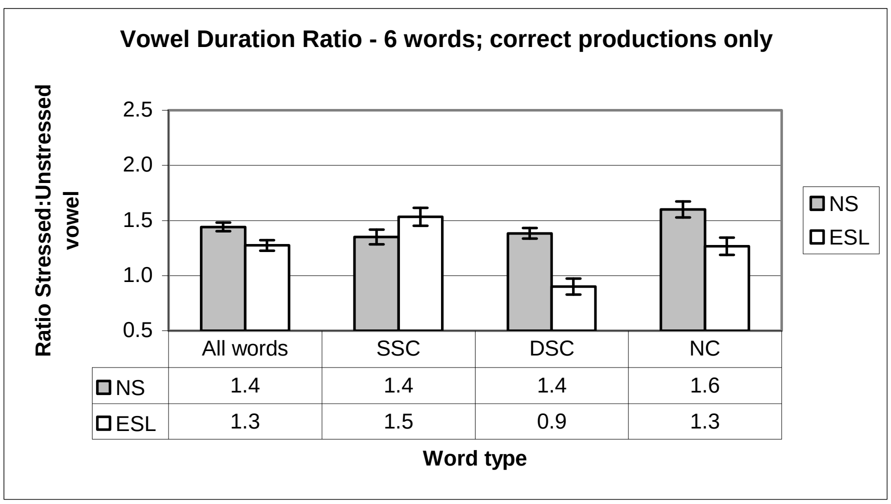
As seen in the 12-word set, the two speaker groups were identical on their production of F0 on all categories except SSCs. It seems that a lexical or sentence effect may be present as it is only in the SSCs that the native speakers tend to increase F0 on the SV consistently. However, it can again be seen that, for all words combined, there is almost no difference for either speaker group. As described before, this similarity is misleading because it was found that the speakers in fact did change F0 by large amounts between the SV and UV on different productions, but when the differences are averaged the mean is close to 0. Therefore, F0 is analyzed in the next section for the correct productions on the 6-word set with consistent F0 patterning by the native speakers.

4.6 Results for 6-word set, words with a reliable increase in F0 on SV by native speakers, ESL speakers' correct productions only

In this section, the productions of the 6-word set by both speaker groups is compared including only correct productions of the target words by the ESL speakers. As in the previous sections, results for vowel duration ratios are given first, followed by results for change in intensity and change in F0.

4.6.1 Vowel Duration Ratio

Figure 4.10: Vowel Duration Ratio. 6-word set with a reliable increase in F0 on the SV by native speakers, ESL speakers' correct productions only

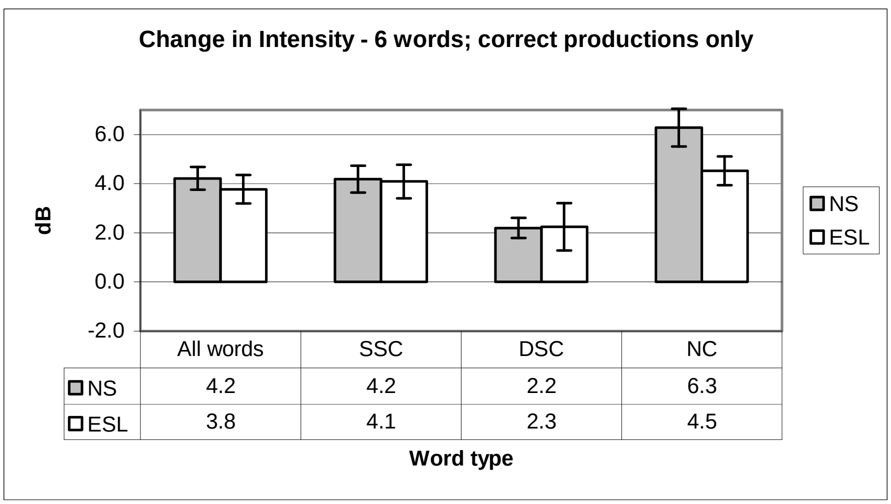


There was a significant difference in the productions between the two speaker groups for: All words, $t(22) = 2.677$, $p < .05$, DSC, $t(22) = 5.728$, $p < .0005$, NC, $t(22) = 3.120$, $p < .01$ with native speakers having greater SV:UV vowel duration ratios than the ESL speakers. The difference between vowel duration ratios for SSCs was not significant ($p > .05$). A Mann-Whitney U test agreed with the t-test in that the results for SSC were not significant ($p > .05$), but the DSCs: U(22) = -3.744, $p < .0005$ and NCs: U(22) = -3.042, $p < .01$ were statistically significant.

As with the other analyses for vowel duration ratio, only SSCs were produced in a similar fashion by the two speaker groups. It is interesting to note that even though this analysis included only correct productions for DSCs, the ESL group still had a vowel duration ratio that was less than 1:1. In other words, they produced these words with just slightly more duration on the UV than on the SV.

4.6.2 Change in Intensity

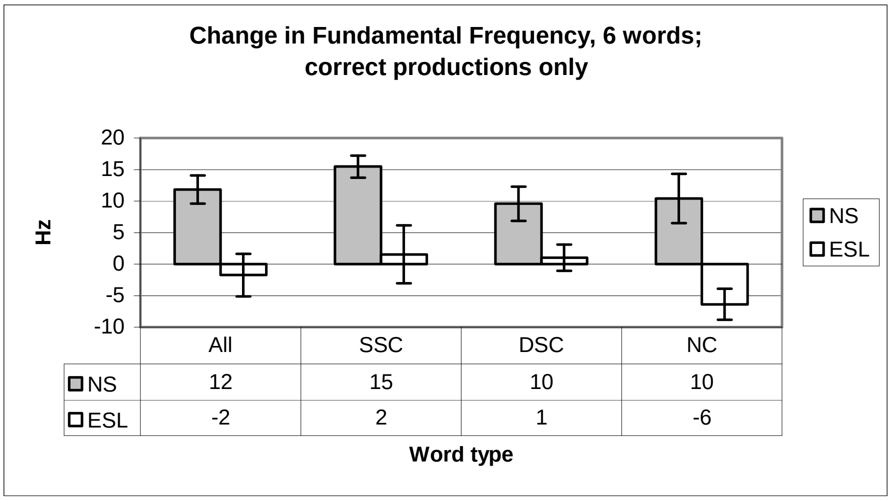
Figure 4.11: Change in Intensity. 6-word set with a reliable increase in F0 on the SV by native speakers, ESL speakers' correct productions only



Unlike the other analyses for intensity, no significant differences were found when comparing all words, or when divided by cognate status ($p > .05$). This finding is very interesting as it shows that when the ESL speakers produced words with correct stress placement for this set of words, no significant differences in intensity existed between the native and non-native speakers. This result differs from what was found for vowel duration ratio and F0, and suggests that of the three variables, the ESL speakers resembled the native speakers most closely in terms of intensity production.

4.6.3 Change in Fundamental Frequency

Figure 4.12: Change in Fundamental Frequency. 6-word set with a reliable increase in F0 on the SV by native speakers, ESL speakers' correct productions only



As with the analysis of F0 for the 6-word set when all productions were included, a significant difference is seen here for change in F0 for: All words, $t(22) = 3.350$, $p < .01$, SSC, $t(21) = 2.834$, $p < .05$, DSC $t(21) = 2.475$, $p < .05$, and NC, $t(22) = 3.642$, $p < .01$ with the native speakers producing greater changes in F0 between SV and UV than ESL speakers. A Mann-Whitney U test for the SSCs confirmed that the difference was significant: U(22) = -3.063, $p < .01$.

As with the 6-word set for all productions, the native speakers produced significantly greater changes in F0 between the SV and UV compared to the ESL speakers for all word types. The ESL speakers made very little difference in F0 between SV and UV on these words. This result was not expected as Spanish speakers have been reported to use F0 movement to a large extent to mark lexical stress in Spanish (Llisterri et al., 2003).

A discussion of the results of the previous four sets of analyses follows in the next section.

4.7 Discussion

The goal of the production study was to investigate and describe how the ESL speakers compare to the native English speakers in the production of the acoustic correlates of lexical stress. In addition to describing the production of the target words and how cognate status could play a role in any differences found, the findings were also used to inform the design of the listening study which is described in the next chapter. The analyses of the production data demonstrate that the ESL speakers produced the acoustic variables very similarly to the native speakers at times, while in other cases they were very different.

The first part of this discussion summarizes the main findings of the production study by word set. The complete 12-word set is discussed first, followed by the findings of the 6-word set. The results of the analyses when considering only correctly pronounced productions then follows for the 12- and 6-word sets. In section 4.8, the results are interpreted in light of the two major research questions for this study as outlined in the Overview of this chapter.

4.7.1 12-word set, all productions

The complete 12-word set presents the most general findings of the production study. The native English and ESL speakers differed in their production of vowel duration ratio for ‘All words.’ The native speakers averaged a 1.9:1 SV:UV ratio, meaning that they produced the SV nearly twice as long as the UV. The ESL speakers, on the other hand, produced a vowel duration ratio of 1.4:1. This difference was statistically significant. When cognate status is taken into account, we see that the two speaker groups produce the SSCs almost identically. The same cannot be said for the DSC or NC category in which statistically significant differences were found between the vowel duration ratios of the two groups. This finding suggests that the ESL speakers were best able to produce vowel duration ratios most similar to native speakers when the word had the same lexical stress pattern in both languages.

For intensity, the results are similar to those for vowel duration ratios. The overall production of change in intensity between the two speaker groups (3.1dB for the native speakers, 1.5dB for the ESL speakers) was significantly different. The difference is also significant for the DSCs and NCs. However, the change is not significantly different for the SSCs, again signaling that similar stress patterns in Spanish and English aid an ESL speaker to produce more native-like patterns when speaking English.

The two groups produced fundamental frequency more similarly overall than they did vowel duration or intensity. No significant differences were found between the speaker groups for ‘All words,’ DSCs or NCs. At first glance, this result also indicates that of the three acoustic correlates of lexical stress examined, the participants showed the least change in F0. For example, in the ‘All words’ category, there is almost no difference in F0 between the SV and UV for either group (1Hz for the Native speakers, -1Hz for the ESL speakers). However, Table 4.2 showed that the Native speakers produced a range of movement between -111 Hz to +103 Hz while the ESL speakers produced a range of -118 Hz to +131 Hz. In other words, the speakers did vary F0, often greatly, between the

SV and UV on individual productions, though it seems these differences disappeared when the averages were calculated in the F0 analysis. For that reason, a subset of the target words in which the native speakers consistently produced increases in F0 on the SV was created and further analyses were conducted on these words. The results of the analyses for vowel duration ratios, intensity and fundamental frequency for this subset of target words are the focus of the following section.

4.7.2 6-word set, all productions

The results for vowel duration ratio and change in intensity were the same in this analysis as they were for the complete 12-word set. Overall, vowel duration ratios are smaller on this set of words than on the 12-word set, being 1.4 for the Native speakers and 1.2 for the ESL speakers. The difference in these ratios, however, is still statistically significant. The vowel duration ratios for DSCs and NCs are also smaller than in the 12-word set, yet the difference is also still statistically significant. The results for SSCs in this set are identical to those for the 12-word set as the two words that qualified as SSCs, *radio* and *October,* are the same in both analyses. The same trend holds for the analysis of change in intensity, with only the SSCs being statistically similar between the two groups. These results again confirm that the two speaker groups are producing vowel duration ratios and change in intensity differently in all categories except the SSCs, signaling that same stress patterns for words in a speaker's L1 and L2 facilitate more native-like production patterning by the ESL speakers.

The principal purpose of this analysis, of course, was to determine how the ESL speakers compared to the native speakers in terms of change in fundamental frequency between the SV and UV on this set of words in which the native speakers consistently increased F0 on the SV. The results of this analysis show that, for these words, the ESL group does not produce the same mean patterns in F0 as the native speaker group. In fact, the ESL group on average produced very little difference in F0 between the SV and UV. The test statistics show that the native speaker group produced greater change in F0 than the ESL speakers between the SV and UV on all word types. Again, this is not to say that the ESL speakers do not use increases of F0 on the SV to mark lexical stress; we know they did at times. But on this subset of words, they produced minimal changes in F0. It seems that although the ESL group produced values for vowel duration ratio and change in intensity that were statistically different from the native speakers, they relied on increases in these two variables to mark lexical stress.

A discussion of these analyses continues in the next section in which only productions by the ESL speakers that were judged as having correct stress placement are included.

4.7.3 12-word set, correct productions only

The next question examined is if there are apparent differences between the two speaker groups purely because mispronunciations of the ESL speakers skewed the results and created differences between their data and that of the native speakers. It was found that less than 12% of the data was pronounced with lexical stress on the incorrect syllable by the ESL group (see Table 4.3), and the majority of these mispronunciations (8.8% of the total data) involved DSCs. The results of the analyses in which incorrect productions were excluded are summarized and discussed below, with a particularly noteworthy finding regarding the production of change in intensity for the 6-word subset.

For vowel duration ratio, the SSC category is again the only one in which the ESL speakers did not differ significantly from the native speakers. This result is not surprising because virtually no SSCs (0.1% of the total data) were mispronounced by the ESL speakers. It is also not surprising that differences in vowel duration ratio for the two speaker groups remained significantly different for the NC category as only 2.3% of the total data was affected by mispronunciations in this category. However, even after removing 8.8% of the total data in the DSC category that had a stress placement error, the differences between the speaker groups remained statistically different for ‘All words,’ DSCs, and NCs at the $p < .05$ level.

The analysis of change in intensity reveals that both the SSCs and NCs were statistically similar between the groups. The DSCs were produced with little average change in intensity between the SV and UV (0.4 dB).

For fundamental frequency, we again see no global differences between the groups except that native speakers increase F0 more on the SV than the UV for SSCs, as was the case with the complete 12-word set. But as in the complete set, differences in F0 disappeared when the data were averaged. Therefore, the 6-word set in which a reliable increase in F0 on the SV was found for the native English speakers is considered again below, including only correct productions by the ESL speakers.

4.7.4 6-word set, correct productions only

In this analysis, the ESL group was again only statistically similar to the Native speakers in terms of SSC production. Their production of vowel duration ratios on the DSCs was just under 1:1. This result confirms that words with the same stress pattern in each language were produced by the ESL speakers most similarly to the native speakers, which might be explained by a positive transfer of patterns between the phonological systems of the speakers' two languages.

The analysis of change in intensity between the SV and UV yields a very informative result as intensity was the only measure on which the two groups did not differ significantly on all word types. When compared with the other analyses done in this production study, it appears that the ESL speakers produce the most native-like production patterns for intensity when lexical stress is correctly placed. This is fortunate because, as will be shown in Chapter 5, producing intensity in a way similar to native speakers is important for speech intelligibility.

For F0, we again see that the native speakers produce greater changes in F0 between the SV and UV than the ESL speakers in all word categories. Thus, incorrect stress placement aside, the ESL group did not show the same pattern for F0 production as the native speakers for this subset of words, and the ESL speakers were more dissimilar from the native English speakers with respect to F0 production than for duration or intensity.

4.8 Summary

In this section, the results of the production data are summarized in terms of the main research questions of this study.

4.8.1 Research question #1: "How does the production of lexical stress by Spanish ESL speakers compare prosodically to native English speakers?"

This question is general in nature. It seeks to determine what, if any, difference occurs between the two speaker groups without being concerned with the cognate status of the word. That is, it is not assumed a priori that transfer of phonological patterns from an ESL speaker's L1 will influence his or her productions of English words. To a large extent, the ESL group produced stress patterns in a way acceptable to native English listeners. That conclusion is supported by the fact that only 11.2% of the data was judged to have incorrect stress placement.

Range of variation in the productions was presented earlier in Table 4.2. This table noted the high and low (extreme) values produced by the two speaker groups in order to show how much they varied on a given variable by considering individual productions. The minimum vowel duration ratio for native speakers was .9, and this ratio occurred on only one token. The ESL speakers, on the other hand, often had values of .3 or .4, equating to a production in which the UV was 3.3 or 2.5 times longer than the SV, respectively. This difference shows that the ESL speakers were more variable in their productions in terms of vowel duration ratios than were the native speaker group.

In terms of intensity, the native speakers' smallest change between the SV and UV was -3 dB while the ESL speakers' was -12.4 dB. On the upper end of the range, the most extreme value occurred with the ESL speakers (15.4 dB for the ESL group compared to 13.1 dB for the NS group). This larger range in intensity differences for the ESL group shows that there was more variability in their production of intensity than for the native English speakers. Their very low values were associated with incorrect placement. For fundamental frequency, the ESL group again produced a wider range of values. F0 differences for the native speaker group ranged from -111 Hz to +103 Hz, and for the ESL group from -118 Hz to +131 Hz.

In addition to yielding information on how the speakers utilized the acoustic correlates of lexical stress in their productions of the target words, the ranges of values were examined to help inform how to vary the manipulations of the stimuli for the listening experiment. Recognizing the non-standard productions by the ESL speakers was needed for the determination of these ranges.

When the data are not separated by cognate status, as in the ‘All words’ analyses, there are consistent statistical differences for vowel duration ratio and intensity between the speaker groups. F0 production was significantly different in the 6-word set only. In the comparisons of correct productions only, we see that statistically there are differences in vowel duration ratios and intensity on the 12-word set, and vowel duration ratios for the 6-word set. However, intensity patterns were statistically identical between the groups when only correct productions of the 6-word set are compared. For F0, the difference was only statistically significant for the 6-word set.

Furthermore, it was hypothesized in Chapter 1 that the ESL speakers would rely heavily on movement in fundamental frequency to mark lexical stress as it has been reported that Spanish speakers use F0 movements extensively to mark lexical stress in Spanish. The transfer of this pattern into English is not attested in this study. In fact, though individual productions by the ESL speakers show a wide range of variation in F0 movement, on average they made little use of F0 to mark stress.

4.8.2 Research Question #2: Does the cognate status of a word affect a Spanish ESL speaker’s production of lexical stress?

The second research question examines exactly where the differences lie in terms of the different target word categories, with the hypothesis that SSCs would lend themselves better to more native-like productions by the ESL participants. The answer to this research question is, “yes.” The same-stress cognates are clearly produced in the most native-like manner by the ESL group, confirming the hypothesis that cognates with the same stress patterns in both languages would facilitate production for the ESL speakers. In fact, the SSCs were always produced in the same manner in terms of vowel duration ratio and intensity. It was only for F0 that the ESL group did not produce the same patterns as the native speakers.

The different-stress cognates were the most problematic for the ESL group, comprising 79% of the incorrectly pronounced words. It is also in this category that the largest differences exist between the native speaker and ESL groups. This strongly suggests that there is a transfer of L1 prosodic patterns to L2 and, when cognates have different stress patterns in the two languages, the L2 speakers diverge most in the acoustic correlates of lexical stress. Though often problematic for the ESL participants, the NC words typically were produced closer to the native speakers' patterns than the DSCs.

As noted earlier in this chapter and in Chapter 3, the findings from this production study were used to choose values for manipulating the stimuli for the listening experiment. The results of the listening experiment, in which five target words spoken by native English speakers were manipulated to resemble the ESL speakers' production of vowel duration ratios, intensity and fundamental frequency as documented here, is the topic of the next chapter.

Chapter 5: Results and Discussion of Listening Task

5.0 Overview

The objective of the listening task was to investigate how American English listeners perceive variations in pronunciation patterns of English by native speakers of Spanish. The focus is specifically on how differing productions of lexical stress affect perceptions of speech intelligibility and nativeness. The method for investigating this question was to test whether the differences measured in the production study (for vowel duration ratios, intensity and fundamental frequency) between the native and non-native speakers affected native English listeners' perceptions of intelligibility and nativeness. As described previously in Chapter 3, the methodology used to explore this question involved manipulating and resynthesizing words that had been pronounced by two of the native English speakers recorded in the production study. The target words chosen for manipulation and resynthesis were spliced out of the original sentence recordings. The manipulations were prosodic in nature in that segmental quality was not altered. The effect of these manipulations on listeners' perceptions is discussed here.

The seven types of manipulations that were performed are listed in Table 5.1. Recall that SV represents the vowel that is marked with primary lexical stress as noted in a dictionary entry, and UV represents the adjacent unstressed vowel. For example, in the word *develop,* the SV is the vowel in the second syllable and the UV is the vowel in the final syllable.

Table 5.1: Seven conditions for stimuli in listening experiment

Condition	Description	Manipulation values
Flat	all flat	SV=UV for duration ratio, intensity, F0
IncDur	increase duration ratio	SV:UV 2.5:1
DecDur	decrease duration ratio	SV:UV 1:2.5
IncInt	increase intensity	SV 6dB > UV
DecInt	decrease intensity	SV 6dB < UV
IncF0	increase F0	SV 20Hz > UV
DecF0	decrease F0	SV 20Hz < UV

Recall that there is no difference in the vowel duration ratio, change in intensity, and change in F0 between the SV and UV in the "Flat" condition. "Increase duration" means that the vowel duration ratio SV:UV was 2.5:1 while "Decrease duration" means that the SV:UV ratio was set to 1:2.5. "Increase intensity" represents a SV which was 6 dB louder than the UV, while "Decrease intensity" represents a SV which was 6dB quieter than the UV. "Increase F0" means that the F0 for the SV was set 20 Hz higher than the UV while "Decrease F0" means that the SV was set 20 Hz lower than the UV. These modifications are relative to the original values of the production selected for manipulation.

The five words selected for manipulation were ones that were at times produced with lexical stress on the incorrect syllable by the non-native speakers in the production study. These five three-syllable words (with location of predicted stress placement marked here with an acute accent) are *bipólar, devélop, disagrée, lúnatic, Octóber*). As described in Chapter 3, the stimuli were presented using the computer program Alvin (Hillenbrand & Gayvert, 2009) and listeners made their ratings by moving a slider on an unmarked visual analog scale (cf. Kreiman et al., 1993). The underlying scale range was 0-700 points, emulating a 7-point likert scale with +/- 100 points between each equally-appearing interval. On this scale, 700 corresponds to the best rating and 0 to the worst rating. The presentation order was balanced so that half of the listeners first rated the stimuli for intelligibility and then for nativeness, and the other half completed the experiment in the opposite order.

To review, the hypothesis was that the stimuli with increases in vowel duration, intensity and fundamental frequency (i.e. IncDur, IncInt, IncF0) on the SV would receive higher ratings for both intelligibility and nativeness than stimuli with all values flattened (i.e. Flat condition) or with increases on the UV (i.e. DecDur, DecInt, DecF0). This hypothesis was made as it has previously been found that increases in those three acoustic dimensions are trademark characteristics of a lexically stressed syllable in English (e.g. Lehiste, 1996) and favor the perception of syllable prominence by native English listeners (Fry, 1955, 1958).

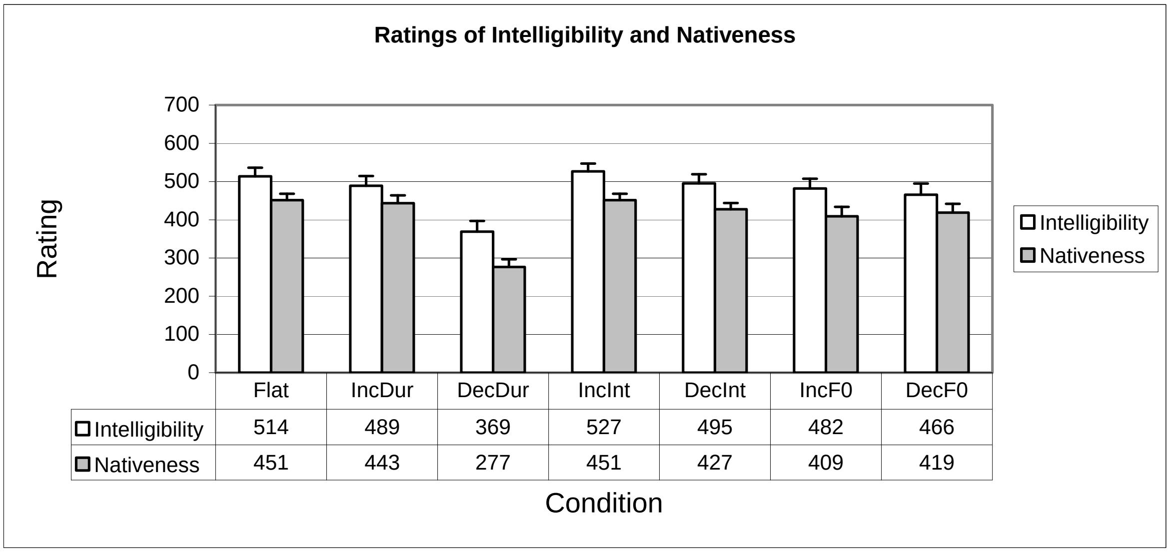
Exclusion of listener:

Twenty-one listeners participated in the experiment. One listener was excluded from the analysis as her responses were found to be outliers in the results for intelligibility and nativeness ratings in four of the seven conditions. Inspection of scatterplots of this listener's responses revealed that for 81 of the 84 tokens in the intelligibility task, she gave a rating of a perfect 700. The three responses that were not rated 700 ranged between 690 and 699. In other words, this listener rated all stimuli as perfectly intelligible, although her ratings for nativeness were much more varied. Since her results were atypical, they were excluded. For all other listeners, the results fell within the normal bounds on the scatterplots for nearly all conditions. These other listeners, when not falling along the trend line on the scatterplots, only fell outside the trend of the other listeners on one of the seven conditions at most.

5.1 Results of the Intelligibility and Nativeness rating tasks

The individual listeners' ratings were compiled in Excel, sorted by condition, and later submitted to statistical analyses in SPSS 17 (SPSS, 2009). Output for the statistical calculations of ratings of intelligibility and nativeness is presented in Appendix O. The 14 repeated tokens that were included to examine intra-rater reliability in each experiment were excluded from the primary data analysis. In total, for each listener there were 10 ratings for each condition (2 talkers (one male, one female) x 5 words), yielding a total of 70 ratings over the seven conditions. An average rating was calculated by condition. The average ratings of intelligibility and nativeness for each condition are shown in Figure 5.1. The bar height for each condition in Figure 1 indicates the average rating for that condition by the 20 listeners. Each bar, then represents the average of 200 ratings per condition. Standard error bars are also shown. A discussion of each result is provided beginning in section 5.5.

Figure 5.1: Ratings of Intelligibility and Nativeness



5.1.1 Intelligibility

The average ratings for each condition were submitted to a one-way repeated measures ANOVA to determine if significant differences existed among the conditions in terms of intelligibility. Results of the ANOVA revealed a significant effect among conditions in intelligibility ratings: F (6, 14) = 25.52, $p < .0005$, multivariate partial eta squared = .916.

In addition to having a statistically significant p value, the effect size, demonstrated in the multivariate partial eta squared value, is very large. This very large effect size is likely due to the fact that ratings of DecDur for intelligibility were well below ratings of the other conditions. Partial eta squared (also referred to as "strength of association") is a statistic used to evaluate effect size which indicates "the proportion of variance of the dependent variable that is explained by the independent variable" (Pallant, 2007, pp. 207-208). Guidelines on effect size indicate that a partial eta squared value of .01 is small, .06 is moderate, and .14 is large. The present result indicates a large effect size, meaning that a great deal of the variance seen in the listener ratings can be explained by the manipulation conditions in the presented stimuli.

As a significant effect was returned from the one-way repeated measures ANOVA, post-hoc pairwise Bonferroni-corrected comparisons were carried out to explore the source of the significant differences among the ratings by condition. The Bonferroni correction on the post-hoc tests is a conservative approach that helps reduce the chance of a Type 1 error when many comparisons are being made by "setting a more stringent alpha level for each comparison [in order to] keep the alpha across all the tests at a reasonable level" Pallant (2007, p. 206).

The result of the Bonferroni post-hoc tests revealed that all conditions were rated higher than DecDur ($p < .0005$). Furthermore, IncInt was rated higher than DecInt ($p < .05$), IncF0 ($p < .05$), and DecF0 ($p < .01$). No other comparisons among conditions were significant.

5.1.2 Nativeness

Ratings of nativeness were submitted to the same type of analysis as those for intelligibility. A one-way repeated measures ANOVA revealed a significant effect among conditions in nativeness ratings: F (6, 14) = 33.17, $p < .0005$, multivariate partial eta squared = .934. Again, the effect size among the conditions is very large. Post-hoc pairwise Bonferroni-corrected comparisons revealed that all conditions were rated higher than DecDur ($p < .0005$). No other significant pairwise differences among conditions were found ($p > .05$).

5.2 Variation in the ratings by listeners

Table 5.2 below gives the low and high ratings for the intelligibility and nativeness tasks by listener. As the values for the main analyses are based on the average ratings by listener, and as such do not consider variation in the listeners' ratings, this Table shows the range of ratings by listener and as such provides insight into how much of the rating scale the listeners utilized. The data in Table 5.2 is sorted by the ratings in the "low" category for each listening task. The highest rating by each listener is listed adjacent to the "low" rating, as is the range of variation between the low and high values. Note that "FL" stands for "female listener" and "ML" stands for "male listener."

Table 5.2: Range of ratings by listener

Intelligibility			
rater	**low**	**high**	**range**
FL1	0	685	**685**
FL6	0	697	**697**
ML9	5	672	**667**
FL3	34	686	**652**
ML1	65	700	**635**
FL8	82	700	**618**
FL4	97	700	**603**
ML6	100	600	**500**
ML3	107	648	**541**
FL11	111	700	**589**
ML7	119	647	**528**
ML10	146	700	**554**
FL2	163	700	**537**
ML4	163	700	**537**
FL9	206	669	**463**
FL10	226	700	**474**
ML8	267	672	**405**
ML2	291	700	**409**
ML5	329	700	**371**
FL5	557	700	**143**
average	**153**	**684**	**530**

Nativeness			
rater	**low**	**high**	**range**
FL11	0	700	**700**
FL6	0	695	**695**
FL8	0	700	**700**
ML2	0	700	**700**
FL5	2	700	**698**
FL3	3	681	**678**
ML7	11	625	**614**
ML9	11	700	**689**
FL1	22	667	**645**
ML6	31	667	**636**
FL4	45	697	**652**
ML10	53	700	**647**
FL10	58	695	**637**
ML1	74	685	**611**
ML8	86	671	**585**
FL9	87	663	**576**
ML4	92	692	**600**
FL2	107	700	**593**
ML3	141	617	**476**
ML5	195	695	**500**
average	**51**	**683**	**632**

The data in Table 5.2 illustrates that, in general, the listeners utilized a large amount of the 700-point scale when rating stimuli for intelligibility and nativeness. These ranges indicate that the average values reported in the main analyses (see Figure 1) are the result of a wide range of ratings along the 700-point scale rather than from a narrow range of ratings toward the middle of the scale. The data in Table 5.2 also indicate that there was less variation overall in the ratings of intelligibility than there was for nativeness, and that the listeners had more positive reactions for the stimuli when rating for intelligibility. When rating for nativeness, the listeners gave ratings that were more often at or toward the lowest end of the scale than were ratings for intelligibility. This result can also be observed by comparing the average ranges for each task. The average range of ratings for intelligibility (530) is lower than that for nativeness (632).

5.3 Correlation analysis of intelligibility and nativeness

The data was subjected to a Pearson product-moment correlation analysis to explore the strength of the relationship between the two continuous variables of the ratings of intelligibility and nativeness. The output of these calculations in SPSS is presented in Appendix P. Scatterplots for pairs of conditions (e.g. IncDur for intelligibility and IncDur for nativeness) were first created in SPSS to check for outliers.

Pearson correlation coefficients (represented as r values) range between -1 to +1. A numeric value close to 0 indicates that there is little or no relationship between the variables, while values closest to +1 (or -1) indicate that there is a strong relationship between the variables. A negative Pearson correlation value indicates that there is a negative correlation between the two variables, that is, as one variable increases, the other decreases. A positive value for the Pearson correlation demonstrates that as one variable increases, so does the other. The result of the Pearson correlation here shows positive values for all conditions, signaling that as ratings of intelligibility increased, so did the ratings of nativeness, and vice-versa. Values of Pearson r are listed below in Table 5.3.

Table 5.3: Results of Pearson product-moment correlation analysis for intelligibility and nativeness

Condition	Pearson *r* correlation	sig. (2-tailed)
Flat	.594**	<.01
IncDur	.713**	<.005
DecDur	.638**	<.005
IncInt	.727**	<.005
DecInt	.496*	<.05
IncF0	.762**	<.005
DecF0	.690**	<.005

**. Correlation is significant at the 0.01 level (2-tailed)
*. Correlation is significant at the 0.05 level (2-tailed)

Pearson product-moment correlation coefficients reveal a positive correlation for each condition with *r* values ranging from .496 for DecInt to .762 for IncF0. The correlation between intelligibility and nativeness ratings is moderate for DecInt, high for IncF0, and fair for all of the other conditions (cf. Berg & Latin, 2004, p. 111). Overall, the Increase conditions had higher correlation coefficients than their Decrease condition counterparts and the Flat condition, meaning that listeners evaluated intelligibility and nativeness for these conditions more similarly than they did for the Decrease and Flat conditions.

5.4 Reliability analysis

Two analyses were carried out to check the reliability of the scale used in the listening experiments, one checking the internal consistency of the scale (inter-rater reliability) and the other checking its temporal ("test-retest") stability (intra-rater reliability).

5.4.1 Inter-rater reliability

A two-way mixed intra-class correlation coefficient (ICC) was calculated to evaluate inter-rater reliability. The output of these statistical calculations for is presented in Appendix Q. The ICC is considered “mixed” as one of the variables (in this case, the independent variable, i.e. the stimulus condition) is fixed and the other variable (in this case, the dependent variable, i.e. the listener ratings) is random (i.e. can receive any score). The ICC assesses “the degree to which the items that make up the scale are all measuring the same underlying attribute” (Pallant, 2007, p. 6). Pallant states that Cronbach’s coefficient alpha, the most common statistic used in this type of analysis, “provides an indication of the average correlation among all of the items that make up the scale,” and returns values ranging from 0 to 1 in which higher numbers indicate higher reliability (Pallant, p. 6). A minimum Cronbach’s alpha score of .7 or higher is recommended as an indication of good internal reliability, with a score of .8 or higher being preferred. The present study reveals a Cronbach’s alpha for the intelligibility task of .966 and for the nativeness task of .956. Both scores indicated very good internal scale consistency (cf. Pallant, 2007; Sheard et al., 1991) indicating that different listeners rated conditions in a similar manner as each other.

5.4.2 Intra-rater reliability

As mentioned earlier, 20% of the stimuli in each task were repeated to assess intra-rater reliability. Therefore, in addition to the 70 stimuli that were analyzed, 14 repetitions of tokens were included in each task, bringing the total number of tokens rated to 84. The 14 repeated tokens were different for both the intelligibility and nativeness tasks.

The degree to which individual listeners rated the same tokens similarly was evaluated in two ways commonly used in studies of speech perception (e.g. Bunton et al., 2001; Tjaden & Wilding, 2004). First, percent close agreement was determined by evaluating how often a listener rated the same token within +/- 100 points. A range of +/- 100 points was chosen as the total scale ranged from 0-700, offering 7 divisions of 100 points akin to the number of divisions on a 7-point likert scale. Second, Pearson product-moment correlation coefficients were obtained between first and second ratings of the same token by the same listener. As possible ratings ranged from 0-700 on the unmarked continuous scale, checking for exact agreement on repeated tokens, as frequently done when using likert scales, would offer little information on listener reliability and therefore was not examined.

Percent-close agreement was 64.3% for intelligibility ratings and 61.1% for nativeness ratings. Pearson *r* coefficients ranged from .25 to .95 (mean = .56) for intelligibility and from -.3 to .87 (mean = .54) for nativeness. The mean scores indicate fair intra-rater reliability for both tasks (Berg & Latin, 2004). The statistical output for this calculation is presented in Appendix R.

5.5 Discussion

In this section, the results of the data analysis are interpreted in terms of how they do or do not confirm the hypotheses of the two listening tests as well as how the results inform the research questions about how differing patterns of pronunciation affect a listener's perception of speech intelligibility and nativeness.

5.5.1 Intelligibility

It was hypothesized that tokens with increases in duration, intensity and fundamental frequency on the vowel that is supposed to receive lexical stress (the SV) would receive higher ratings for intelligibility than the conditions in which increases occurred on the vowel that was not supposed to receive stress (the UV), or in the condition in which all of the variables had the same values on the SV and on the UV (the 'Flat' condition). In other words, the conditions IncDur, IncInt, and IncF0 were hypothesized to be more intelligible than their Decrease counterparts DecDur, DecInt, and DecF0 or the Flat condition. The conditions involving changes of vowel duration ratios, intensity, fundamental frequency and the Flat condition will be discussed here in turn. Discussion of significance, reported in *p* values, for differences between conditions refers to results returned from the post-hoc Bonferroni-corrected comparisons of the one-way repeated measures ANOVA explained earlier.

First, an increase in the vowel duration ratio (IncDur) received higher ratings (mean = 489) than did a decrease in duration ratio (DecDur) (mean = 369). This difference is statistically significant ($p < .005$). Given the overall range of means for the intelligibility ratings, it seems that what drives this result is the decrease in ratio which motivated low ratings. Listeners gave the DecDur conditions the lowest ratings for intelligibility out of the seven conditions. The large difference between DecDur and the other six conditions suggests that durational patterns have a strong effect on listeners' perception of speech intelligibility. If the duration of the SV is increased relative to the UV, or if equal, listeners find the speech approximately equally intelligible. However, increasing the duration of the UV relative to the SV is strongly disfavored.

As with duration, the condition with increased intensity on the SV (IncInt) received significantly higher ratings than its counterpart with a reduction of intensity on the SV (DecInt) ($p < .05$). This result indicates that listeners noticed the increased intensity (+6 dB) on the SV and found the speaker to be easier to understand with this increase. This finding supports the hypothesis that increasing intensity on the SV would improve speech intelligibility, suggesting that a speaker should increase intensity on the vowel in the syllable that is supposed to be stressed to increase his or her intelligibility. Increasing intensity on an unstressed syllable lowers speech intelligibility.

Third, increasing fundamental frequency on the SV (IncF0) was rated no differently for intelligibility than was increasing F0 on the UV (DecF0). This result does not confirm the hypothesis that raising pitch on the SV would lead to higher ratings of intelligibility. Thus, though higher F0 has been observed to be a characteristic of prominent and stressed syllables in English (Fry, 1958; Lehiste, 1996), it does not appear to alter intelligibility compared to a lowered F0 on the SV. On one hand, this result is surprising because it does not coincide with the assumption that increasing all three dimensions – duration, intensity and F0 – will help to mark prominence on a stressed syllable. On the other hand, the result makes sense when considering what native speakers of English can do (and in fact *did* in the production study) when speaking. Pitch accent is a lexical property that aligns with the stressed syllable in English, and lowering F0 on the SV was a common production pattern by the native speakers in this experiment. Therefore, the native listeners presumably interpreted a lowering of F0 on the SV as a production with a low pitch accent associated with that word. If this is so, it would help explain why the native listeners were equally accepting of decreased F0 or increased F0 on the SV.

The most surprising result in the intelligibility task comes with the ‘Flat’ condition. It was hypothesized that the Flat condition would receive higher ratings than the ‘decrease’ versions, but not as high as the ‘increase’ versions. In numerical terms, the ‘Flat’ condition (mean = 514) was rated second only to IncInt (mean = 527). In terms of significance, however, only DecDur was rated significantly more poorly than the 'Flat' condition (p<.005). This result is surprising because the tokens in the 'Flat' condition sound completely monotone and in fact do not sound like a typical native speaker's voice of any language. However, the participants found the intelligibility of this condition to be comparable to all of the other conditions with the exception of DecDur.

5.5.2 Nativeness

In this section, the results of the nativeness ratings and how they relate to the hypotheses and research questions are discussed. As in the discussion of the intelligibility ratings, the conditions based on changes of vowel duration ratios, intensity, fundamental frequency and the flat condition will be discussed here in turn.

As with ratings of intelligibility, duration plays an important role in listeners' perceptions and ratings of nativeness. The hypothesis was confirmed that increasing duration on the SV (IncDur) would be rated significantly higher than increasing the UV's proportion of duration (DecDur). In fact, the results show that of the seven study conditions presented to the listeners, only an increased duration on the vowel that is supposed to be unstressed (the UV) was strongly disfavored by native English listeners (p<.005). This result suggests that in order for a speaker to sound most native-like, he or she should be sure not to produce the UV longer than the SV.

Changes in intensity did not impact the listeners' ratings of nativeness, so the hypothesis that increasing intensity on the SV (IncInt condition) would lead to higher ratings of nativeness than increasing intensity on the UV syllable (DecInt condition) was not supported. In light of the findings from the intelligibility experiment, this result was unexpected because the IncInt condition was rated significantly higher than DecInt in the speech intelligibility task, indicating that the +/- 6dB manipulation was acoustically perceptible. This result suggests that the listeners may have interpreted changes in intensity differently when rating for intelligibility than for nativeness.

As with speech intelligibility, differences in fundamental frequency failed to contribute significantly to listeners' perception of nativeness. Although this result does not confirm the hypothesis that increased F0 on the SV would lead to higher ratings of nativeness, it is again understandable in the sense that native English speakers can and do produce both H* and L* pitch accents when speaking and the listeners may have equated the differing movements on the SV with acceptable high or low pitch accents on the stimuli.

Finally, the Flat condition was rated equally native-sounding as all other conditions with the exception of DecDur. This result fails to confirm the hypothesis that the 'Flat' condition would be rated more poorly than the 'increase' ones yet better than the 'decrease' ones. This result was not expected because completely monotone speech is something that native English speakers virtually never produce in normal speaking situations. To the researcher's ear, these flat tokens sound quite robotic because there is no prosodic movement. On the other hand, the segmental quality of these tokens was unchanged and thus completely native. The fact that listeners rated the flat tokens just as highly as all other conditions except DecDur suggests again that durational and segmental qualities of the stimuli in this experiment were most important for assigning high ratings of nativeness.

5.5.3 Correlation of intelligibility and nativeness ratings

Ratings of intelligibility and nativeness correlated positively for all conditions, indicating that listeners perceive changes in the acoustic correlates of stress similarly for these two rating tasks, and that as ratings increased in one task (e.g. intelligibility of a condition) they did for the other task as well (e.g. nativeness of the same condition). Therefore, changes that were favored for one condition in the intelligibility rating task were also favored for the nativeness rating task.

The Pearson *r* values given in Table 5.3 show a range in correlation from .496 for DecInt to .762 for IncF0, indicating correlation strength between moderate and high for each of the conditions (Berg & Latin, 2004). It is interesting to note that a pattern emerges when doing pairwise comparisons between the ‘increase’ and ‘decrease’ conditions. The ‘increase’ condition for each variable has a higher correlation between intelligibility and nativeness than its ‘decrease’ counterpart. That is, the *r* value for IncDur is .713 compared to .638 for DecDur; It is .727 for IncInt compared to .496 for DecInt; and it is .762 for IncF0 compared to .690 for DecF0. This result may suggest that the “increased” modifications were more expected by the listeners and were therefore rated with more consistency.

The largest difference in these ranges is found in the condition involving changes in intensity, suggesting that the listeners assigned ratings for the intensity conditions less consistently between the two tasks than they did for duration and fundamental frequency. In addition, the ‘Flat’ condition had the second weakest correlation at .594, indicating that the listeners may have perceived the ‘Flat’ condition somewhat differently with respect to intelligibility than with respect to nativeness.

5.5.4 Inter-rater reliability

An evaluation of inter-rater reliability seeks to determine if different individual listeners gave similar ratings to each other for the various conditions. As stated earlier, Cronbach's alpha for intelligibility ratings was .966 and was .956 for the nativeness ratings. Both scores indicate very good internal scale consistency reliability (cf. Pallant, 2007, p. 98) suggesting that different listeners rated the same condition in a similar manner to each other.

5.5.5 Intra-rater reliability

The assessment of intra-rater reliability seeks to determine how consistent individual listeners are when rating the same token on more than one occasion. As mentioned previously, 20% of the tokens were repeated for each rating task, with the 20% of repeated tokens in the intelligibility rating task not the same as those repeated for the nativeness rating task. All of the tokens for each task, including the repeats, were randomized by the computer program Alvin. Each rating task lasted approximately 10 minutes. Consequently, a listener might hear two examples of the same token within a minute of each other, or up to 10 minutes apart. This is a very short test-retest interval.

The first measure used to assess intra-rater reliability was percent-close agreement. This was calculated as a percentage of how often ratings of a first and second listening were assigned a value on the unmarked sliding scale within 100 points of each other. For the intelligibility and nativeness ratings, percent-close agreement was 64.3% and 61.1%, respectively. A Pearson product-moment correlation analysis between first and second ratings was also calculated. The correlations varied widely from listener to listener, with r values ranging from .25 to .95 on repeated ratings of tokens for intelligibility and -.3 to .87 for repeated tokens in the nativeness task. The positive *r* values indicate that the repeated tokens were judged similarly on both listenings. One listener, however, varied so much between first and second listenings that her intra-rater reliability score was negative. Twelve of the twenty listeners had significant correlations ($p<.05$) between the first and second listenings on the intelligibility ratings, and twelve listeners (not all the same as in the intelligibility task) had significant correlations between the first and second listenings on the nativeness ratings ($p<.05$). Seven of the twenty listeners had significant correlations at the $p<.05$ level for both intelligibility and nativeness ratings. Therefore, it can be said that there was partial validation of intra-rater reliability (cf. Munro & Derwing, 1995, p. 297).

It is worth noting that the range of *r* values for intra-rater reliability on the intelligibility ratings is smaller than that for the nativeness ratings, even though the mean values are nearly identical (.54 and .56, respectively). This larger range in Pearson *r* values for the nativeness ratings suggests that perception of nativeness was less reliable than for intelligibility, or that listeners had greater differences in their approach to rating the tokens for nativeness than they did for intelligibility. It is possible that rating nativeness may be more a subjective exercise than rating intelligibility and therefore more open to variation.

5.6 Summary

The results of this listening study show that changes in these acoustic correlates of lexical stress in English have a significant effect on the perception of speech intelligibility and nativeness. A longer vowel duration ratio between the SV than the UV, and increased intensity on the SV, are particularly favorable for improving judgments of speech intelligibility. Only durational patterns in which the UV was the longer vowel were found to be significantly detrimental to the perception of nativeness. This suggests that native English listeners are more sensitive to expected relative vowel durations when combined with appropriate segmental quality than to differences in intensity or F0 when making judgments of speech nativeness. Changes in F0 on the SV did not affect either intelligibility or nativeness ratings. At least for nativeness, variation in F0 movement, both by increasing and decreasing F0 on the SV, might actually help to signal that the speaker is native and could possibly explain why no differences were found between the two conditions for F0.

Some final considerations of the results from this listening experiment, along with those from the production experiment, will be given in the concluding chapter, along with implications of these results for language teaching. Limitations of the study, as well as directions for future research, will also be discussed.

Chapter 6: Conclusion

6.0 Overview

The first goal of this dissertation was to compare how native English speakers and Spanish ESL speakers produce three acoustic correlates of lexical stress in English. The results from this comparison were then used to address the second goal, which was to investigate how different productions of these acoustic cues affect native English listeners' perception of speech intelligibility and nativeness. To do so, vowel durations and differences in intensity and fundamental frequency were controlled individually to examine which, if any, most affected listeners' perceptions. A summary of the main findings of the dissertation is given in this chapter. Implications of these findings are then discussed, followed by an assessment of the strengths and limitations of the dissertation. In closing, avenues for future research related to the present study as well as some final comments are offered.

6.1 Summary of the results

In the following two sections, a summary of the results is given, first for the production study and then for the listening study.

6.1.1 Production study

The first research question of the production study sought to determine how the production of lexical stress by the Spanish ESL speakers compared to that of the native English speakers. The hypothesis was that the ESL speakers would rely most heavily on F0 movement to mark lexical stress. It was found that the ESL speakers produced a wider range of vowel duration ratios and greater variation in intensity and F0 than did the native speakers. Considering all productions of all the tokens, it was found that differences between the stressed and unstressed vowels in both duration and intensity were statistically significant between speaker groups. Change in F0 was statistically different only for the set of six words in which native speakers consistently produced an increase in F0 on the stressed vowel.

When analyzing only correctly stressed productions by the ESL speakers, vowel duration ratios were again found to be different for both the complete data set and the six-word set. Change in intensity was statistically different only for the complete data set, and F0 was only significantly different for the six-word set. These results indicate that in those productions where the ESL speakers placed lexical stress on the correct syllable, they produced values for change in intensity similar to those of the native speakers. Furthermore, it was found that the ESL speakers did not rely on increasing F0 on the stressed vowel to mark lexical stress. This was particularly clear when considering the analysis of the six words in which the native speakers reliably did so. In fact, the ESL speakers used F0 movement between stressed and unstressed syllables very little in the data analyzed in this study, contrary to expectations. Therefore, although the ESL speakers often differed significantly from the native speakers in their vowel duration ratios and intensity patterns, it appears that they relied on these cues more heavily than F0 for marking lexical stress.

The second research question of the production study asked if the cognate status of a word affects a Spanish ESL speaker's production of lexical stress. It was hypothesized that the ESL speakers would produce cognates with the same stress pattern in each language (SSCs) (e.g. '***ra***-*di-o* / '***ra***-*di-o*) more similarly to the native speakers than cognates with different stress patterns in each language (DSCs) (e.g. '***lu***-*na-tic* / *lu-*

'***na***-*ti*-*co*). Further, it was hypothesized that non-cognates would be produced in a manner more similar to the native speakers than the DSCs, but not so similarly as the SSCs.

Overall, the ESL group produced the SSCs most similarly to the native speakers. This result was true for vowel duration ratio and intensity, but not consistently for F0 patterns. Particularly when analyzing the six-word set in which native speakers realized a reliable increase of F0 on the stressed vowel, results clearly differed for the ESL speakers. Overall, it appears that positive transfer of phonological patterns from the ESLs' L1 to L2 can explain the fact that it was the SSCs that the ESLs produced in a manner most similar to the native speakers. The hypothesis that DSCs would be produced in a manner most dissimilar to the native speakers was also confirmed. In all analyses, the ESL group differed statistically from the native speakers for this word type. This result suggests that negative transfer took place and drove these differences. Finally, although NCs were often problematic, in general they were produced by the ESLs in a manner closer to the native speaker production than the DSCs. In particular, in correctly produced tokens of NCs the ESL group produced a difference in intensity between the stressed and unstressed vowels that was not significantly different from the native speaker group.

6.1.2 Listening study

The first research question of the listening study asked if independently increasing vowel duration, intensity or F0, on the stressed vowel would impact a native English listener's perception of speech intelligibility and nativeness. It was hypothesized that increasing the value of any one of these variables on the stressed vowel would improve listeners' ratings. This hypothesis was only partially confirmed. For the ratings of speech intelligibility, it was found that decreasing duration on the stressed vowel had the most adverse effect on ratings and that increasing intensity on the stressed vowel had a positive effect. Increasing duration on the stressed vowel, or having the same duration on the stressed vowel and the unstressed vowel, led to ratings for intelligibility that were not significantly different. Increasing or decreasing F0 on the stressed vowel also failed to affect ratings of intelligibility.

Ratings of nativeness were not improved by increasing vowel duration, intensity or F0 of the SV. Only decreasing duration on the SV led to statistically lower ratings than the other conditions. This indicates that, in order to sound most native-like, it is important to have adequate duration on the syllable that native listeners expect to receive lexical stress.

The second research question asked if a positive correlation exists between ratings of intelligibility and ratings of nativeness. The hypothesis was that there would be no correlation, as a token could sound intelligible yet not native. The results demonstrate, however, that a positive correlation between intelligibility and nativeness ratings did exist. In other words, conditions that led to higher ratings of intelligibility also led to higher ratings of nativeness. This finding indicates that pronouncing words in a native-like manner will aid listeners in their understanding of an utterance. Therefore, aiming for native-like pronunciation should be considered important when teaching pronunciation as it increases a speaker's level of intelligibility.

In the next section, some implications of the study findings are suggested.

6.2 Implications of the research findings

It is hoped that the findings of the present dissertation not only contribute to general linguistic knowledge of prosody in first and second languages, but that they can also be applied to work on ESL pedagogy or other fields such as speech disorders. The potential application of the results here could benefit second language learners or people with a speech disorder as they strive to be most intelligible and, if they desire, to sound most native-like. The data yields several implications for these speech communities.

First, producing appropriate durational patterns for the stressed and unstressed vowels of a word should be emphasized. It was found that decreasing duration on the vowel that should receive lexical stress is detrimental to ratings of both speech intelligibility and nativeness. Instructors or speech coaches should ensure that their students learn to produce these two vowels with the same durations, at a minimum, or preferably with an increase of duration on the vowel that should be stressed. The present results suggest that unexpected duration ratios are very salient and listeners react strongly to them.

It is uncertain why increasing duration did not have a more positive effect on ratings, particularly because longer vowel duration on the stressed syllable is an important and noticeable cue of lexical stress. It is possible that for some tokens the duration on the stressed vowel was increased more than might be expected in typical speaking conditions. Such a situation could have lowered some listeners' ratings for this condition. This possibility could be investigated in the future by examining the trends of listeners' ratings for each target word in the stimuli set.

Secondly, the listeners found that productions with an increase of intensity on the stressed vowel relative to the unstressed vowel are significantly more understandable, at least in pairwise comparisons with tokens that do not have this increase. In fact, increasing intensity on the stressed vowel was the only condition which listeners found to be helpful for improving speech intelligibility. Therefore, working with learners to increase intensity on the stressed vowel should prove rewarding as it positively impacted ratings of speech intelligibility.

Increased intensity on the stressed vowel was only found to be statistically important in ratings of speech intelligibility. It is unclear why this movement in intensity did not receive significantly high ratings for nativeness. Perhaps this is akin to how, when a listener asks for a repetition of an utterance, an interlocutor may raise his or her voice when repeating the utterance. Teachers are known to do the same when they feel that their students do not understand a particular piece of information; in such a situation, a teacher often speaks more loudly. Perhaps loudness helps a person understand the message more clearly, but does not have an impact on the perceived nativeness of the speaker.

Regarding F0 movement, the results of this study do not indicate that differences between the stressed and unstressed vowels are important for improving ratings of intelligibility or nativeness. As stated before, increasing and decreasing F0 on the stressed vowel is something native speakers do naturally. It is essential for the production of pitch accent and can also be related to paralinguistic aspects of speech such as expression of emotion. Therefore, producing varied changes in F0 may help to mark a speaker as native and thus may explain why neither increases nor decreases in F0 on the stressed vowel affected the ratings.

It is curious that the ESL speakers did not vary F0 movement much overall compared to the native speakers, particularly because Spanish speakers apparently do so when speaking in their native language. It appears that there was not a transfer of F0 production from their L1 to their L2. It is possible that the speakers in this study did not vary F0 because they have not perceived that native speakers often do this, perhaps perceiving English as sounding more monotonous than Spanish. If this is so, they may have strived for overall flatter F0 contours than the native English speakers. Further investigation of this topic is needed to determine the extent to which Spanish speakers use increases or decreases in F0 on the stressed vowel to mark lexical stress when speaking in Spanish.

A final implication is that teaching pronunciation with the aim of sounding like a native speaker still appears to be a worthwhile endeavor. As mentioned in Chapter 2 of this dissertation, attaining native-like pronunciation, while of great importance to many language learners, has increasingly become considered of lesser importance by pronunciation researchers and teachers. This is because speech intelligibility is now considered the primordial goal of pronunciation training. The results of this study demonstrate that ratings of intelligibility correlate positively with ratings of nativeness. Therefore, learning to pronounce words like a native speaker should also enhance a language learner's ability to be understood easily by his or her listener and deserves to retain a level of importance in the language classroom.

To summarize, the results suggest that focusing primarily on vowel durations and intensity will be most beneficial for language learners as these cues appear to be the most obviously related to perceived intelligibility and nativeness.

6.3 Strengths and limitations of the study

This dissertation has strengths and limitations, some of which are related to each other. Some of these limitations may be addressed with future research on this topic as will be discussed in the following section.

One strength of the production study is that the stimuli were carefully controlled and it was possible to collect six repetitions of each target word from each speaker. Having multiple repetitions of each word by each group of speakers gives insight into the amount of variation in the data and also makes a statistical analysis possible. A limitation of this choice is that read speech is less natural than spontaneous discourse. Therefore, the data here can not inform about more typical speaking styles such as casual conversation. Furthermore, only single target words were extracted from each sentence and analyzed, thus ignoring any possible patterns that may exist in the connected speech.

Another strength of the research design is that each acoustic correlate of lexical stress was controlled individually when preparing the conditions for the listening experiment. This level of control is beneficial as it reveals with more certainty how much each variable contributes to the perception of intelligibility and nativeness. However, in order to exercise this control, the stimuli had to be manipulated and resynthesized. The resynthesized speech does not sound entirely natural, particularly because changes to the F0 patterns of the word made the words sound more computerized. Reducing the naturalness of the speech via resynthesis may have influenced listeners' ratings of the stimuli in the listening task.

Another limitation of the study is that the results are based on a small set of words. Only twelve different three-syllable words (totaling 1,728 individual productions over the 24 speakers) were analyzed for the production study. Of these twelve words, only five were manipulated and resynthesized for the listening task. The limitation here, obviously, is that the results can not be easily generalized as the number of unique words is small.

Though the limitations of the study are numerous, the final methodology behind the dissertation incorporates advances from many initial attempts to discover the optimal way of approaching and investigating the research questions at hand. It is felt that this methodology proved to be effective for addressing issues relating to the study of lexical stress.

6.4 Future Work

In the process of working on this dissertation, a great number of possibilities for future research related to the production study and the listening study were encountered. For the production study, a large amount of data was collected, much of which was not analyzed. Investigating the same research questions proposed here on a larger set of words would be valuable in order to expand the data set and increase the generalizability of the findings. Examining a larger set of words may also prove particularly useful by potentially reducing the influence of lexical effects. Furthermore, analyzing the sentences as a whole, or phrases within each sentence, could help reveal patterns of stress placement in more connected speech. The analysis of recordings of the Rainbow Passage that were also collected would be useful in this matter.

Another piece of data that was collected but not analyzed is recordings of the Spanish speakers reading sentence stimuli and a reading passage in Spanish. It would be interesting to examine if the native Spanish speakers in this study produced similar patterns when reading in Spanish as they did when reading in English. Furthermore, incorporating data from ESL speakers of differing proficiency levels could be revealing. Only highly proficient speakers participated in this study because it was important that the speakers be able to read the stimuli fluently. Yet examining the speech of intermediate or beginning learners could reveal patterns not witnessed in the present data and could help inform implications for ESL pedagogy for speakers of differing proficiency levels.

The listening study also inspired its share of possibilities for future research. First, the selection of listeners in this study was not controlled, except that the listeners were predominately members of the university community. This was done with the intention to obtain a general impression of listener reactions rather than soliciting ratings from a potentially more homogenous audience, such as a group of ESL teachers. However, it would be very interesting to control the listener groups to examine if listeners with very little contact with second language speech react differently to the stimuli than listeners who have frequent contact with ESL speakers. Similarly, second language learners often state that it is easier to understand other second language speakers than it is to understand native English speakers. It would be informative, then, to ask L2 speakers to participate in the listening study to examine how their responses compare to those of the native English listeners.

The stimuli for the listening task were based on manipulations where the acoustic correlates of lexical stress were increased, decreased, or flattened. This ignores another possible production, one which did occur with the ESL speakers in this study: placing too much stress on the *expected* stressed syllable. Presenting stimuli with more-than-typical stress on the stressed vowel could demonstrate on how listeners react to yet another non-standard production. Furthermore, the conditions in the listening experiment represented only single changes of each study variable. This was useful to determine if any single manipulation affected the listener ratings. However, these acoustic cues, when produced in typical speaking situations, are often varied in combination by speakers. Thus, it could be profitable to investigate how changes to two or more variables simultaneously affect listener perceptions (e.g. increasing both intensity and duration on the SV).

Many more ideas come to mind for future research in the listening experiment, such as questioning sentence position of a target word affects the overall intelligibility of a sentence, or if there is a correlation between reaction times and ratings for stimuli (i.e. if slower reaction times correlate with lower ratings). All of these avenues for future research can be addressed with the current research methodology. But perhaps a more difficult question to address, yet of great practical relevance, regards how to take the findings from this study and apply them to teaching pedagogy, and to later evaluate their effectiveness. For example, how do we best go about teaching students to produce appropriate changes in intensity between stressed and unstressed vowels and subsequently test its effectiveness on the perception of the student's degree of intelligibility and nativeness? Certainly, working toward a practical outlet for the implications of the study is of great importance.

6.5 Closing Comments

At the outset of working on this dissertation, the task at hand seemed rather simple and straightforward. Hopefully the material herein has been described in a way that makes it still seem that way. During the process, however, there were many twists and turns, most of them unexpected by the researcher. Yet these developments made the process not only challenging but also interesting. It also gave rise to seemingly endless possibilities for further research.

The study of prosody, in general, and prosody of the speech of second language learners, in particular, is in its early stages and much is left to be learned. At present, few studies of prosody and second language speech have addressed the role of fundamental frequency and intensity, perhaps due to the complexity of evaluating such features. It is hoped that the work here both stimulates further study in this fascinating area by others interested in linguistics and speech and hearing sciences and adds some knowledge to the study of prosody as it relates to native and non-native speakers. And it is hoped that the results of this study and future work find their way to language learners and users, whether they encounter themselves in the classroom, a speech clinic, or on the street, as we work to support language communities as a whole.

Appendices

Appendix A: Speaker demographics

Key: MNS = Male native speaker; FNS = Female native speaker; MESL = Male ESL speaker; FESL = Female ESL speaker

Native speakers			**Time in New Mexico**	
Speaker	**From**	**Age**	**Years**	**Months**
MNS1	New Mexico	30	25	
MNS2	Missouri	43	2	
MNS3	New Mexico	53	28	
MNS4	Montana	31	1	
MNS5	Indiana	36	10	
MNS6	New Mexico	30	25	
FNS1	Arizona	32	1	6
FNS2	New Mexico	27	2	6
FNS3	New Mexico	23	18	
FNS4	Washington	32	1	2
FNS5	Florida	36		6
FNS6	Ohio	31	2	7

ESL speakers			**Time in U.S.**	
Speaker	**From**	**Age**	**Years**	**Months**
MESL1	Ecuador	22	2	2
MESL2	Colombia	21	1	10
MESL3	Ecuador	30	17	9
MESL4	Peru	26		10
MESL5	Peru	38	5	
MESL6	Peru	25	1	6
FESL1	Colombia	45	1	8
FESL2	Colombia	26		4
FESL3	Ecuador	31		3.5
FESL4	Peru	33	2	
FESL5	Venezuela	26	3	1
FESL6	Peru	26	1	

Appendix B: Language background questionnaire - Native English speakers

1. Name ____________________________________
2. Age ________
3. Where are you from? ____________________________________
4. How long have you lived in Albuquerque? ___________________
5. What is your highest level of education?
    a. ___ completed high school
    b. ___ some college
    c. ___ completed a technical or vocational school program
    d. ___ completed an undergraduate program at a university
    e. ___ completed a graduate program at a university
6. Are you aware of any speech impairment you might have? (circle one) yes no
7. Are you aware of any hearing impairment you might have? (circle one) yes no

Appendix C: Language background questionnaire - Non-native English speakers

1. Name ____________________________________
2. Age ________
3. Where are you from (country/city)? ____________________________________
4. When did you first come to the US? ______________
5. How long have you lived in the US? _____ years _______ months
6. Were you first exposed to English *in the US* or *in your home country*? (circle one)
7. How long did you study English before coming to the US? _____ years _______ months
8. How long have you studied English in the US? _____ years _______ months
9. What percentage of time do you speak English each day? ________
10. What is your highest level of education?

    ___ completed high school

    ___ some college

    ___ completed a technical or vocational school program

    ___ completed an undergraduate program at a university

    ___ completed a graduate program at a university
11. How would you rate your English ability?

    ___ Fluent, with native or near-native ability ___ Advanced learner

    ___ Intermediate learner ___ Beginning learner
12. If you are studying English at CELAC at UNM, which class level are you currently enrolled in? (*Low-intermediate, Intermediate, High-intermediate, Advanced,* or *Academic Bridge*) _____________________________
13. If you are a regular UNM student, are you in an *undergraduate* or a *graduate* program? (circle one). What is your major? _________________________
14. Are you aware of any speech impairment you might have? (circle one) yes no
15. Are you aware of any hearing impairment you might have? (circle one) yes no

Appendix D: Non-cognate target words and sentence stimuli

The 24 non-cognate target words are listed in this Appendix. Data for all target words was collected. Note that only the three-syllable words were analyzed in this dissertation. In the table below, “Stressed syllable” refers to the syllable in the word that should receive lexical stress according to the dictionary. For “Word Class,” N = noun, V = verb, J = adjective, B = adverb; For “Position,” I = initial (meaning that the word appeared after the first 3 syllables of the carrier sentence) and F = final (meaning that the word appeared 3 syllables before the end of the carrier sentence).

Two syllables	Stressed syllable	Word Class	Position	Spanish word
glasses	1	N	F	lentes, vasos
friendly	1	J	I	amable
teacher	1	N	I	maestro
instead	2	B	F	en vez de
arrange	2	V	F	ordenar
degree	2	N	I	grado, título

Three syllables	Stressed syllable	Word Class	Position	Spanish word
management	1	N	F	gerencia
easily	1	B	I	fácilmente
location	2	N	F	lugar
develop	2	V	I	desarrollar
disagree	3	V	F	no estar de acuerdo
unafraid	3	J	I	sin temor

Four syllables	Stressed syllable	Word Class	Position	Spanish word
watermelon	1	N	F	sandia
anybody	1	N	I	cualquier persona
dependable	2	J	F	confiable
development	2	N	I	desarrollo
recognition	3	N	F	reconocimiento
underwater	3	J	I	bajo de agua

Five syllables	Stressed syllable	Word Class	Position	Spanish word
unanswerable	2	J	F	incontestable
uncompromising	2	J	I	intransigente
liability	3	N	F	responsabilidad
microwavable	3	J	I	se puede meter en la microondas
magnification	4	N	F	aumento
underdeveloped	4	J	I	subdesarrollado

Sentence Stimuli for Non-cognates

Target word is marked below in **bold** type.

2 syllable words
Susanna didn't forget to bring her new **glasses** on the trip.
She was not **friendly** with the new girls at our neighborhood party.
Matt is a **teacher** by day and a country musician by night.
Matthew decided to order the salad **instead** of the soup.
Jane had to take time off from work so she could **arrange** the wedding.
He earned his **degree** by working two jobs and studying at night.

3 syllable words
The owners are currently looking for good **management** for the store.
She cannot **easily** forget her bad experience with the dog.
The international students found a good **location** for the dance.
One way to **develop** the ideas is by brainstorming a list.
Several of the students in class liked to **disagree** constantly.
The boy was **unafraid** of the consequences of getting bad grades.

4 syllable words
Children and adults both love to eat lots of **watermelon** at picnics.
He hasn't met **anybody** that wants to work at the governor's office.
The head of the Physics department is a **dependable** professor.
It was a **development** we were not expecting to arise that day.
It's incredible that she did not receive **recognition** for her work.
Nancy swam **underwater** for nearly a minute to impress her friends.

5 syllable words
The questions the professor asked seemed almost **unanswerable** for Linda.
Larry is **uncompromising** with his dedication to help the students.
The two young drivers said they would not assume **liability** for the crash.
The plate is **microwavable** but the new plastic glasses and bowls are not.
Microscopes now come with higher levels of **magnification** than before.
The zone is **underdeveloped** in comparison to other neighborhoods.

Appendix E: Cognate target words and sentence stimuli

The 24 Spanish/English cognate target words are listed in this Appendix. Data for all target words was collected. Note that only the three-syllable words were analyzed in this dissertation. In the table below, “Stressed syllable” refers to the syllable in the word that should receive lexical stress according to the dictionary. “Same stress” refers to if the word has the same stress pattern in both languages (Y = yes, N = no); “Identical cognates indicates if the cognate is *orthographically* identical in both languages; For “Word Class,” N = noun, V = verb, J = adjective, B = adverb; For “Position,” I = initial (meaning that the word appeared after the first 3 syllables of the carrier sentence) and F = final (meaning that the word appeared 3 syllables before the end of the carrier sentence).

Two syllables	Stressed syllable	Same stress?	Identical cognate?	Word Class	Position	Spanish word
kilo	1	Y	Y	N	F	kilo
purpose	1	N	N	N	I	propósito
tennis	1	Y	N	N	I	tenis
receive	2	N	N	V	F	recibir
admit	2	N	N	V	F	admitir
control	2	Y	Y	N/V	I	control

Three syllables	Stressed syllable	Same stress?	Identical cognate?	Word Class	Position	Spanish word
lunatic	1	N	N	N	F	lunático
radio	1	Y	Y	N	I	radio
bipolar	2	N	Y	J	F	bipolar
October	2	Y	N	N	I	octubre
engineer	3	N	N	N	F	ingeniero
introduce	3	Y	Y	V	I	introducir/ introduce

Four syllables	Stressed syllable	Same stress?	Identical cognate?	Word Class	Position	Spanish word
necessary	1	N	N	J	F	necesario
literature	1	N	N	N	I	literatura
photographer	2	Y	N	N	F	fotógrafo
combustible	2	N	Y	N	I	combustible
sentimental	3	N	Y	J	F	sentimental
televisions	3	N	Y	N	I	televisiones

Five syllables	Stressed syllable	Same stress?	Identical cognate?	Word Class	Position	Spanish word
intoxicating	2	N	N	J	F	intoxicante
depository	2	N	N	N	I	depositorio
perpendicular	3	N	Y	J	F	perpendicular
testimonial	3	N	Y	N	I	testimonial
organization	4	N	N	N	I	organización
tuberculosis	4	Y	Y	N	F	tuberculosis

Sentence Stimuli for Cognates

Target word is marked below in **bold** type.

2 syllable words

I bought two kilos of grapes for me and a **kilo** for Mary.
His only **purpose** in life seems to be annoying his neighbor.
She plays in **tennis** tournaments around the country on weekends.
Nobody ever thought that one day they would **receive** Paul's notice.
In the federal courtroom today Mike must **admit** that he lied.
You have to **control** the amount of activity in the class.

3 syllable words

The tired student was walking around like a **lunatic** yesterday.
Diana's **radio** stopped working in the middle of the party.
The medication is designed for a strong **bipolar** disorder.
School starts in **October** for some school districts in the northern county.
My brother-in-law got his degree as an **engineer** this summer.
He would not **introduce** the woman he was with at the restaurant.

4 syllable words

The one secretary who thought the work was **necessary** kept calling.
Some modern **literature** can be very difficult to understand.
At the last basketball game the newspaper **photographer** fell asleep.
New forms of **combustible** for car engines hope to increase gas mileage.
Our next-door neighbor sat down to comfort a **sentimental** child today.
Buying new **televisions** can be most expensive for sports stadiums.

5 syllable words

The smell coming from the burning building was **intoxicating** the children.
A city **depository** for storing illegal weapons is needed.
Pennsylvania and Park Avenue don't run **perpendicular** to Lomas.
The student's **testimonial** got the attention of the academy.
Melissa's **organization** did a lot to get the legislation passed.
The newest doctor was slow to explain how **tuberculosis** kills people.

Appendix F: ICC for Stress Placement Ratings

Inter-rater reliability with professional phonetician #1: Male ESL speakers

Case Processing Summary

		N	%
Cases	Valid	66	91.7
	Excluded[a]	6	8.3
	Total	72	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.956	2

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.916[b]	.866	.948	22.785	65	65	.000
Average Measures	.956[c]	.928	.973	22.785	65	65	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Inter-rater reliability with professional phonetician #1: Female ESL speakers

Case Processing Summary

		N	%
Cases	Valid	69	95.8
	Excluded[a]	3	4.2
	Total	72	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.916	2

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.845[b]	.761	.901	11.925	68	68	.000
Average Measures	.916[c]	.865	.948	11.925	68	68	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Inter-rater reliability with professional phonetician #2: Male ESL speakers

Case Processing Summary

		N	%
Cases	Valid	69	95.8
	Excluded[a]	3	4.2
	Total	72	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
1.000	2

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	1.000[b]	1.000	1.000	.	68	.	.
Average Measures	1.000[c]	1.000	1.000	.	68	.	.

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Inter-rater reliability with professional phonetician #2: Female ESL speakers

Case Processing Summary

		N	%
Cases	Valid	69	95.8
	Excluded[a]	3	4.2
	Total	72	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
1.000	2

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	1.000[b]	1.000	1.000	1.927E15	68	68	.000
Average Measures	1.000[c]	1.000	1.000	1.927E15	68	68	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Appendix G: ICC for Pitch Accent Ratings

Case Processing Summary

		N	%
Cases	Valid	96	100.0
	Excluded[a]	0	.0
	Total	96	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.844	2

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.730[b]	.621	.811	6.400	95	95	.000
Average Measures	.844[c]	.766	.896	6.400	95	95	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Appendix H: Target Word Familiarity Scale

Target word familiarity rating

Name _______________________________________

Please rate how often you use the words listed below:

1 = rarely or never
2 = once in a while (e.g. 2 or 3 times a month)
3 = weekly (e.g. 2 or 3 times a week)
4 = daily

1.	admit	1	2	3	4
2.	anybody	1	2	3	4
3.	arrange	1	2	3	4
4.	bipolar	1	2	3	4
5.	combustible	1	2	3	4
6.	control	1	2	3	4
7.	degree	1	2	3	4
8.	dependable	1	2	3	4
9.	depository	1	2	3	4
10.	develop	1	2	3	4
11.	development	1	2	3	4
12.	disagree	1	2	3	4
13.	easily	1	2	3	4
14.	engineer	1	2	3	4
15.	friendly	1	2	3	4
16.	glasses	1	2	3	4

17. instead	1	2	3	4
18. intoxicating	1	2	3	4
19. introduce	1	2	3	4
20. kilo	1	2	3	4
21. liability	1	2	3	4
22. literature	1	2	3	4
23. location	1	2	3	4
24. lunatic	1	2	3	4
25. magnification	1	2	3	4
26. management	1	2	3	4
27. microwavable	1	2	3	4
28. necessary	1	2	3	4
29. October	1	2	3	4
30. organization	1	2	3	4
31. perpendicular	1	2	3	4
32. photographer	1	2	3	4
33. purpose	1	2	3	4
34. radio	1	2	3	4
35. receive	1	2	3	4
36. recognition	1	2	3	4
37. sentimental	1	2	3	4
38. teacher	1	2	3	4
39. televisions	1	2	3	4

40. tennis	1	2	3	4
41. testimonial	1	2	3	4
42. tuberculosis	1	2	3	4
43. unafraid	1	2	3	4
44. unanswerable	1	2	3	4
45. uncompromising	1	2	3	4
46. underdeveloped	1	2	3	4
47. underwater	1	2	3	4
48. watermelon	1	2	3	4

Appendix I: Parametric (t-test) statistical output for production study from SPSS

NB: For Group Statistics: Language 1 = native speakers, Language 2 = ESL speakers

Vowel Duration Ratio: 12-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
DurAllWords3	1	12	1.925	.1545	.0446
	2	12	1.525	.2261	.0653
DurSSCs3	1	12	1.350	.2316	.0669
	2	12	1.525	.2734	.0789
DurDSCs3	1	12	2.208	.2275	.0657
	2	12	1.658	.3965	.1145
DurNCs3	1	12	1.917	.2443	.0705
	2	12	1.400	.2296	.0663

Vowel Duration Ratio: 12-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DurAllWords3	Equal variances assumed	.591	.450	5.060	22	.000	.4000	.0791	.2360	.5640
	Equal variances not assumed			5.060	19.431	.000	.4000	.0791	.2348	.5652
DurSSCs3	Equal variances assumed	.280	.602	-1.692	22	.105	-.1750	.1034	-.3895	.0395
	Equal variances not assumed			-1.692	21.420	.105	-.1750	.1034	-.3899	.0399
DurDSCs3	Equal variances assumed	2.023	.169	4.168	22	.000	.5500	.1320	.2763	.8237
	Equal variances not assumed			4.168	17.534	.001	.5500	.1320	.2722	.8278
DurNCs3	Equal variances assumed	.102	.752	5.338	22	.000	.5167	.0968	.3159	.7174
	Equal variances not assumed			5.338	21.916	.000	.5167	.0968	.3159	.7174

Change in Intensity: 12-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
IntAllWords3	1	12	3.075	1.0350	.2988
	2	12	1.525	.8966	.2588
IntSSCs3	1	12	4.192	1.8971	.5476
	2	12	4.075	2.3784	.6866
IntDSCs3	1	12	2.492	1.0405	.3004
	2	12	-.408	1.4909	.4304
IntNCs3	1	12	3.108	1.2930	.3732
	2	12	1.950	1.2288	.3547

Change in Intensity: 12-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
IntAllWords3	Equal variances assumed	.183	.673	3.921	22	.001	1.5500	.3953	.7302	2.3698
	Equal variances not assumed			3.921	21.562	.001	1.5500	.3953	.7293	2.3707
IntSSCs3	Equal variances assumed	1.459	.240	.133	22	.896	.1167	.8782	-1.7047	1.9380
	Equal variances not assumed			.133	20.964	.896	.1167	.8782	-1.7099	1.9433
IntDSCs3	Equal variances assumed	.375	.547	5.526	22	.000	2.9000	.5248	1.8116	3.9884
	Equal variances not assumed			5.526	19.661	.000	2.9000	.5248	1.8040	3.9960
IntNCs3	Equal variances assumed	.221	.643	2.250	22	.035	1.1583	.5149	.0904	2.2262
	Equal variances not assumed			2.250	21.943	.035	1.1583	.5149	.0903	2.2264

Change in F0: 12-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
FoAllWords3	1	12	1.083	9.9665	2.8771
	2	12	-1.317	4.2383	1.2235
FoSSCs3	1	12	15.458	6.0282	1.7402
	2	12	1.542	15.9052	4.5914
FoDSCs3	1	12	-.500	10.6640	3.0784
	2	12	-3.475	6.5748	1.8980
FoNCs3	1	12	-2.650	12.8804	3.7182
	2	12	-.908	8.7665	2.5307

Change in F0: 12-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FoAllWords3	Equal variances assumed	8.932	.007	.768	22	.451	2.4000	3.1264	-4.0838	8.8838
	Equal variances not assumed			.768	14.853	.455	2.4000	3.1264	-4.2696	9.0696
FoSSCs3	Equal variances assumed	4.203	.052	2.834	22	.010	13.9167	4.9101	3.7336	24.0997
	Equal variances not assumed			2.834	14.096	.013	13.9167	4.9101	3.3922	24.4411
FoDSCs3	Equal variances assumed	2.140	.158	.823	22	.420	2.9750	3.6165	-4.5251	10.4751
	Equal variances not assumed			.823	18.307	.421	2.9750	3.6165	-4.6138	10.5638
FoNCs3	Equal variances assumed	2.407	.135	-.387	22	.702	-1.7417	4.4977	-11.0694	7.5861
	Equal variances not assumed			-.387	19.391	.703	-1.7417	4.4977	-11.1427	7.6594

Vowel Duration Ratio: 6-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
DurAllWords3	1	12	1.442	.1379	.0398
	2	12	1.175	.1765	.0509
DurSSCs3	1	12	1.350	.2316	.0669
	2	12	1.525	.2734	.0789
DurDSCs3	1	12	1.383	.1642	.0474
	2	12	.783	.1992	.0575
DurNCs3	1	12	1.600	.2486	.0718
	2	12	1.267	.2741	.0791

Vowel Duration Ratio: 6-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DurAllWords3	Equal variances assumed	.164	.689	4.125	22	.000	.2667	.0646	.1326	.4007
	Equal variances not assumed			4.125	20.786	.000	.2667	.0646	.1321	.4012
DurSSCs3	Equal variances assumed	.280	.602	-1.692	22	.105	-.1750	.1034	-.3895	.0395
	Equal variances not assumed			-1.692	21.420	.105	-.1750	.1034	-.3899	.0399
DurDSCs3	Equal variances assumed	.540	.470	8.050	22	.000	.6000	.0745	.4454	.7546
	Equal variances not assumed			8.050	21.226	.000	.6000	.0745	.4451	.7549
DurNCs3	Equal variances assumed	.000	1.000	3.120	22	.005	.3333	.1068	.1118	.5549
	Equal variances not assumed			3.120	21.793	.005	.3333	.1068	.1116	.5550

Change in Intensity: 6-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
IntAllWords3	1	12	4.133	1.5059	.4347
	2	12	2.767	1.6177	.4670
IntSSCs3	1	12	4.192	1.8971	.5476
	2	12	4.075	2.3784	.6866
IntDSCs3	1	12	2.200	1.4104	.4071
	2	12	-.208	2.5650	.7404
IntNCs3	1	12	6.28	2.658	.767
	2	12	4.52	2.009	.580

Change in Intensity: 6-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
IntAllWords3	Equal variances assumed	.105	.749	2.142	22	.044	1.3667	.6380	.0435	2.6898
	Equal variances not assumed			2.142	21.888	.044	1.3667	.6380	.0431	2.6902
IntSSCs3	Equal variances assumed	1.459	.240	.133	22	.896	.1167	.8782	-1.7047	1.9380
	Equal variances not assumed			.133	20.964	.896	.1167	.8782	-1.7099	1.9433
IntDSCs3	Equal variances assumed	3.148	.090	2.850	22	.009	2.4083	.8450	.6559	4.1607
	Equal variances not assumed			2.850	17.094	.011	2.4083	.8450	.6263	4.1904
IntNCs3	Equal variances assumed	.699	.412	1.828	22	.081	1.758	.962	-.236	3.753
	Equal variances not assumed			1.828	20.477	.082	1.758	.962	-.245	3.762

Change in F0: 6-word set

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
FoAllWords3	1	12	11.833	7.7940	2.2499
	2	12	-1.483	10.7906	3.1150
FoSSCs3	1	12	15.458	6.0282	1.7402
	2	12	1.542	15.9052	4.5914
FoDSCs3	1	12	9.58	9.407	2.716
	2	12	-.67	11.150	3.219
FoNCs3	1	12	10.417	13.5191	3.9026
	2	12	-6.375	8.5043	2.4550

Change in F0: 6-word set

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FoAllWords3	Equal variances assumed	.677	.419	3.466	22	.002	13.3167	3.8426	5.3477	21.2857
	Equal variances not assumed			3.466	20.022	.002	13.3167	3.8426	5.3018	21.3316
FoSSCs3	Equal variances assumed	4.203	.052	2.834	22	.010	13.9167	4.9101	3.7336	24.0997
	Equal variances not assumed			2.834	14.096	.013	13.9167	4.9101	3.3922	24.4411
FoDSCs3	Equal variances assumed	.490	.491	2.434	22	.024	10.250	4.211	1.516	18.984
	Equal variances not assumed			2.434	21.393	.024	10.250	4.211	1.502	18.998
FoNCs3	Equal variances assumed	.490	.491	3.642	22	.001	16.7917	4.6106	7.2299	26.3534
	Equal variances not assumed			3.642	18.527	.002	16.7917	4.6106	7.1249	26.4584

Vowel Duration Ratio: 12-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
DurAllWords3	1	12	1.925	.1545	.0446
	2	12	1.633	.2060	.0595
DurSSCs3	1	12	1.350	.2316	.0669
	2	12	1.533	.2774	.0801
DurDSCs3	1	12	2.208	.2275	.0657
	2	12	1.925	.3793	.1095
DurNCs3	1	12	1.917	.2443	.0705
	2	12	1.492	.1975	.0570

Vowel Duration Ratio: 12-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DurAllWords3	Equal variances assumed	1.173	.290	3.924	22	.001	.2917	.0743	.1375	.4458
	Equal variances not assumed			3.924	20.401	.001	.2917	.0743	.1368	.4465
DurSSCs3	Equal variances assumed	.453	.508	-1.757	22	.093	-.1833	.1043	-.3997	.0330
	Equal variances not assumed			-1.757	21.320	.093	-.1833	.1043	-.4001	.0334
DurDSCs3	Equal variances assumed	6.238	.020	2.219	22	.037	.2833	.1277	.0186	.5481
	Equal variances not assumed			2.219	18.006	.040	.2833	.1277	.0151	.5516
DurNCs3	Equal variances assumed	.252	.620	4.686	22	.000	.4250	.0907	.2369	.6131
	Equal variances not assumed			4.686	21.075	.000	.4250	.0907	.2364	.6136

Change in Intensity: 12-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
IntAllWords3	1	12	3.075	1.0350	.2988
	2	12	1.958	1.0049	.2901
IntSSCs3	1	12	4.192	1.8971	.5476
	2	12	4.092	2.3608	.6815
IntDSCs3	1	12	2.492	1.0405	.3004
	2	12	.442	1.2450	.3594
IntNCs3	1	12	3.108	1.2930	.3732
	2	12	2.133	1.2309	.3553

Change in Intensity: 12-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
IntAllWords3	Equal variances assumed	.041	.842	2.681	22	.014	1.1167	.4164	.2530	1.9803
	Equal variances not assumed			2.681	21.981	.014	1.1167	.4164	.2530	1.9804
IntSSCs3	Equal variances assumed	1.357	.256	.114	22	.910	.1000	.8743	-1.7132	1.9132
	Equal variances not assumed			.114	21.026	.910	.1000	.8743	-1.7180	1.9180
IntDSCs3	Equal variances assumed	.488	.492	4.377	22	.000	2.0500	.4684	1.0786	3.0214
	Equal variances not assumed			4.377	21.328	.000	2.0500	.4684	1.0769	3.0231
IntNCs3	Equal variances assumed	.081	.778	1.892	22	.072	.9750	.5153	-.0937	2.0437
	Equal variances not assumed			1.892	21.947	.072	.9750	.5153	-.0939	2.0439

Change in F0: 12-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
FoAllWords3	1	12	1.083	9.9665	2.8771
	2	12	-.858	4.9687	1.4343
FoSSCs3	1	12	15.458	6.0282	1.7402
	2	12	1.542	15.9052	4.5914
FoDSCs3	1	12	-.500	10.6640	3.0784
	2	12	-2.708	7.6787	2.2167
FoNCs3	1	12	-2.650	12.8804	3.7182
	2	12	-.392	9.2541	2.6714

Change in F0: 12-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FoAllWords3	Equal variances assumed	6.313	.020	.604	22	.552	1.9417	3.2148	-4.7254	8.6087
	Equal variances not assumed			.604	16.150	.554	1.9417	3.2148	-4.8683	8.7516
FoSSCs3	Equal variances assumed	4.203	.052	2.834	22	.010	13.9167	4.9101	3.7336	24.0997
	Equal variances not assumed			2.834	14.096	.013	13.9167	4.9101	3.3922	24.4411
FoDSCs3	Equal variances assumed	1.362	.256	.582	22	.566	2.2083	3.7934	-5.6588	10.0755
	Equal variances not assumed			.582	19.990	.567	2.2083	3.7934	-5.7049	10.1216
FoNCs3	Equal variances assumed	1.970	.174	-.493	22	.627	-2.2583	4.5784	-11.7534	7.2367
	Equal variances not assumed			-.493	19.967	.627	-2.2583	4.5784	-11.8098	7.2931

Vowel Duration Ratio: 6-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
DurAllWords3	1	12	1.442	.1379	.0398
	2	12	1.275	.1658	.0479
DurSSCs3	1	12	1.350	.2316	.0669
	2	12	1.533	.2774	.0801
DurDSCs3	1	12	1.38	.164	.047
	2	10	.90	.231	.073
DurNCs3	1	12	1.600	.2486	.0718
	2	12	1.267	.2741	.0791

Vowel Duration Ratio: 6-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DurAllWords3	Equal variances assumed	.336	.568	2.677	22	.014	.1667	.0623	.0375	.2958
	Equal variances not assumed			2.677	21.292	.014	.1667	.0623	.0373	.2960
DurSSCs3	Equal variances assumed	.453	.508	-1.757	22	.093	-.1833	.1043	-.3997	.0330
	Equal variances not assumed			-1.757	21.320	.093	-.1833	.1043	-.4001	.0334
DurDSCs3	Equal variances assumed	2.885	.105	5.728	20	.000	.483	.084	.307	.659
	Equal variances not assumed			5.551	15.877	.000	.483	.087	.299	.668
DurNCs3	Equal variances assumed	.000	1.000	3.120	22	.005	.3333	.1068	.1118	.5549
	Equal variances not assumed			3.120	21.793	.005	.3333	.1068	.1116	.5550

Change in Intensity: 6-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
IntAllWords3	1	12	4.133	1.5059	.4347
	2	12	3.775	1.9982	.5768
IntSSCs3	1	12	4.192	1.8971	.5476
	2	12	4.092	2.3608	.6815
IntDSCs3	1	12	2.20	1.410	.407
	2	10	2.25	3.053	.965
IntNCs3	1	12	6.28	2.658	.767
	2	12	4.52	2.009	.580

Change in Intensity: 6-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
IntAllWords3	Equal variances assumed	1.568	.224	.496	22	.625	.3583	.7223	-1.1397	1.8563
	Equal variances not assumed			.496	20.448	.625	.3583	.7223	-1.1463	1.8629
IntSSCs3	Equal variances assumed	1.357	.256	.114	22	.910	.1000	.8743	-1.7132	1.9132
	Equal variances not assumed			.114	21.026	.910	.1000	.8743	-1.7180	1.9180
IntDSCs3	Equal variances assumed	9.709	.005	-.051	20	.960	-.050	.985	-2.104	2.004
	Equal variances not assumed			-.048	12.172	.963	-.050	1.048	-2.329	2.229
IntNCs3	Equal variances assumed	.699	.412	1.828	22	.081	1.758	.962	-.236	3.753
	Equal variances not assumed			1.828	20.477	.082	1.758	.962	-.245	3.762

Change in F0: 6-word set, correct productions

Group Statistics

	Language	N	Mean	Std. Deviation	Std. Error Mean
FoAllWords3	1	12	11.833	7.7940	2.2499
	2	12	-1.750	11.6840	3.3729
FoSSCs3	1	12	15.458	6.0282	1.7402
	2	12	1.542	15.9052	4.5914
FoDSCs3	1	12	9.58	9.407	2.716
	2	11	1.02	6.854	2.066
FoNCs3	1	12	10.417	13.5191	3.9026
	2	12	-6.375	8.5043	2.4550

Change in F0: 6-word set, correct productions

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FoAllWords3	Equal variances assumed	1.350	.258	3.350	22	.003	13.5833	4.0544	5.1749	21.9917
	Equal variances not assumed			3.350	19.171	.003	13.5833	4.0544	5.1024	22.0642
FoSSCs3	Equal variances assumed	4.203	.052	2.834	22	.010	13.9167	4.9101	3.7336	24.0997
	Equal variances not assumed			2.834	14.096	.013	13.9167	4.9101	3.3922	24.4411
FoDSCs3	Equal variances assumed	.299	.590	2.475	21	.022	8.565	3.460	1.369	15.761
	Equal variances not assumed			2.510	20.037	.021	8.565	3.412	1.448	15.682
FoNCs3	Equal variances assumed	.490	.491	3.642	22	.001	16.7917	4.6106	7.2299	26.3534
	Equal variances not assumed			3.642	18.527	.002	16.7917	4.6106	7.1249	26.4584

Appendix J: Non-parametric (Mann-Whitney U test) statistical output for production study from SPSS

12-word set

Ranks

	Language	N	Mean Rank	Sum of Ranks
DurAllWords3	1	12	17.33	208.00
	2	12	7.67	92.00
	Total	24		
DurDSCs3	1	12	16.71	200.50
	2	12	8.29	99.50
	Total	24		
FoNCs3	1	12	12.42	149.00
	2	12	12.58	151.00
	Total	24		

Test Statistics[b]

	DurAllWords3	DurDSCs3	FoNCs3
Mann-Whitney U	14.000	21.500	71.000
Wilcoxon W	92.000	99.500	149.000
Z	-3.388	-2.926	-.058
Asymp. Sig. (2-tailed)	.001	.003	.954
Exact Sig. [2*(1-tailed Sig.)]	.000[a]	.002[a]	.977[a]

a. Not corrected for ties.

b. Grouping Variable: Language

6-word set

Ranks

	Language	N	Mean Rank	Sum of Ranks
DurDSCs3	1	12	18.50	222.00
	2	12	6.50	78.00
	Total	24		
DurNCs3	1	12	16.83	202.00
	2	12	8.17	98.00
	Total	24		

Test Statistics[b]

	DurDSCs3	DurNCs3
Mann-Whitney U	.000	20.000
Wilcoxon W	78.000	98.000
Z	-4.184	-3.042
Asymp. Sig. (2-tailed)	.000	.002
Exact Sig. [2*(1-tailed Sig.)]	.000[a]	.002[a]

a. Not corrected for ties.

b. Grouping Variable: Language

12-word set, correct productions

Ranks

	Language	N	Mean Rank	Sum of Ranks
DurAllWords3	1	12	16.71	200.50
	2	12	8.29	99.50
	Total	24		
FoDSCs3	1	12	13.96	167.50
	2	12	11.04	132.50
	Total	24		
FoNCs3	1	12	12.17	146.00
	2	12	12.83	154.00
	Total	24		

Test Statistics[b]

	DurAllWords3	FoDSCs3	FoNCs3
Mann-Whitney U	21.500	54.500	68.000
Wilcoxon W	99.500	132.500	146.000
Z	-2.961	-1.011	-.231
Asymp. Sig. (2-tailed)	.003	.312	.817
Exact Sig. [2*(1-tailed Sig.)]	.002[a]	.319[a]	.843[a]

a. Not corrected for ties.

b. Grouping Variable: Language

6-word set, correct productions

Ranks

	Language	N	Mean Rank	Sum of Ranks
DurSSCs3	1	12	10.21	122.50
	2	12	14.79	177.50
	Total	24		
DurDSCs3	1	12	16.17	194.00
	2	10	5.90	59.00
	Total	22		
DurNCs3	1	12	16.83	202.00
	2	12	8.17	98.00
	Total	24		
FoSSCs3	1	12	16.92	203.00
	2	12	8.08	97.00
	Total	24		

Test Statistics[b]

	DurSSCs3	DurDSCs3	DurNCs3	FoSSCs3
Mann-Whitney U	44.500	4.000	20.000	19.000
Wilcoxon W	122.500	59.000	98.000	97.000
Z	-1.604	-3.744	-3.042	-3.063
Asymp. Sig. (2-tailed)	.109	.000	.002	.002
Exact Sig. [2*(1-tailed Sig.)]	.114[a]	.000[a]	.002[a]	.001[a]

a. Not corrected for ties.

b. Grouping Variable: Language

Appendix K: Variability in F0 production by native speakers in production study

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	October	SSC	33	0	2	1	36
FNS	October	SSC	23	2	9	2	36
Total			**56**	**2**	**11**	**3**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	radio	SSC	26	1	9	0	36
FNS	radio	SSC	24	3	9	0	36
Total			**50**	**4**	**18**	**0**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	bipolar	DSC	18	3	10	5	36
FNS	bipolar	DSC	11	2	17	6	36
Total			**29**	**5**	**27**	**11**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	engineer	DSC	14	5	8	9	36
FNS	engineer	DSC	4	23	5	4	36
Total			**18**	**28**	**13**	**13**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	introduce	DSC	11	7	17	1	36
FNS	introduce	DSC	1	30	3	2	36
Total			**12**	**37**	**20**	**3**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	lunatic	DSC	32	0	4	0	36
FNS	lunatic	DSC	22	4	6	4	36
Total			**54**	**4**	**10**	**4**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	develop	NC	0	6	20	10	36
FNS	develop	NC	0	19	11	6	36
Total			**0**	**25**	**31**	**16**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	disagree	NC	0	16	7	13	36
FNS	disagree	NC	2	26	1	7	36
Total			**2**	**42**	**8**	**20**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	easily	NC	26	4	6	0	36
FNS	easily	NC	14	16	6	0	36
Total			**40**	**20**	**12**	**0**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	location	NC	22	1	7	6	36
FNS	location	NC	16	4	13	3	36
Total			**38**	**5**	**20**	**9**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	management	NC	18	0	14	4	36
FNS	management	NC	18	2	13	3	36
Total			**36**	**2**	**27**	**7**	72

speaker	word	type	**>5Hz**	**<5Hz**	**wi 5Hz**	**excluded**	**total**
MNS	unafraid	NC	9	13	14	0	36
FNS	unafraid	NC	0	36	0	0	36
Total			**9**	**49**	**14**	**0**	72

Appendix L: Background questionnaire for listening experiment

Language background questionnaire

1. **First Name only ___________________________________**
(Only the code FNL "Female native listener" or MNL "Male native listener" + participant number will be used in all data records and writings, e.g. FNL2, MNL3)

2. Age ________

3. Where are you from? ___________________________________

4. How long have you lived in Albuquerque? __________________

5. Besides English, what other language(s) do you speak?

   ____

   ______

6. What is your highest level of education?

   ___ completed high school

   ___ some college

   ___ completed a technical or vocational school program

   ___ completed an undergraduate program at a university

   ___ completed a graduate program at a university

7. Are you aware of any speech impairment you might have? (circle/highlight one)

   yes   no

8. Are you aware of any hearing impairment you might have? (circle/highlight one)

   yes   no

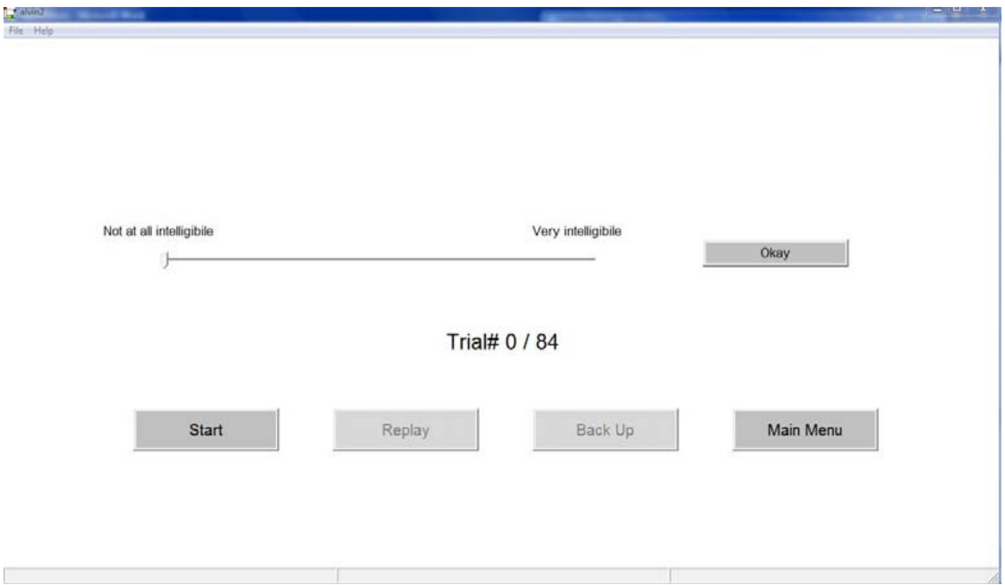
Appendix M: Listener demographics

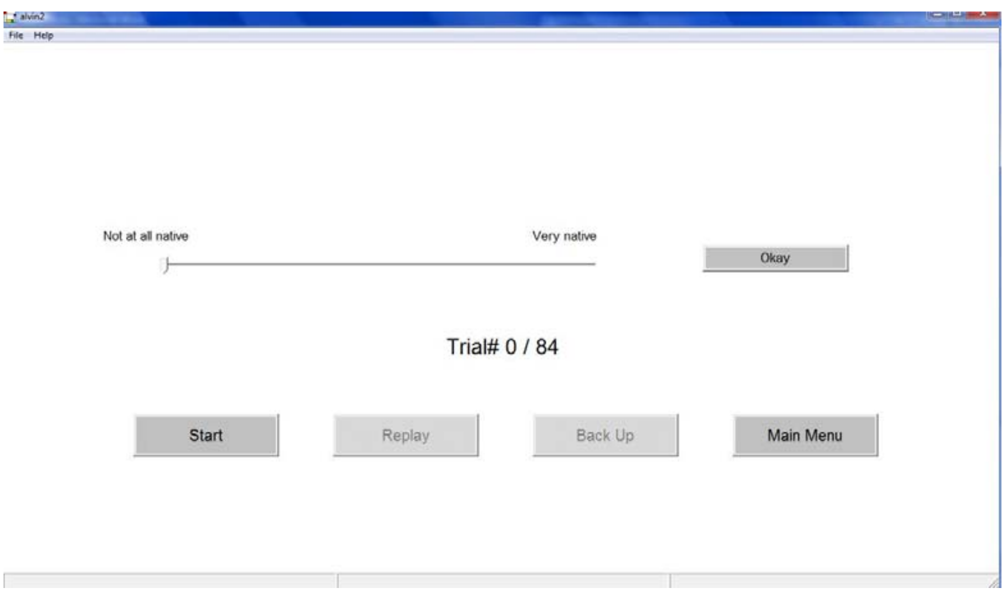
Key: ML = Male listener; FL = Female listener

Listener	From	Age	Years in NM	Other languages
FL1	Missouri	33	9	German
FL2	n/a - USA	51	1	Russian, Chinese, Korean, French
FL3	New Mexico	21	3	
FL4	Texas	29	0.75	Spanish, ASL, Arabic
FL5	New Mexico	23	17	Spanish
FL6	Illinois	24	2	Spanish
FL7	Illinois	24	1	Spanish
FL8	Ohio	31	7	Spanish
FL9	New Mexico	29	12	French
FL10	California	43	12	Spanish, Portuguese, Apache
FL11	Illinois	29	0.75	Spanish
ML1	Massachusetts	27	5	
ML2	Louisiana	33	0.25	Portuguese
ML3	New Mexico	29	29	
ML4	New Mexico	27	2	Spanish, Dutch, Chinese
ML5	New Mexico	53	10	Spanish
ML6	Colorado	40	0.75	Spanish, Italian, French, Swedish
ML7	Indiana	26	3	French, German, Spanish
ML8	Michigan	34	4	
ML9	Illinois	36	11	French
ML10	Washington	27	2	Spanish
Note: FL7 was excluded from the analysis				

Appendix N: Screen shots of rating scales for intelligibility and nativeness ratings

The participants were instructed to rate each token for intelligibility and nativeness by moving the slider on the unmarked scale. When doing so, the participant was asked to consider "How intelligible does the speaker sound" and "How much does the speaker sound like a native speaker of English?"





Appendix O: Statistical output for one-way ANOVA for listening study from SPSS

One-way ANOVA, Listening experiment

Intelligibility:

Itype1 = Condition one, flat
Itype2 = Condition 2, Increase duration
Itype3 = Condition 3, Decrease duration
Itype4 = Condition 4, Increase Intensity
Itype5 = Condition 5, Decrease Intensity
Itype6 = Condition 6, Increase F0
Itype7 = Condition 7, Decrease F0

Descriptive Statistics

	Mean	Std. Deviation	N
Itype1	513.95	100.253	20
Itype2	488.65	117.235	20
Itype3	368.95	125.754	20
Itype4	526.75	90.994	20
Itype5	495.60	107.751	20
Itype6	481.85	114.874	20
Itype7	465.60	131.996	20

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
factor1	Pillai's Trace	.916	25.519[a]	6.000	14.000	.000	.916
	Wilks' Lambda	.084	25.519[a]	6.000	14.000	.000	.916
	Hotelling's Trace	10.937	25.519[a]	6.000	14.000	.000	.916
	Roy's Largest Root	10.937	25.519[a]	6.000	14.000	.000	.916

a. Exact statistic

b. Design: Intercept
Within Subjects Design: factor1

Pairwise Comparisons

Measure:MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a] Lower Bound	Upper Bound
1	2	25.300	16.397	1.000	-32.132	82.732
	3	145.000*	16.988	.000	85.497	204.503
	4	-12.800	8.770	1.000	-43.517	17.917
	5	18.350	11.346	1.000	-21.391	58.091
	6	32.100	13.814	.659	-16.287	80.487
	7	48.350	14.071	.058	-.934	97.634
2	1	-25.300	16.397	1.000	-82.732	32.132
	3	119.700*	9.022	.000	88.099	151.301
	4	-38.100	14.577	.359	-89.159	12.959
	5	-6.950	14.332	1.000	-57.150	43.250
	6	6.800	9.313	1.000	-25.821	39.421
	7	23.050	12.994	1.000	-22.464	68.564
3	1	-145.000*	16.988	.000	-204.503	-85.497
	2	-119.700*	9.022	.000	-151.301	-88.099
	4	-157.800*	16.626	.000	-216.036	-99.564
	5	-126.650*	14.666	.000	-178.019	-75.281
	6	-112.900*	10.940	.000	-151.217	-74.583
	7	-96.650*	13.608	.000	-144.313	-48.987
4	1	12.800	8.770	1.000	-17.917	43.517
	2	38.100	14.577	.359	-12.959	89.159
	3	157.800*	16.626	.000	99.564	216.036
	5	31.150*	8.313	.029	2.032	60.268
	6	44.900*	12.433	.039	1.350	88.450
	7	61.150*	14.579	.010	10.086	112.214
5	1	-18.350	11.346	1.000	-58.091	21.391
	2	6.950	14.332	1.000	-43.250	57.150

	3	126.650*	14.666	.000	75.281	178.019
	4	-31.150*	8.313	.029	-60.268	-2.032
	6	13.750	11.855	1.000	-27.773	55.273
	7	30.000	11.949	.446	-11.853	71.853
6	1	-32.100	13.814	.659	-80.487	16.287
	2	-6.800	9.313	1.000	-39.421	25.821
	3	112.900*	10.940	.000	74.583	151.217
	4	-44.900*	12.433	.039	-88.450	-1.350
	5	-13.750	11.855	1.000	-55.273	27.773
	7	16.250	12.048	1.000	-25.951	58.451
7	1	-48.350	14.071	.058	-97.634	.934
	2	-23.050	12.994	1.000	-68.564	22.464
	3	96.650*	13.608	.000	48.987	144.313
	4	-61.150*	14.579	.010	-112.214	-10.086
	5	-30.000	11.949	.446	-71.853	11.853
	6	-16.250	12.048	1.000	-58.451	25.951

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

*. The mean difference is significant at the .05 level.

One-way ANOVA, Listening experiment

Nativeness:

Ntype1 = Condition one, flat
Ntype2 = Condition 2, Increase duration
Ntype3 = Condition 3, Decrease duration
Ntype4 = Condition 4, Increase Intensity
Ntype5 = Condition 5, Decrease Intensity
Ntype6 = Condition 6, Increase F0
Ntype7 = Condition 7, Decrease F0

Descriptive Statistics

	Mean	Std. Deviation	N
Ntype1	451.35	73.384	20
Ntype2	443.40	91.054	20
Ntype3	276.65	90.168	20
Ntype4	451.40	74.542	20
Ntype5	427.40	73.438	20
Ntype6	408.90	111.626	20
Ntype7	418.50	103.598	20

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
factor1	Pillai's Trace	.934	33.173[a]	6.000	14.000	.000	.934
	Wilks' Lambda	.066	33.173[a]	6.000	14.000	.000	.934
	Hotelling's Trace	14.217	33.173[a]	6.000	14.000	.000	.934
	Roy's Largest Root	14.217	33.173[a]	6.000	14.000	.000	.934

a. Exact statistic

b. Design: Intercept
Within Subjects Design: factor1

Pairwise Comparisons

Measure:MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	7.950	17.245	1.000	-52.453	68.353
	3	174.700*	16.581	.000	116.623	232.777
	4	-.050	10.410	1.000	-36.511	36.411
	5	23.950	15.992	1.000	-32.066	79.966
	6	42.450	24.175	1.000	-42.227	127.127
	7	32.850	16.713	1.000	-25.690	91.390
2	1	-7.950	17.245	1.000	-68.353	52.453
	3	166.750*	14.565	.000	115.734	217.766
	4	-8.000	13.706	1.000	-56.008	40.008
	5	16.000	18.994	1.000	-50.530	82.530
	6	34.500	21.336	1.000	-40.234	109.234
	7	24.900	19.382	1.000	-42.989	92.789
3	1	-174.700*	16.581	.000	-232.777	-116.623
	2	-166.750*	14.565	.000	-217.766	-115.734
	4	-174.750*	13.573	.000	-222.292	-127.208
	5	-150.750*	19.164	.000	-217.875	-83.625
	6	-132.250*	16.500	.000	-190.044	-74.456
	7	-141.850*	17.577	.000	-203.415	-80.285
4	1	.050	10.410	1.000	-36.411	36.511
	2	8.000	13.706	1.000	-40.008	56.008
	3	174.750*	13.573	.000	127.208	222.292
	5	24.000	17.725	1.000	-38.085	86.085
	6	42.500	21.662	1.000	-33.374	118.374
	7	32.900	17.238	1.000	-27.479	93.279
5	1	-23.950	15.992	1.000	-79.966	32.066
	2	-16.000	18.994	1.000	-82.530	50.530

	3	150.750*	19.164	.000	83.625	217.875
	4	-24.000	17.725	1.000	-86.085	38.085
	6	18.500	22.378	1.000	-59.882	96.882
	7	8.900	17.373	1.000	-51.952	69.752
6	1	-42.450	24.175	1.000	-127.127	42.227
	2	-34.500	21.336	1.000	-109.234	40.234
	3	132.250*	16.500	.000	74.456	190.044
	4	-42.500	21.662	1.000	-118.374	33.374
	5	-18.500	22.378	1.000	-96.882	59.882
	7	-9.600	23.436	1.000	-91.690	72.490
7	1	-32.850	16.713	1.000	-91.390	25.690
	2	-24.900	19.382	1.000	-92.789	42.989
	3	141.850*	17.577	.000	80.285	203.415
	4	-32.900	17.238	1.000	-93.279	27.479
	5	-8.900	17.373	1.000	-69.752	51.952
	6	9.600	23.436	1.000	-72.490	91.690

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

*. The mean difference is significant at the .05 level.

Appendix P: Correlation analysis of intelligibility and nativeness

Descriptive Statistics

	Mean	Std. Deviation	N
Itype1	513.95	100.253	20
Ntype1	451.35	73.384	20

Correlations

		Itype1	Ntype1
Itype1	Pearson Correlation	1	.594**
	Sig. (2-tailed)		.006
	N	20	20
Ntype1	Pearson Correlation	.594**	1
	Sig. (2-tailed)	.006	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype2	488.65	117.235	20
Ntype2	443.40	91.054	20

Correlations

		Itype2	Ntype2
Itype2	Pearson Correlation	1	.713**
	Sig. (2-tailed)		.000
	N	20	20
Ntype2	Pearson Correlation	.713**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype3	368.95	125.754	20
Ntype3	276.65	90.168	20

Correlations

		Itype3	Ntype3
Itype3	Pearson Correlation	1	.638**
	Sig. (2-tailed)		.002
	N	20	20
Ntype3	Pearson Correlation	.638**	1
	Sig. (2-tailed)	.002	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype4	526.75	90.994	20
Ntype4	451.40	74.542	20

Correlations

		Itype4	Ntype4
Itype4	Pearson Correlation	1	.727**
	Sig. (2-tailed)		.000
	N	20	20
Ntype4	Pearson Correlation	.727**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype5	495.60	107.751	20
Ntype5	427.40	73.438	20

Correlations

		Itype5	Ntype5
Itype5	Pearson Correlation	1	.496*
	Sig. (2-tailed)		.026
	N	20	20
Ntype5	Pearson Correlation	.496*	1
	Sig. (2-tailed)	.026	
	N	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype6	481.85	114.874	20
Ntype6	408.90	111.626	20

Correlations

		Itype6	Ntype6
Itype6	Pearson Correlation	1	.762**
	Sig. (2-tailed)		.000
	N	20	20
Ntype6	Pearson Correlation	.762**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Descriptive Statistics

	Mean	Std. Deviation	N
Itype7	465.60	131.996	20
Ntype7	418.50	103.598	20

Correlations

		Itype7	Ntype7
Itype7	Pearson Correlation	1	.690**
	Sig. (2-tailed)		.001
	N	20	20
Ntype7	Pearson Correlation	.690**	1
	Sig. (2-tailed)	.001	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix Q: Statistical output from ICC for Inter-rater reliability

Intelligibility:

Case Processing Summary

		N	%
Cases	Valid	7	46.7
	Excluded[a]	8	53.3
	Total	15	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.966	19

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.597[b]	.357	.882	29.148	6	108	.000
Average Measures	.966[c]	.913	.993	29.148	6	108	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Inter-rater reliability ICC

Nativeness:

Case Processing Summary

		N	%
Cases	Valid	7	100.0
	Excluded[a]	0	.0
	Total	7	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.956	20

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.522[b]	.287	.847	22.830	6	114	.000
Average Measures	.956[c]	.890	.991	22.830	6	114	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Appendix R: Intra-rater Reliability

FL1 = Female Listener 1, ML1 = Male Listener 1, etc.
"First" = first listening of token
"Repeat" = second listening of token

Correlations

		FL1_first	FL1_repeat
FL1_first	Pearson Correlation	1	.302
	Sig. (2-tailed)		.294
	N	14	14
FL1_repeat	Pearson Correlation	.302	1
	Sig. (2-tailed)	.294	
	N	14	14

Correlations

		FL2_first	FL2_repeat
FL2_first	Pearson Correlation	1	.530
	Sig. (2-tailed)		.051
	N	14	14
FL2_repeat	Pearson Correlation	.530	1
	Sig. (2-tailed)	.051	
	N	14	14

Correlations

		FL3_first	FL3_repeat
FL3_first	Pearson Correlation	1	.315
	Sig. (2-tailed)		.273
	N	14	14
FL3_repeat	Pearson Correlation	.315	1
	Sig. (2-tailed)	.273	
	N	14	14

Correlations

		FL4_first	FL4_repeat
FL4_first	Pearson Correlation	1	.687**
	Sig. (2-tailed)		.007
	N	14	14
FL4_repeat	Pearson Correlation	.687**	1
	Sig. (2-tailed)	.007	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		FL5_first	FL5_repeat
FL5_first	Pearson Correlation	1	.921**
	Sig. (2-tailed)		.000
	N	14	14
FL5_repeat	Pearson Correlation	.921**	1
	Sig. (2-tailed)	.000	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		FL6_first	FL6_repeat
FL6_first	Pearson Correlation	1	.591*
	Sig. (2-tailed)		.026
	N	14	14
FL6_repeat	Pearson Correlation	.591*	1
	Sig. (2-tailed)	.026	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		FL8_first	FL8_repeat
FL8_first	Pearson Correlation	1	.651*
	Sig. (2-tailed)		.012
	N	14	14
FL8_repeat	Pearson Correlation	.651*	1
	Sig. (2-tailed)	.012	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		FL9_first	FL9_repeat
FL9_first	Pearson Correlation	1	.578*
	Sig. (2-tailed)		.030
	N	14	14
FL9_repeat	Pearson Correlation	.578*	1
	Sig. (2-tailed)	.030	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		FL10_first	FL10_repeat
FL10_first	Pearson Correlation	1	.251
	Sig. (2-tailed)		.387
	N	14	14
FL10_repeat	Pearson Correlation	.251	1
	Sig. (2-tailed)	.387	
	N	14	14

Correlations

		FL11_first	FL11_repeat
FL11_first	Pearson Correlation	1	.282
	Sig. (2-tailed)		.328
	N	14	14
FL11_repeat	Pearson Correlation	.282	1
	Sig. (2-tailed)	.328	
	N	14	14

Correlations

		ML1_first	ML1_repeat
ML1_first	Pearson Correlation	1	.404
	Sig. (2-tailed)		.152
	N	14	14
ML1_repeat	Pearson Correlation	.404	1
	Sig. (2-tailed)	.152	
	N	14	14

Correlations

		ML2_first	ML2_repeat
ML2_first	Pearson Correlation	1	.621*
	Sig. (2-tailed)		.018
	N	14	14
ML2_repeat	Pearson Correlation	.621*	1
	Sig. (2-tailed)	.018	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		ML3_first	ML3_repeat
ML3_first	Pearson Correlation	1	.721**
	Sig. (2-tailed)		.004
	N	14	14
ML3_repeat	Pearson Correlation	.721**	1
	Sig. (2-tailed)	.004	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		ML4_first	ML4_repeat
ML4_first	Pearson Correlation	1	.718**
	Sig. (2-tailed)		.004
	N	14	14
ML4_repeat	Pearson Correlation	.718**	1
	Sig. (2-tailed)	.004	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		ML5_first	ML5_repeat
ML5_first	Pearson Correlation	1	.879**
	Sig. (2-tailed)		.000
	N	14	14
ML5_repeat	Pearson Correlation	.879**	1
	Sig. (2-tailed)	.000	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		ML6_first	ML6_repeat
ML6_first	Pearson Correlation	1	.255
	Sig. (2-tailed)		.379
	N	14	14
ML6_repeat	Pearson Correlation	.255	1
	Sig. (2-tailed)	.379	
	N	14	14

Correlations

		ML7_first	ML7_repeat
ML7_first	Pearson Correlation	1	.735**
	Sig. (2-tailed)		.003
	N	14	14
ML7_repeat	Pearson Correlation	.735**	1
	Sig. (2-tailed)	.003	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		ML8_first	ML8_repeat
ML8_first	Pearson Correlation	1	.318
	Sig. (2-tailed)		.267
	N	14	14
ML8_repeat	Pearson Correlation	.318	1
	Sig. (2-tailed)	.267	
	N	14	14

Correlations

		ML9_first	ML9_repeat
ML9_first	Pearson Correlation	1	.659*
	Sig. (2-tailed)		.010
	N	14	14
ML9_repeat	Pearson Correlation	.659*	1
	Sig. (2-tailed)	.010	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

		ML10_first	ML10_repeat
ML10_first	Pearson Correlation	1	.880**
	Sig. (2-tailed)		.000
	N	14	14
ML10_repeat	Pearson Correlation	.880**	1
	Sig. (2-tailed)	.000	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

REFERENCES

Anani, M. (1989). Incorrect stress placement in the case of Arab learners of English. *International Review of Applied Linguistics in Language Teaching (IRAL), 27*(1), 15-21.

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

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris Publications.

Benrabah, M. (1997). Word stress: A source of unintelligibility in English. *International Review of Applied Linguistics in Language Teaching (IRAL), 35*(3), 157-166.

Berg, K. E., & Latin, R. W. (2004). *Essentials of research methods in health, physical education, exercise science, and recreation* (2nd ed.). Baltimore: Lippincott Williams and Wilkins.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer (Version 4.4.04).

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287-308). Amsterdam: John Benjamins Publishing Company.

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

Brookshire, R. H. (2003). *Introduction to neurogenic communication disorders*. St. Louis, MO: Mosby.

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics and Phonetics, 15*(3), 181-193.

Bunton, K., Kent, R. D., Kent, J. F., & Rosenbek, J. C. (2000). Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clinical Linguistics and Phonetics, 14*(1), 13-24.

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

Chun, D. M., Hardison, D. M., & Pennington, M. C. (2008). Technologies for prosody in context: Past and future of L2 research and practice. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 323-346). Amsterdam: John Benjamins Publishing Company.

Costello, R. B. (Ed.). (1992). *The American Heritage Dictionary* (Third ed.). New York: Dell Publishing.

Derwing, T. M. (2003). What do ESL students say about their accents? *The Canadian Modern Language Review, 59*(4), 547-566.

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 347-369). Amsterdam: John Benjamins Publishing Company.

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., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393-410.

Duffy, J. R. (2005). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier Mosby.

Fairbanks, G. (1960). *Voice and articulation drillbook* (Second ed.). New York: Harper and Row.

Fernández Planas, A. M., & Martínez Celdrán, E. (2003). El tono fundamental y la duración: Dos aspectos de la taxonomía prosódica en dos modalidades de habla (enunciative e interrogativa) del español. *Estudios de fonética experimental, 12,* 165-200.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of Acoustic Society of America, 76*(3), 692-707.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of Acoustic Society of America, 93*(3), 1589-1608.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-277). Baltimore: York Press.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995a). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1-26.

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

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of Acoustic Society of America, 27*(4), 155-158.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 126-152.

Gilbert, J. B. (1993). *Clear speech: Pronunciation and listening comprehension in North American English* (Second ed.). Cambridge: Cambridge University Press.

Grabe, E., Rosner, B. S., García-Albea, J. E., & Zhou, X. (2003). Perception of English intonation by English, Spanish, and Chinese listeners. *Language and Speech, 46*(4), 375-401.

Grant, L. (2000). *Well Said*. Boston: Heinle and Heinle.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition, 7*(3), 207-226.

Hahn, L. (1999). *Native speakers' reaction to non-native stress in English discourse.* Unpublished Doctoral dissertation, University of Urbana-Champaign, Urbana-Champaign, IL.

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

Hewings, M. (2004). *Pronunciation practice activities: A resource book for teaching English pronunciation*. Cambridge: Cambridge University Press.

Hewings, M., & Goldstein, S. (1999). *Pronunciation plus: Practice through interaction (North American English)*. Cambridge: Cambridge University Press.

Hillenbrand, J. M., & Gayvert, R. T. (2009). Alvin (Version 2 (beta)).

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of Acoustic Society of America, 97*(5), 3099-3111.

Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge: Cambridge University Press.

Hubicka, O. (1981). Phonology: Stress. *Practical English Teaching, 1*(3), 20-23.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (Vol. 269-293). Harmondsworth: Penguin.

Jenkins, J. (1998). Which pronunciation norms and models for English as an International Language. *ELT Journal, 52*(2), 119-126.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics in Language Teaching, 28*(2), 99-118.

Kiesling, S., Dilley, L., & Raymond, W. D. (2006). *The variation in conversation (ViC) project: Creation of the Buckeye Corpus of Conversational Speech*. Columbus, OH: The Ohio State University.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Ladefoged, P. (1993). *A course in phonetics* (3rd ed.). Fort Worth: Harcourt Brace Jovanovich, Inc.

Lehiste, I. (1996). Suprasegmental features of speech. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 226-244). St. Louis, MO: Mosby.

Lehiste, I., & Fox, R. A. (1992). Perception of prominence by Estonian and English listeners. *Language and Speech, 35*(4), 419-434.

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

Llisterri, J., Machuca, M., de la Mota, C., Riera, M., & Ríos, A. (2003). *The perception of lexical stress in Spanish.* Paper presented at the International Congress of Phonetic Science, Barcelona.

Major, R. C. (2008). Transfer in second language phonology: A review. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 63-94). Amsterdam: John Benjamins Publishing Company.

Moore, B. J. (2003). *An introduction to the psychology of hearing* (5th ed.). Oxford: Elsevier.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly*, *25*(3), 481-520.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453-467.

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

Munro, M. J., & Derwing, T. M. (1995). 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. *Studies in second language acquisition, 23*, 451-468.

Okobi, A. O. (2006). *Acoustic correlates of word stress in American English.* Unpublished Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge.

Olsen, A. E., & Biley, P. C. (2002a). *Academic vocabulary: Academic words*. New York: Longman.

Olsen, A. E., & Biley, P. C. (2002b). *Active vocabulary: General and academic words*. New York: Longman.

Orion, G. F. (1997). *Pronouncing American English: Sounds, stress, and intonation* (Second ed.). Boston: Heinle and Heinle.

Ortega-Llebaría, M. (2006). Phonetic cues to stress and accent in Spanish. In M. Díaz-Campos (Ed.), *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology* (pp. 104-118). Somerville, MA: Cascadilla Proceedings Project.

Pallant, J. (2007). *SPSS Survival Manual* (2nd ed.). New York: McGraw-Hill.

¿Qué hay de nuevo? Learn what's new in serving those of diverse backgrounds. (2006, February). *NSSLHA Now (The newsletter of the national student speech language hearing association,* 3-4.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (Second ed.). Cambridge: Cambridge University Press.

Rogers, C. L., & Dalby, J. (2005). Forced-Choice Analysis of Segmental Production by Chinese-Accented English Speakers. *Journal of Speech, Language, and Hearing Research, 48*, 306-322.

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly, 40*(4), 715-737.

Schairer, K. E. (1992). Native speaker reaction to non-native speech. *The Modern Language Journal, 76*(3), 309-319.

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

Seddoh, S. A. (2004). Prosodic disturbance in aphasia: Speech timing versus intonation production. *Clinical Linguistics and Phonetics, 18*(1), 17-38.

Shah, A. (2002). *Temporal characteristics of Spanish-accented English: Acoustic measures and their correlation with accentedness ratings.* Unpublished Dissertation, The City University of New York, New York City.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193-247.

Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34*, 285-293.

Spoelders, M., Guiora, A. Z., Van Leeuwen, K., Meuleman, C., Van Besien, F., Vyt, A., et al. (1996). A psycholinguistic study of pronunciation of bilinguals: The case of Belgium. In T. Hickey & J. Williams (Eds.), *Language, education, and society in a changing world* (pp. 117-126). Cleveland: Multilingual Matters.

SPSS. (2009). SPSS: Statistical Package for Social Sciences (Version 17).

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

Tarone, E. (2005). Speaking in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 485-502). Mahway, NJ: Lawrence Erlbaum Associates.

Timmis, I. (2002). Native-speaker norms and International English: A classroom view. *ELT Journal*, *56*(3), 240-249.

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research, 47*, 766-783.

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *Journal of Acoustic Society of America, 99*(6), 3782-3790.

United States Census 2000. Retrieved April 5, 2006, from http://www.census.gov

United States Department of Labor, Occupational Safety and Health Administration; Appendix to Section 1910.134: Fit Testing Procedures. (2009). Retrieved April 17, 2009, from http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=9780

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in second language acquisition*, *4*(2), 114-136.

Wenk, B. J. (1985). Speech rhythms in second language acquisition. *Language and Speech*, *28*(2), 157-175.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition*, *20*, 1-25.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. New York: Oxford University Press.

Zampini, M. L. (2008). L2 speech production research: Findings, issues, and advances. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219-249). Amsterdam: John Benjamins Publishing Company.

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