
Carl Tyler Johnson

Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

Recommended Citation
https://docs.lib.purdue.edu/open_access_dissertations/1573

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair
    Department of English
Dr. Mary Niepokuj
    Department of English
Dr. Margie Berns
    Department of English
Dr. Amy Neel
    Department of Speech and Hearing Sciences, University of New Mexico

Approved by:

Dr. Dorsey Armstrong
    Head of the Graduate Program
Dedicated to Cris Elder
ACKNOWLEDGMENTS

I would like to thank my advisor and chair, April Ginther, for guiding me through this process, and my committee members, Mary Niepokuj, Margie Berns, and Amy Neel for helpful suggestions. I would like to thank Chris Koops for help with Praat scripts and statistics, Susan Paine, John Pesko and Anastasiia Kim for help with statistical analyses, and Chuck Paine for the computer I used to write this dissertation.

Thank you to Willa, Mo, Thom, Wolfgang, Emma, Ruby, Sona, Issa, Leo, Oscar, Moose, Tilley, Guy Noir, and Werner. Thank you to Cris for support and endurance.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................................................... viii
LIST OF FIGURES ......................................................................................................................................... ix
GLOSSARY .................................................................................................................................................... x
LIST OF ABBREVIATIONS................................................................................................................................ xi
LIST OF SYMBOLS ..................................................................................................................................... xiii
ABSTRACT ................................................................................................................................................... xiv

CHAPTER 1. INTRODUCTION .................................................................................................................. 1
  Motivation for the study ............................................................................................................................... 2
  Acoustic terminology .................................................................................................................................. 4
  Previous studies .......................................................................................................................................... 6
  Outline of the study .................................................................................................................................... 12

CHAPTER 2. OVERVIEW OF PARTICIPANTS ......................................................................................... 13
  Rhythmic classes of languages .................................................................................................................... 14
  The Context of the Data ............................................................................................................................... 18
    “The Foreign TA Problem” in the United States ......................................................................................... 19
    Purdue’s response to the conflict .............................................................................................................. 21
    Purdue’s OEPT ........................................................................................................................................ 25
  Prosody of ITAs ......................................................................................................................................... 28
    ITAs and sentence stress ............................................................................................................................ 28
    ITAs and intonation ................................................................................................................................... 29
    ITAs and proficiency test criteria ............................................................................................................. 31

CHAPTER 3. AN INFORMAL DEFINITION OF PROSODY ........................................................................ 33
  Definitions of prosody ................................................................................................................................. 34
  Definitions of prosodic terminology ........................................................................................................... 38
    Lexical stress/prosodic word stress .......................................................................................................... 38
    Phrasal stress/prosodic phrasal stress ....................................................................................................... 39
    Sentence stress/intonational phrase ......................................................................................................... 39
    Focus ........................................................................................................................................................ 40
Prosodic prominence ........................................................................................................ 42
The Role of Prosody in Comprehending Speech: A perceptual description of prosody .... 43
Word detection within a lexical word .............................................................................. 44
Word detection in the prosodic phrase ......................................................................... 45
Prosody and the given-new contract ............................................................................ 46
Word detection in the intonational phrase ..................................................................... 48
Summary of the perceptual importance of prosody ....................................................... 49
An Acoustic Description of Prosody ............................................................................ 50
Problems in acoustic studies of prosody ....................................................................... 50
Acoustic measurements of lexical stress ....................................................................... 51
Measurements of primary lexical stress, secondary lexical stress, and no stress .......... 53
Measurements of sentence stress and prosodic prominence ........................................ 54
Summary of description of acoustic prosody ................................................................. 56
A Phonological Description of Prosody ....................................................................... 57
Information structure, function and content words ....................................................... 59
Prosodic phonology in Generative Grammar ............................................................... 60
Prosodic phonology in Optimality Theory ................................................................... 61
Summary of relevant findings from perceptual, acoustic and phonological descriptions of prosody ...................................................................................................................................... 65

CHAPTER 4. METHOD ...................................................................................................... 67
Participants..................................................................................................................... 67
Syntactic constraints on data collection ....................................................................... 67
Sentence initial constituents ......................................................................................... 68
Paratone ......................................................................................................................... 68
Sentence-final constituents ......................................................................................... 69
Pausing ........................................................................................................................... 71
Single-syllable words ................................................................................................. 73
Trading relationships ................................................................................................. 73
Effect of adjacent stress on function words ............................................................... 74
Summary of constraints on data .................................................................................. 75
Data collection .............................................................................................................. 75
## LIST OF TABLES

Table 1: Common Lay and Acoustic Terminology ................................................................. 4
Table 2: Syllable-timing in English Phrases ........................................................................ 8
Table 3: Data from Liberman (2008) .................................................................................. 15
Table 4: PVI values for four varieties of English (Chen, 2015) ............................................. 16
Table 5: PVI values for six languages (Mok, 2009) ............................................................... 18
Table 6: Disambiguation of relationship among perceptual, acoustic, and phonological viewpoints towards prosody .................................................................................. 37
Table 7: Birch and Clifton’s (1995) Given-New Contract and Processing Difficulty .......... 47
Table 8: Eligibility of data .................................................................................................... 75
Table 9: Token type and total ............................................................................................... 76
Table 10: Frequency counts by OEPT level and examinee .................................................... 77
Table 11: Ordered Logistic Regression ............................................................................... 88
LIST OF FIGURES

Figure 1: Waveform of "cats" showing intensity and duration ................................................................. 5
Figure 2: Cats chase mice. (3 syllables) .................................................................................................. 9
Figure 3: The cats have chased the mice. (6 syllables) ......................................................................... 9
Figure 4: Comparison of syllable duration variation .............................................................................. 15
Figure 5: Mean PVI values of four varieties of English .......................................................................... 17
Figure 6: PVI values for six languages (Mok, 2009) ................................................................................ 17
Figure 7: The Apartment Building Effect ............................................................................................... 33
Figure 8: A conceptual schematic of prosodic levels .............................................................................. 41
Figure 9: A conceptual schematic of prosodic levels with acoustic cues .................................................. 57
Figure 10: Acoustic measurements for (2)...[BOXed]\text{\textphi} [in the ]CROWD]\text{\textphi} .................................. 63
Figure 11: Acoustic measurements for (3)...[BOXed]\text{\textphi} [IN]\text{\textphi} [the ]CROWD]\text{\textphi} .................... 64
Figure 12: Acoustic example of high paratone, which begins with "Let's skip..." ......................................... 69
Figure 13: The effect of pausing on F0 and duration .................................................................................. 72
Figure 14: Boxplots for Duration, F0, and Intensity .................................................................................. 86
Figure 15: Density plots for duration, F0, and intensity ......................................................................... 86
Figure 16: QQ-plots for duration, F0, and intensity ................................................................................. 87
Figure 17: Percentage of Acoustically Distinct Tokens that Meet or Exceed a 1:1.2 Ratio ...................... 90
Figure 18: Percentage of Acoustically Distinct Tokens that Meet or Exceed a 1:1.5 Ratio ...................... 91
Figure 19: Percentage of Acoustically Distinct Tokens that Meet or Exceed thresholds for F0 and Intensity .................................................................................................................................... 95
Figure 20: Visual representation of Gernsbacher's cataphoric "boldfacing" on structure building mechanisms ................................................................................................................................. 100
Figure 21: Simplified schematic of three stress levels ............................................................................ 101
Figure 22: Schematic of relationship of multiple systems in linguistics and cognitive psychology ............ 104
Figure 23: Schematic of collapsed columns based on lack of acoustic distinction between function and content words ............................................................................................................. 106
GLOSSARY

**Duration** in speech acoustics is exactly what it sounds like – a temporal measurement of discrete sound segments, words, or pauses, etc. Duration is typically measured in milliseconds, which produces positive integers rather than decimals that are often much shorter than 1 second. An example of this is the duration of an aspiration from the release of a \([t^h]\) until the onset of a following vowel, which is about 70 ms, as in \([t^h]ap\). The duration of an unstressed vowel might be around 50 ms, for example, while an adjacent stressed vowel might be 80 ms, or more. The perceptual analog to duration is length. There is an inexact relationship between duration and length in speech perception, as Turk and Sawusch (1996) demonstrated by digitally manipulating values of duration and intensity on vowels. Their results showed that the duration of a vowel affected the perception of intensity and intensity of a vowel affected the perception of duration. For example, while holding intensity values constant but varying durational values, participants reported changes in perceived loudness although intensity had, in fact, remained constant. As this study measures vowels, and because the term, “length,” in lay terminology refers to tense and lax vowels such as /i/ and /ɪ/, to avoid confusion, “length” is not used in this study.

**Intensity** is the physical pressure exerted by the perturbation of air. Undisturbed air molecules transmit a tiny amount of physical pressure, or intensity, on a sound recording device, which produces a faint white noise. When a person speaks, the vibration of the vocal folds and the air within the speech tract cause the physical air pressure to change as the sound waves propagate through the surrounding air. Sliding your hand across a tabletop produces low intensity levels as any vibration of air is localized at the point of contact between hand and table. Banging your hand on a tabletop produces high intensity levels because the entire table vibrates from the blow, which propagates more air molecules to vibrate. Changes in acoustic intensity are caused by anything that perturbs the air: the wind, a car engine, a whoopee cushion, a knock on a door, stereo speakers, etc.

**Loudness** is the perceptual analog to physical intensity. What is physically produced as intensity is perceived as loudness via the auditory system, which includes ears, eardrums (ossicles),
auditory nerves, and speech-sensitive areas of the brain. As mentioned above regarding duration and length, there is also an inexact relationship between intensity and loudness.

**Fundamental frequency (F0)** is the rate of vibration of vocal folds. Human listeners do not hear the F0 in vowels directly. We actually hear a concentration of frequencies that are higher than the F0, the source frequency. This is due to the sound filter that is our vocal tract. The source of speech sound waves is air that is pushed through the vocal folds, which vibrate and perturb the air. The vocal tract filters the vibration of the vocal folds (i.e., fundamental frequency) to produce thousands of frequencies, which include vowels, consonants and other speech sounds. We don’t hear the unfiltered frequency of the vocal folds because the vocal tract prevents listeners from directly hearing the sound source. In fact, vibrating vocal folds that are unfiltered by the vocal tract sound similar to a kazoo or a buzzing insect when one listens to direct recordings. We calculate the fundamental frequency (the greatest common denominator in sound wave frequencies) based on multiple frequencies that are present in the speech signal. We can contrast the measurement of F0 with the measurement of other sounds. Any noise – the clinking of glasses, footsteps on a wooden floor, a squeaking door hinge – produces sound waves, and we can measure the frequency, not the fundamental frequency, of those sound waves. When we hear ambient sounds such as the clinking of glasses, we are usually experiencing the sound source relatively unfiltered. That is, the sound waves are travelling directly from the sound source to our ears.

**Pitch** is auditory perception of fundamental frequency. Our auditory system detects that acoustic energy is concentrated around certain frequencies, for a vowel, for example. The speech perception area of the brain simultaneously perceives where acoustic energy is concentrated in the audio spectrum (thus producing vowel formants and a particular vowel), and also perceives the least common denominator of all of the frequencies present in the speech signal, which is the fundamental frequency. So the pitch (fundamental frequency) can rise and fall because of changes in rates of vibration of vocal folds while the areas of acoustic energy (vowel formants) remain stable because of the shape of the mouth and tongue. This phenomenon is evident in a surprised question like, “REALLY?” where two instances of [i] are produced with rising pitch where the first [i] would have a lower fundamental frequency than the second [i].
LIST OF ABBREVIATIONS

dB = decibels (a measurement of sound pressure created by sound waves)
ESL = English as a second language
F0 = Fundamental frequency
Hz = Hertz (measured by vibrations of vocal folds per second)
IGS = International graduate student
L1 = First language, native language
L2 = Second language, non-native language
ms = milliseconds
NES = Native English speaker
NNES = Non-native English speaker
OEPP = Oral English Proficiency Program
OEPT = Oral English Proficiency Test
OT = Optimality Theory
SLH = Strict Layer Hypothesis
LIST OF SYMBOLS

IP = Phonological intonational phrase
Φ = Phonological prosodic phrase
ω = Phonological prosodic word
LARGE CAPITAL TEXT denotes primary word stress
SMALL CAPITAL TEXT denotes secondary stress
lowercase text denotes no stress
ABSTRACT

Author: Johnson, Carl, Tyler. PhD.
Institution: Purdue University
Degree Received: December 2017
Major Professor: April Ginther

English prosody works as a structural and semantic glue that establishes relationships among words and phrases within a sentence, and among sentences within a larger discourse. This dissertation hypothesizes and demonstrates an association between acoustic measurements of English prosody and holistic measures of English proficiency. To test this hypothesis, acoustic data was used from 10 examinees each of low, medium, and high oral English proficiency groups of L1 Chinese speakers who took Purdue’s Oral English Proficiency Test (OEPT). Prosodic measurements of duration, F0, and intensity were gathered from adjacent function and content words in the OEPT audio data and compared with holistic OEPT scores. An ordered logistic regression found a significant difference \( p = 2.00 \times 10^{-16} \) among the three groups for how groups used durational differences between adjacent function and content words. Parallels of mental mapping of information are proposed between acoustic treatment of function and content words and the suppression and enhancement mechanisms of Gernsbacher’s (1997a) Structure Building Framework.
CHAPTER 1. INTRODUCTION

This study investigates the associations between holistic measures of oral proficiency of non-native English speakers, whose first-language (L1) is Chinese, and acoustic measurements of their non-native English prosody. English prosody is commonly exemplified by phonological features such as word stress, sentence stress, and intonation, and may be described with a number of acoustic cues, but prosody is most often acoustically described with measurements of pitch, loudness, and durational relationships among speech sounds\(^1\). By measuring acoustic cues of individual phonological prosodic features and combining those findings into a comprehensive picture, acoustic studies have progressed in the representation of the prosodic features of L1 English speakers and, to a lesser extent, prosodic features of non-native varieties of English. This study provides further evidence of the acoustic features of non-native English prosody of L1 Chinese speakers by measuring the acoustic cues of one phonological prosodic feature. This study investigates how overall English language proficiency of L1 Chinese speakers is associated with a prosodic phonological rule, which states that content words (e.g., nouns and verbs) are more prosodically prominent than adjacent function words (e.g., articles, prepositions, and auxiliary verbs) (Selkirk, 2004).

---

\(^1\) An accurate description of phonological, perceptual, and acoustic prosody or (suprasegmentals) will be given in Chapter 3. In the meantime, lay terminology such as stress and intonation (when speaking of phonological prosody) and pitch, loudness, and duration (when speaking of listener perception of prosody) are sufficient to introduce the context for this acoustic study of prosody among non-native English speakers.
Motivation for the study

Graduate programs across the US typically employ graduate students as teaching assistants (TAs). The starting point of this study came from my own work as a TA at Purdue University’s Oral English Proficiency Program (OEPP), which is a program that screens (through language proficiency testing) and helps current and future international teaching assistants (ITAs) improve their effectiveness in undergraduate classrooms. If the OEPP’s oral English proficiency screening determines that potential ITAs’ oral proficiency is too low to lead an undergraduate classroom without language support, the OEPP supports these graduate students to improve their academic communication skills, oral English skills, and cultural awareness of American university classrooms. The OEPP course curriculum consists of a classroom component that concentrates on the skills mentioned above, as well as once-weekly individual meetings with both the classroom instructor and a tutor, in which OEPP students may prepare for presentations, assess personal progress that has been made or is still required, and focus on individual goals and needs.

In these one-on-one sessions with students, some of my colleagues at the OEPP focused their efforts on pronunciation of discrete sound segments, perhaps because there is a widespread assumption that accurate pronunciation is a major factor in maximizing intelligibility\(^2\) (J. Sereno, Lammers, & Jongman, 2016), although, in fact, all segment types may not be equally important to promote intelligibility (Fogerty, Kewley-Port, & Humes, 2012). During this same time, I noticed that in my own one-on-one meetings, one sentence spoken with two different prosodic patterns resulted in different degrees of comprehensibility\(^2\). If the speaker focused attention on

---

\(^2\) I define the term *intelligibility* as ability to understand the individual words in an utterance while *comprehensibility* (Kang, 2010; Kang, Rubin, & Pickering, 2010) refers to ability to understand the combined meaning of the words of an utterance in context.
accuracy of pronunciation of segments but the sentence was delivered with relatively equal stresses on all or most words, for me that sentence was less comprehensible than the same sentence that had segmental mispronunciations but was delivered with expected patterns of phrasal stress. In my own experience, prosody seemed to be more important than accurate segmental pronunciation.

My anecdotal experience was partially confirmed by quantitative studies (Anderson-Hsieh, Johnson, & Koehler, 1992; Hahn, 2004; Kang, 2010), and these studies confirmed the influence of discourse prosody on comprehensibility at a discourse level and also global measurements of prosody on comprehensibility. I had already observed in the classroom that my OEPP students could use large prosodic gestures of discourse intonation that accompanied the intention of the speaker to emphasize a main point or to contrast two ideas, for example. But my one-on-one conference experiences drew my curiosity towards subtler prosodic relationships among words and led me to ask whether my anecdotal experiences of phrasal stress differences were empirically supported. This hunch led me to investigate whether higher oral English proficiency test scores are associated with expected patterns of phrasal stress (i.e., where content words are more prosodically prominent\(^3\) than adjacent function words), and whether lower oral English proficiency test scores are associated with relatively equal stresses on the words in a phrase (i.e., where content words are prosodically undistinguished from adjacent function words). This study poses the following research questions:

1. Do L1 Chinese speakers of non-native English acoustically distinguish content and function words with prosodic features?

\(^3\) Prosodic prominence, along with lexical stress, phrasal stress, sentence stress, and prosodic focus will be defined and discussed in chapter 3.
2. Which prosodic acoustic measurements distinguish content and function words in the speech of L1 Chinese speakers of non-native English?

3. How are acoustic measurements of prosodic prominence associated with measurements of English language proficiency of L1 Chinese speakers of non-native English?

**Acoustic terminology**

Pitch, loudness, and length are commonly used terms by both acoustic phoneticians and lay speakers. When a lay speaker describes someone as speaking fast, loudly, and in a high-pitched voice, that is a subjective description based on human perception of physical acoustics. The motivation for inquiry into speech prosody is to understand more about the complex phenomena of perception and production of speech prosody. Despite the sensitivity of human speech perception, acoustic phoneticians need an objectively measurable counterpart to those subjective perceptions. *Fundamental frequency (F0), intensity, and duration* are the objectively measurable counterparts to perceptions of *pitch, loudness, and length*, the relationships of which are shown in Table 1.

<table>
<thead>
<tr>
<th>Subjective perception:</th>
<th>Acoustic cues</th>
<th>Objective acoustic measurement:</th>
<th>Measured in:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Described in lay terms as:</td>
<td>Pitch</td>
<td>Loudness</td>
<td>Length</td>
</tr>
<tr>
<td>High to low</td>
<td>Intensity</td>
<td>Duration</td>
<td></td>
</tr>
<tr>
<td>Loud to soft</td>
<td>Hertz (Hz)</td>
<td>Decibels (dB)</td>
<td></td>
</tr>
<tr>
<td>Long to short</td>
<td>Milliseconds (ms)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1 shows a common visual representation of sound waves of speech. In this example, we see two acoustic aspects of the word, “cats”. Human speech creates sound waves, which create variations in air pressure, which listeners perceive as loudness\(^4\). Those variations in air pressure are measured in decibels as intensity. In Figure 1, we can see by the large vertical oscillations of air pressure that the segment, \([\text{æ}]\), is the loudest, or most intense, part of the word, followed by \([s]\), \([\text{kh}]\), and \([t]\). Dashed vertical lines that indicate segment boundaries and the relative length of the segments in “cats”. Again, the longest segment is \([\text{æ}]\), followed by \([s]\), \([\text{kh}]\), and \([t]\).

\[\begin{array}{c}
\text{Time (s)} \\
2.109 & 2.466 & -0.02139 & 0.01767
\end{array}\]

\(\begin{array}{c}
\text{Sound pressure variation} \\
-0.02139 & 0 & 0 & 0.01767
\end{array}\)

**Figure 1**: Waveform of "cats" showing intensity and duration

A visual representation of fundamental frequency is not shown in Figure 1 since the acoustic calculation of F0 is somewhat complicated. Regarding pitch, however, lay listeners\(^4\) The perception of loudness is affected by the frequency of the speech signal. For example, if two tones differ in frequency but have the same intensity, the higher-frequency signal will be perceived as louder.
generally understand that the pitch of one’s voice is a product of the vibration of a speaker’s vocal folds, that voice pitch rises and falls throughout a clause or discourse as intonation (e.g., to indicate questions or statements), and pitch variation co-occurs with the words in normal speech. As stated above, the acoustic counterpart to pitch is fundamental frequency, which is simply the rate of vibration of the vocal folds. So, if a speaker’s vocal folds are vibrating 75 cycles per second, the F0 is 75 Hz. For more explanation of the relationship between pitch and fundamental frequency, see the discussion on “Pitch” in the Glossary.

The purpose of making these objective versus subjective distinctions of acoustic cues is to minimize the effects of human subjectivity on the data for this study, which can present problems of validity and reliability for quantifying speech data. None of the above discussion is meant to suggest that objective measurement is superior to subjective evaluation: There are, of course, methodological situations where human subjectivity is preferred to objective measurement. Speech prosody is, after all, a perceptual linguistic phenomenon. However, the methodological context described in Chapter 4 will demonstrate why only objective measurements are appropriate for this study.

**Previous studies**

The results of this study are intended to be descriptive and to provide information that adds to the representation of the prosody of L1 Chinese non-native English speakers (NNESs). Two previous studies have looked at prosodic treatment of function and content words of L1 Chinese NNESs. One previous study (Baker et al., 2011) measured duration of function and content words in read-aloud text and another has looked at the intonational treatment of function and content words in extemporaneous academic speech. These studies are relevant because the
present study shares the prosodic measurement of function and content words of L1 Chinese NNESs but differs with the first in terms of the use of extemporaneous speech data rather than read-aloud data, and differs with the second in terms of targeting the effects of stress rather than intonation.

Baker, et al. (2011) studied temporal measurements of English speech segments from 12 native English speakers, 20 L1 Chinese NNESs, and 20 L1 Korean NNESs. The research team used the “Please call Stella” script (Weinberger, 2015) and Gina’s Pizza Shop script (Baker & Bradlow, 2009) to investigate whether durations of function and content words differed among the three groups. They found that while the durations of content words among all three groups was similar, the durations of function words between the native English speakers and NNESs was very different. Specifically, the team found that the native English speakers used greater durational difference between content and function words than the L1 Chinese NNESs and the L1 Korean NNESs. There was no difference, however, in the durational differences of content and function words between the two NNES groups. Baker et al. also had 50 native English-speaking raters judge severity of accent, from “native” to “foreign,” for all of the above 52 participants. They found that the best predictor of accent rating among NNESs was variance in word duration. That is, the more heterogeneous the participant’s word duration measurements were, the closer the accent judgment was to the “native”-sounding end of the accentedness scale, and the participants with more homogenous word duration measurements were more likely to be at the “foreign”-sounding end of the scale.

This second finding indicates whether a speaker’s language is more syllable-timed or stress-timed. Equal durational measurements of syllables in a phrase occurs in syllable-timed language such as Chinese or Italian, while variable durational measurements of syllables in a
phrase describe a stress-timed language such as English or Russian. [The rhythmic status of Korean is unclear (Arvaniti, 2009).] Another way to describe the nature of linguistic rhythm is by the total time that is required to utter a given number of syllables. If the time required depends on the number of syllables in the phrase, then the language is more syllable-timed. So, a syllable-timed language should require about half as long to utter a 3-syllable phrase compared to a 6-syllable phrase. But if the time required depends on the number of stressed elements in the phrase, then the language is more stress-timed (Celce-Murcia, Brinton, & Goodwin, 1996). The example text in Table 2, and segmentations of spoken waveforms in Figure 2 and Figure 3 demonstrate how in English, two similar phrases with three stressed elements take about the same amount of time to say regardless of whether the phrase has three (1.41 seconds) or six syllables (1.388 seconds). This rhythmic distinction will be instrumental in the choice of participants for this study.

Table 2: Syllable-timing in English Phrases

<table>
<thead>
<tr>
<th>Word class</th>
<th>Function</th>
<th>Content</th>
<th>Function</th>
<th>Content</th>
<th>Function</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 syllables:</td>
<td>CATS</td>
<td></td>
<td>CHASE</td>
<td></td>
<td></td>
<td>MICE</td>
</tr>
<tr>
<td>6 syllables:</td>
<td>The</td>
<td>CATS</td>
<td>have</td>
<td>CHASE</td>
<td>the</td>
<td>MICE</td>
</tr>
</tbody>
</table>
Figure 2: Cats chase mice. (3 syllables)

Figure 3: The cats have chased the mice. (6 syllables)
Although Chinese lies closer to the syllable-timing end of the continuum of these two categories of rhythmic timing, Chinese does have sensitivity to stressed and unstressed elements. Shi (1995) recorded and measured the speech of two English and two Mandarin Chinese speaking mother-child dyads, both speaking their respective native languages. One hundred function words (which carry no stress) and 100 content words (which carry lexical word stress) were randomly collected from each of the mothers’ speech, regardless of sentence position, yielding a total of about 400 words for each language. In the English measurements, duration, intensity and vowel quality showed significant differences between function and content words, while pitch was not significant. The measurements for the Mandarin mothers were very similar to the English results. Duration and intensity showed significant effects while pitch was insignificant. Shi found that both English and Mandarin distinguish content from function words by employing differences of duration and intensity. This study is interesting because although these two Mandarin speakers clearly distinguished function and content words in their native language, the 20 L1 Chinese NNESs in Baker et al. (2011) did not make similar distinctions between English function and content words.

A methodological concern of Shi’s (1995) study is that the researcher set no syntactic or discourse constraints on data collection, so words were collected from any position in a sentence. This means that data occurring at the beginning and end of a sentence were included in data collection for these studies. The physiology of the human vocal tract and the capacity of the lungs constrain the length of intonational phrases. Pitch values crosslinguistically typically start higher and gradually decline or down-step over the course of a sentence as the lungs empty of air (Pierrehumbert, 1980; Vaissière, 2005; Wennerstrom, 2001). “The evidence for baseline declination derives principally from the fact that unaccented syllables at the beginning of a
sentence are often at a higher level than unaccented syllables at the end of a sentence” (Cruttenden, 1997). Intonational declination means that a randomly selected function word from the beginning of a sentence may have a value equal to or higher than a randomly selected content word from the end of a sentence, which is of methodological relevance for data collection.

Wennerstrom (1998) measured the cohesive use of English intonation in academic lectures that were given by 18 native Mandarin Chinese speaking ITAs of a variety of English proficiency levels. In one of her four experiments, Wennerstrom measured pitch values on function and content words, but limited content words only to newly introduced information while she excluded given information content words from her data. Wennerstrom avoided the problem discussed in Shi’s (1995) study by eliminating function words that occurred at phrase-initial boundaries. She collected the first 20 pairs of function and content words from each participant, but it is unclear (in both the journal article and the dissertation that the article is based on) whether these function and content words occurred adjacent to one another. Knowing this information is important because the “pitch [of function words] is interpolated from surrounding tones” (Wennerstrom, 1998, p. 10). Wennerstrom found that all of her participants used higher intonation (about 40-60% higher) on new information content words than on function words, although there was no difference among participants of different English proficiency. As Wennerstrom was specifically investigating the effects of new information on intonation, it was appropriate to collect data from the end of the sentence, as the unmarked location for new information is at the end of a sentence (Halliday & Matthiessen, 2013).

The present study, however, is investigating somewhat the opposite of Wennerstrom’s (1998) function-content research question. The present study attempts to eliminate variables that are known to affect acoustic prosodic measurements, such as sentence stress or given/new
information status. In the attempt to measure only the effects of one specified prosodic phonological stress rule, which states that content words are more prosodically prominent than adjacent function words (Selkirk, 2004), this study will strictly constrain data collection. A complete description of the constraints to data collection will be given in Chapter 4, but with regard to the previous studies that have measured prosodic treatment of function and content words of L1 Chinese NNESs, the present study limits discourse structure effects by setting specific syntactic constraints that eliminate the likelihood of new information content words from entering the data, and it removes physiological phrase boundary effects that were identified in Shi (1995).

Outline of the study

Chapter 2 gives an overview of the ITA participants of the study, provides a historical context for ITAs in the United States, and describes the circumstances that created the proficiency test, which the study draws its data from. Chapter 2 ends with a review of acoustic studies of ITAs’ oral English prosody and overall proficiency. Chapter 3 discusses definitions of prosody, and provides a theoretical framework that motivates the selection of content and function words as a focus for prosodic measurement. Chapter 4 describes the nature of the prosodic data that is collected, describes the restrictions to data collection, and describes the restrictions to acoustic measurements. Chapter 5 provides the results of the study, an interpretation the quantitative results, and connections from this study to future research.
CHAPTER 2. OVERVIEW OF PARTICIPANTS

International graduate students (IGSs) currently account for 35% of graduate students in the United States (Okahana & Allum, 2015). At Purdue University, where the data for this study originates, the IGS population constituted substantially more at 40.3% in 2015 (“International students and scholars enrollment & statistical report fall 2015,” 2015). Many IGSs in the United States are ITAs, and therefore teach undergraduate courses, lead laboratory class components, tutor undergraduates, assist with research, or perform similar duties during their graduate study; that is, ITAs regularly interact with undergraduates, most of whom are native English speakers. For all of these duties, the ability to communicate clearly and efficiently is essential.

The participants for this study are non-native English speaking prospective ITAs who identify their L1 as Mandarin Chinese. This self-reported data is only somewhat reliable as there is a cultural tradition of conflating all Sino-Tibetan languages within the political boundaries of China as “dialects” of Mandarin Chinese although these dialects are often not mutually intelligible (Handel, 2015). Conflating dialects of one language with different languages becomes more logical when one learns that “all Chinese dialects share the same written language and essentially the same grammar” (Duanmu, 2000, p. 2). So, the linguistic reality of the L1s of the participants in this study may include languages that are not verbally mutually intelligible with Mandarin. Therefore, this study assumes that all of its participants are native speakers of at least one language in the Sino-Tibetan language family, which as a typological group, contrast prosodically with English. For the sake of simplicity, the participants in this study will be referred to as Chinese speakers, in keeping with the participants’ self-identified language group.
Rhythmic classes of languages

This group of participants has been selected because 1) Chinese speakers make up the largest linguistic demographic within the IGS community at both Purdue and the United States, and 2) Mandarin Chinese has been described as a syllable-timed language, which contrasts typologically with English prosodic rhythm (H. C. Chen, 2015; Mok, 2009). It should be intuitive to anyone with foreign language experience that languages differ rhythmically from one another. Very generally, according to the Rhythm Class Hypothesis (Grabe & Low, 2002), syllable-timed languages (e.g., Chinese [and its cultural dialects], Spanish, and Cantonese) prefer prosodic uniformity across syllables in a phrase while stress-timed languages (e.g., English, Dutch, Farsi and Arabic) prefer prosodic contrast across syllables in a phrase. Intuition and perception are valid indicators of prosodic rhythm, but exactly how prosodic rhythms differ acoustically among languages has remained an open question since the Rhythm Class Hypothesis was proposed (Galaczi, Post, Li, Barker, & Schmidt, 2017). While perceptual judgments categorically separate languages and dialects into syllable-timed and stress-timed prosodic groups (Fuchs, 2016), clear acoustic boundaries that delineate those groups have been inconclusive (Rathcke & Smith, 2015). For that reason, a continuum that includes four prosodic rhythmic tendencies (syllable-timed, stress-timed, rhythmically mixed, and mora-timed) has been adapted from Grabe and Low’s original work (Fuchs, 2016; Li & Post, 2014).

Taking the above caveats of rhythmic categorization into consideration, Figure 4 presents a simple acoustic example of the rhythmic contrast between an archetypical stress-timed language (English) and an archetypical syllable-timed language (Spanish) (Liberman, 2008). In a blog post meant for illustrative purposes, Liberman measured the duration of the first 120 syllables (following the self-introduction) of a news podcast spoken by the same bilingual
newscaster who presents in English and Spanish. The syllable duration plot of along the y-axis of English is visually more variable than that of Spanish. The English data visually ranges from 100 ms to 500 ms while the Spanish data ranges from 100 ms to 300 ms.

![Figure 4: Comparison of syllable duration variation for a bilingual individual speaking English and Spanish (Liberman, 2008)](image)

**Table 3: Data from Liberman (2008)**

<table>
<thead>
<tr>
<th>Language</th>
<th>Mean syllable duration</th>
<th>Standard deviation</th>
<th>Percentage of standard deviation to Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>204 ms</td>
<td>107 ms</td>
<td>52%</td>
</tr>
<tr>
<td>Spanish</td>
<td>189 ms</td>
<td>59 ms</td>
<td>31%</td>
</tr>
</tbody>
</table>

The data of Liberman’s experiment show in Table 3 that while the mean duration of syllables was similar between the two languages (204 ms in English versus 189 ms in Spanish),
the standard deviations were very different (107 ms and 59 ms), and thus, the percentage that the standard deviation represents to the syllable mean was also very different between the two languages (52% and 31%). These two percentages represent one type of measure of durational variability among all of the syllables in the speech samples, and show one measure of rhythmic contrast between the two languages.

An commonly used alternative measure of speech rhythm is the Pairwise Variability Index, developed by Grabe and Low (2002), allows for plotting rhythmic classes of languages along a continuum, that is, more syllable-timed, more stress-timed, or somewhere in between. Larger PVI values correlate with human perception of stress-timing and smaller PVI values correlate with human perception of syllable-timing. Chen (2015) tested the PVI values for four varieties of English. She asked a total of 40 participants to read a script aloud: 30 were speakers of L2 English from three different locales and L1 Chinese dialects: Hong Kong (Cantonese), Beijing (Mandarin Chinese), and Taiwan (Taiwanese Chinese), as well as 10 L1 American English speakers from California. After gathering data of stressed and unstressed syllables from all four groups, Chen found a mean PVI value of 52.3 for the L1 English speakers while the means of the PVI values were in the 30s for the participants with an L1 of one of the Chinese dialect groups (see Table 4). Chen’s PVI values are plotted in Figure 5.

<table>
<thead>
<tr>
<th>L1 of Participants</th>
<th>Mandarin</th>
<th>Taiwanese</th>
<th>Cantonese</th>
<th>Native English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean PVI</td>
<td>32.64</td>
<td>37.55</td>
<td>39.76</td>
<td>52.3</td>
</tr>
</tbody>
</table>
Mok (2009) used a read aloud script translated into different languages to compare the
rhythmic timing of Cantonese and Mandarin with four European languages. The PVI values from
her study are plotted in Figure 6: PVI values for six languages and the data is presented in Table 5.
Table 5: PVI values for six languages (Mok, 2009)

<table>
<thead>
<tr>
<th>Language (Number of participants)</th>
<th>Hong Kong Cantonese (6)</th>
<th>Beijing Mandarin (6)</th>
<th>French (6)</th>
<th>Italian (3)</th>
<th>German (15)</th>
<th>British English (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized PVI</td>
<td>34.32</td>
<td>45.02</td>
<td>49.47</td>
<td>54.78</td>
<td>56.42</td>
<td>69.67</td>
</tr>
</tbody>
</table>

The intention of this discussion is not to stake claims about rhythm class categories, but rather to demonstrate the prosodic tendencies of syllable-timed Sino-Tibetan languages, which contrast prosodically (both perceptually and acoustically) from stress-timed languages like English (Grabe & Low, 2002; Nazzi, Bertoncini, & Mehler, 1998; Nazzi & Ramus, 2003; Ramus, Nespor, & Mehler, 1999). As the present study investigates NNESs’ prosodic production, this study enrols participants who must acquire a contrasting prosodic rhythm in their second language, English.

The Context of the Data

The prosodic data measured and analyzed in this study comes from audio recordings from Purdue’s Oral English Proficiency Test (OEPT). Purdue’s OEPT is administered to prospective non-native English speaking ITAs with the objective of ensuring that ITAs possess the oral English proficiency required to successfully carry out their job. Language proficiency testing was not always the policy at Purdue, which resulted in undergraduate classes that were taught by ITAs who lacked the requisite language skills. This lack of English language proficiency among some ITAs was the reason that Purdue and many other U.S. universities and
colleges found themselves in the midst of a situation that came to be known as “the foreign TA problem”.

“The Foreign TA Problem” in the United States

During the 1980s, a conflict developed at U.S. universities among international teaching assistants, American undergraduates and their parents, and the administrators and faculty members of universities. The conflict centered on a mismatch of communicative, cultural, and educational expectations of the above three parties (Bailey, Pialorsi, & Zukowski/Faust, 1984). Prior to the emergence of this conflict, both domestic and international graduate students had been employed at universities, but international graduate students (i.e., NNESs) were usually employed in research positions that limited their contact with university undergraduates, while domestic native English speaking graduates were normally employed as TAs or do other work that required regular contact with undergraduates.

Contemporary newspaper accounts attributed the conditions for the conflict to emerge to structural and demographic changes in the academic and professional science communities that dissuaded domestic graduates from continuing into graduate studies and instead led them toward lucrative professional careers in fields such as science, math and engineering (Fiske, 1985; Kelley, 1982). Faced with a TA void to fill, universities turned to ITAs to do the work that had been performed by native English speaking graduate students: teaching undergraduate classes, leading labs and recitations, and fulfilling other responsibilities that put ITAs in the position of educating undergraduates and answering student questions.

The conflict emerged because, prior to language proficiency testing, many ITAs did not have adequate oral English proficiency to succeed in these teaching positions. A professor of
mathematics at M.I.T. described the decision-making problem that graduate programs faced when choosing graduate candidates:

The result is a dilemma for graduate faculties: do they take the candidates with the best credentials or opt for less qualified Americans? For many institutions, the choice is simple. Research comes first. Then we look for some evidence that [ITAs are] not hopeless in the classroom (Fiske, 1985).

Doctoral universities that are recognized by the Carnegie Classification as having the Highest Research Activity, which includes schools like M.I.T. and Purdue, are especially motivated to accept candidates who are likely to be the most valuable in support of research. Administrators also have long-term motives of increasing the schools’ global profile through published research, which in turn ensures maintenance or increase in tuition revenue from international students. In addition to research and fiscal motives, Purdue’s mission includes the promotion of diversity of people, culture, and ideas (Ginther & Allen, 2015).

The combination of R1 universities’ missions along with a new role for IGSs led to exasperation and disparaging complaints about ITAs’ English communication skills at R1 schools, which appeared in university newspapers, letters to university administrators and government officials, and nationally circulated newspapers (Bailey, 1983b, 1984; Fiske, 1985; Kelley, 1982; Shaw, 1982; Swanbeck, 1981; Timmerman, 1981). While ensuring adequate oral English proficiency of ITAs is a valid requirement for appointment to a teaching position, a tone of ethnocentrism was evident in these complaints (Bailey, 1983b). The shared negative tone in articles that ascribed blame for the conflict to ITAs is exemplified by an inset cartoon that appeared alongside one university newspaper article. The cartoon shows a professor’s lectern at which a space alien with flailing tentacles, three eyes, and mouth located in its abdomen, says,
“EEK BUP ONX ELP…” to which a university student comments to his classmate, “It’s all alien to me” (Krause, 1981; Timmerman, 1981). Language education specialists (e.g., teachers and scholars) voiced their disapproval of how public sentiment focused blame on ITAs while ignoring the complexity of the problem, which also included accountability from universities, but the same language education specialists acknowledged the legitimate problem of communicative, cultural, and educational mismatches (Bailey, 1983a; Fisher, 1985).

**Purdue’s response to the conflict**

The same mismatches of expectations and classroom realities that were occurring at the national level were also occurring at Purdue, which,

“[were] exposed in the late 1970s through loud complaints by American undergraduate students and their parents about the undergraduates' inability to understand the spoken English of their ITAs. Their charge, which university administrations had to address, was that ITAs' poor communication skills were having a negative affect on undergraduates' academic progress” (Cassell, 2007, p. 12).

In 1987, Purdue’s University Senate approved a remedy to the situation by establishing (through Purdue’s Graduate School) an Oral English Proficiency Program, which was given two primary duties. The University Senate’s mandate introduced language testing for prospective ITAs to demonstrate English proficiency before receiving a teaching assignment. Results from the proficiency test were reported back to examinees’ respective departments of study to aid departmental administrators in making the decision to appoint an ITA to a teaching position. Those prospective ITAs who scored below a certain threshold on the oral English proficiency
test could be required by their departments to enroll in the new English language program, which was the second part of the OEPP’s mandate, before being given a teaching assignment. Dr. Margie Berns, the founding director of the OEPP, identified the goals of the ITA-focused course as, “cultivating awareness of verbal and non-verbal behavior, concentrates on pronunciation and acquisition of technical and non-technical vocabulary relating to courses, presentation and interaction, and focusing on cultural differences in the American classroom” (Cassell, 2007, pp. 31–32).

The current goals of the OEPP ITA development course, *Classroom Communication for International Graduate Assistants*, include those original goals as well as an emphasis on TA-student interaction through question and answer procedures and elaboration on key concepts (College of Liberal Arts, 2016). The OEPP course is now a 5-credit hour, 600-level graduate course. OEPP classes are capped at 8 students to maximize both teacher-student discussion and student-student interaction. Classes meet twice per week for a total of 4 hours per week. To ensure that class goals align with individual student goals and abilities, the OEPP administration attempts to create homogenous proficiency level classes by placing students of similar English proficiency levels together. These proficiency levels are described in detail on the OEPT scale (see Appendix). In addition to class meetings, each individual OEPP student meets weekly with the course instructor for 30 minutes and an OEPP tutor for 50 minutes. Individual meetings are designed to identify personal goals, work towards those goals, and assess whether OEPP students are making progress. Instructors and tutors are typically IGSs from colleges and programs related to language and teaching, such as Education; Speech, Language, and Hearing Sciences; English Language & Linguistics; and Second Language Studies/ESL.
OEPP students are primarily assessed via four formal presentations and three ICs, which stand for “interactive conversation” or “informal conversation”. Both of these assessment instruments include a self-assessment that oftentimes references points of progress that were identified in individual meetings with the tutor and instructor. IC topics of conversation are general and typical of small talk among international residents (their opinion of university housing accommodations, midterm/finals week preparation, grocery shopping in the US, etc.). IC topics are presented to pairs of students, and are intended to be a conversation starter from which the conversation is expected to proceed anywhere the students take it. ICs usually last about 15 minutes.

The 4 academic presentations have a time limit of 15 minutes plus 5 minutes for question and answer. The audience for presentations is volunteer undergraduates from a wide variety of academic areas. The presentations include the goals: 1. To allow OEPP Ss to practice explaining academic concepts in accessible, everyday language to a general audience, and 2. To allow ITAs in training a chance to practice interacting with Purdue undergraduates during question and answer procedures.

There is an emphasis in OEPP classroom instruction on explaining academic concepts in non-complex language, using easily understandable examples, summarizing, paraphrasing, and rephrasing, fielding questions from the audience, audience-generated question and answer procedures, teacher-initiated comprehension questions, and how expectations from an American undergraduate classroom may be different from the expectations of the ITA’s native academic culture.

From the perspective of mostly East Asian IGSs, American undergraduates are more likely to treat Q&A as a conversational exchange, and undergraduates are more likely to ask
questions during a lecture. Likewise, university instructors are more likely to expect and encourage this type of interaction. Regarding these topics about American classroom culture, the intimate OEPP class size encourages discussion and advice to come from both teacher and ITA colleagues-in-training alike.

During formal presentations, non-major undergraduate questions are especially indicative of the success of the presentation. Sometimes questions are specific and based on information from the presentation, and are requests for further elaboration or clarification. Requests for specific information would be an indication of a successful presentation on the academic topic. However, sometimes an undergraduate question can be paraphrased simply as, “I don’t understand,” which may indicate that the topic was not accessible to a general audience. This type of student question is also vague, which should elicit probing questions from the ITA in training to discover exactly what the undergraduate does not understand, and to what extent that can be remedied during the Q&A. In addition, in the self-assessment component of the presentation assignment, OEPP ITAs in training are encouraged to think about how they might have been more effective in the situation above.

After the creation of the OEPP in 1987, the Educational Testing Service’s Speaking Proficiency Assessment Kit (ETS SPEAK) was initially used as a proficiency test for potential ITAs at Purdue. ETS published research results that showed the SPEAK test to be a good predictor of overall language proficiency, but not a good predictor of teaching ability (Sarwark, Smith, MacCallum, & Cascallar, 1994). Hoekje and Linnell (1994) did not dispute ETS’s claims of the predictive power that SPEAK has for general English language use, but they did warn users of the SPEAK test that, “…it is no longer acceptable to have language tests which are statistically valid and reliable but inauthentic in their tasks…” (Hoekje & Linnell, 1994, p. 122).
Hoekje and Linnell were specific in their critique of using a test of general oral English vis-à-vis a specialized group of examinees; “In our evaluation, the SPEAK test is lacking as an assessment measure for ITA performance because it contains few opportunities to elicit the sort of discourse competence that has been shown to be a crucial part of comprehensibility in instructional language” (Hoekje & Linnell, 1994, p. 121).

**Purdue’s OEPT**

It was because of this decontextualized general language that the SPEAK test used, as well as the difficulty of administering the SPEAK test, that in 2001, the OEPP completed its own oral English proficiency test, “…that is specific to the context at Purdue University [which] provides the basis for a more valid measurement of ITA English proficiency required for teaching at the university…” (R. Yang, 2010). There have been two versions of the OEPT, titled OEPT1 and OEPT2, which differ in rating scales, rating descriptors, and several test items (Ginther, Redden, Mishima, Cheng, & Thirakunkovit, 2013). This study uses data from the OEPT2, which will henceforth be referred to simply as the OEPT.

“The OEPT is a [12-item] computer-based test used by the Oral English Proficiency Program to screen prospective teaching assistants for language proficiency” (“Oral English proficiency test,” 2012). The content of the test is situated in interactions that a graduate teaching assistant is likely to have in the context of an American University. Ten items prompt examinees to respond extemporaneously while the other two items require examinees to read a printed text aloud. Some of the extemporaneous items require a response to a written prompt, such as giving a reaction to a campus newspaper headline or giving advice to an undergraduate student who is having a problem with one of his/her classes. Other items require a response to an audio prompt,
such as relaying a voicemail message to a colleague at work or summarizing a conversation between a professor and a student. Each test item has written instructions that are also read by a narrator’s voice. Examinees are offered up to 2 minutes of preparation time, but they may start recording their response as soon as they are ready to do so. Likewise, if examinees do not click to proceed to the recording, the recording starts automatically when the preparation time limit is reached. Recorded oral responses have a 2-minute time limit, except summarizing a short academic lecture, which has a 3-minute limit. If the examinee does not stop the recording, the recording stops automatically when the time limit is reached.

After the test administration, two trained human raters assign individual scores to test items as well as an overall holistic test score based on the item-level scores. English language proficiency, which is determined by the holistic test score, is rated on a 6-point scale that ranges from 35 to 60 in 5-point increments, which can briefly be described as: 35 (Restricted performance), 40 (Not ready for the classroom), 45 (Borderline performance), 50 (Adequate performance), 55 (Very good performance), 60 (Excellent performance). A detailed description of the performance levels can be found in the Appendix. Examinees pass or fail “certification,” or sufficiently proficient to work in a teaching environment with Purdue undergraduates, based on this 6-point scale. Examinees who score 50, 55, or 60 are certified to serve as ITAs. Examinees who score 45 may serve as a TA but they must concurrently enroll in the OEPP course. Examinees who score 35 and 40 on the OEPT must enroll in the OEPP and be certified before receiving an appointment as a TA. Should the two raters’ OEPT scores disagree, and the disagreement spans the cut-off between passing and failing certification, a third rater is assigned to break the tie (Ginther et al., 2013). OEPT test scores and classroom certification results are reported to the OEPT examinee’s or OEPP student’s graduate school to serve as a guide when
the graduate school decides on what type of TA appointment or whether to appoint the IGS to a position that would entail interaction with Purdue undergraduates. Students who enroll in the OEPP must be certified before they are given a TA assignment. The OEPP director makes the ultimate decision to certify an OEPP student, which is based on recommendations from the OEPP student’s classroom instructor and tutor.

OEPT raters are trained systematically during the academic year to ensure reliability. The OEPT rating rubric (see Appendix) shows that raters are expected to consider many factors in their evaluations of overall English proficiency: intelligibility, coherence, comprehensibility, correctness of grammar, appropriate complexity of syntax, appropriateness and breadth of vocabulary usage, segmental pronunciation, fluency measures such as unexpected pausing and speech rate, and prosody. Among all of these contributors to English proficiency that raters may attend to, OEPT raters often describe examinees’ prosody as monotonic or staccato (Ginther, 2016). And this description of prosody is not limited to test evaluations from OEPT raters, as this description in Wennerstrom (2000) exemplifies:

To summarize the characteristics of the low-fluency speakers, we’ve seen several examples of the tendency to give relatively equal pitch to each word regardless of its role in the information structure of the discourse. This can occur as many sequential high-pitched words or as a flat monotonous string of words, creating a choppy, word-by-word effect (Wennerstrom, 2000, p. 118).

It is unsurprising that raters attend to prosody because it plays a role in speech perception and comprehension. We will see in Chapter 3 that a listener’s general impression of prosody is important for comprehension of the meaning of an utterance because prosody provides meaningful structure to the speaker’s message as words and phrases are produced. Prosodic
structure lays out an acoustic map for the listener that aligns with syntactic structure and points to key information. One can imagine though, that if prosody is staccato, then the acoustic map draws attention to every word or syllable in a sentence, and thus, highlights none of the words. Or if prosody is monotonic, then likewise, the acoustic map doesn’t draw attention to any word in particular. However, expected English prosody would draw listener attention to things like nouns and verbs at the sentence level, and new information and contrasting information at the discourse level. Using prosody to signal phrase structure and discourse structure is especially important in a public setting such as a university lecture, where the speaker does not get immediate spontaneous feedback for unclear or ambiguous speech, as one might get in a conversation.

**Prosody of ITAs**

**ITAs and sentence stress**

Hahn (2004) provides an example of the effects of inappropriate prosody from an ITA on university undergraduate comprehension. Hahn studied the comprehension effects that were produced by different configurations of sentence stress on new and given information. Sentence stress is acoustically signaled with longer durational values and elevated pitch values (Calhoun, 2010). Raised acoustic values perceptually signal new information and contrastive information.

Hahn recruited one highly English proficient former ITA with an L1 of Korean to compose three versions of a lecture that differed only in application of sentence stress. The first version had expected sentence stress on new information. The second version had inappropriate sentence stress on given information. The third version was delivered with relatively flat stress that had no sentence stresses.
Ninety native English speaking undergraduate participants in Hahn’s study listened to the lectures and responded to two questionnaires that assessed listener comprehension by tallying main ideas and details remembered from the lecture. The subjects that listened to the first version with appropriate sentence stress recalled significantly more information than the subjects that listened to the second or third versions with inappropriate or flat stress, respectively. The same pattern emerged for both measures of comprehension: recollection of main ideas, and recollection of details. Version 1 (appropriate stress) scored the highest, version 2 (inappropriate stress) scored the lowest, and version 3 (flat stress) scored slightly higher than group 2. Moreover, 9 of 30 listeners of the third stressless version said that the lecturer spoke too fast, although the overall time and pauses between words and sentences was strictly controlled to be equal with sound editing software. The study concludes that appropriate stress not only allow listeners to comprehend and retain information, but also appropriate stresses allow listeners to comprehend oral text quickly and efficiently. Hahn argues that classrooms that specialize in instruction to international graduate students should include explicit instruction on prosody.

**ITAs and intonation**

In addition to prosodic treatment of function and content words that was described in Chapter 1, Wennerstrom (1998) also tested the hypothesis that intonation was instrumental to discourse structure. She recruited 18 L1 Mandarin Chinese NNES ITAs whose general oral English proficiency had been evaluated with ETS’s SPEAK test and whose scores had spanned the evaluation scale. Wennerstrom measured the cohesive use of English intonation in academic lectures given by the ITA participants. She assumed that English intonation was a system of rule-
governed, intonational cues that indicates relationships among syntactic constituents and informational units.

The intonational environments that Wennerstrom investigated were word class (function vs. content words, discussed above), phrase boundaries, contrastive information vs. given information, and paratone. The phrase boundary that Wennerstrom studied was a high-rising, phrase-final intonational unit that signals, 1. the speaker’s intention to continue and, 2. that the information that comes before and after the tone to be thematically connected. Contrastive information typically carries steeply rising, high intonation on syntactically parallel constituents (e.g., two noun phrases). Given information typically carries a low pitch accent. In a public speaking context, paratone includes elevated intonational range, which is beyond a speaker’s average pitch range, connects to the larger discourse, and spans multiple words of a new topic or redirection to a previous topic. An example of a phrase that would carry paratone in the context of an academic lecture is, “And now I want to turn our attention to NEW TOPIC.”

A multiple regression analysis found that the participants’ use of paratone was predictive of the participants’ SPEAK proficiency scores. The more often the ITA used paratone, the higher the ITA’s SPEAK proficiency score. There was no effect for phrase boundaries or contrastive information vs. given information. These findings are consistent with Calhoun’s (2010) findings, which will be discussed in Chapter 4.

Although the participants in the following study were not ITAs, Wennerstrom (Wennerstrom, 2000) conducted a study with participants in an academic environment, which disagreed with her earlier results (Wennerstrom, 1998). She compared English fluency test scores of 10 NNESs, who were enrolled in an intensive English as a second language program, to acoustic measurements of their intonational production. The oral fluency test was very similar to
the ETS SPEAK test, and the two raters who participated in the study were trained SPEAK raters. The oral data for the discourse analysis came from an informal one-on-one conversation conducted and recorded by the participants themselves as part of an intensive ESL assignment. A discourse analysis was carried out on the participants’ recorded speech to identify: 1. High information status, new and contrastive information, 2. Anaphoric, given or contextually retrievable information, and 3. Low information function words. Finally, Wennerstrom collected peak F0 values for the strongest syllable in all of these words. Wennerstrom found that the three lowest English proficiency participants, whose L1s were Korean, Thai, and Japanese, made very little intonational distinction between content and function words, as well as little intonational distinction between new or contrastive information and given information. The high fluency speakers (1 Italian, 3 Koreans, 2 Japanese, and 1 Chinese) used F0 values that clearly distinguished new and contrastive information from given information, and F0 values that clearly distinguished function and content words.

**ITAs and proficiency test criteria**

Plough, Briggs, and Van Bonn (2010) quantitatively and qualitatively analyzed recordings of 44 potential ITAs (23 of whom were L1 Chinese speakers) who took an oral English proficiency test. The proficiency test was used to determine whether potential ITAs have language skills sufficient to receive TA assignments at an American university. The interactive test included four parts: 1. an evaluator conducts a general interview with the examinee; 2. The examinee gives a lesson presentation that the examinee has chosen and prepared in advance of the test; 3. a role play in which an evaluator takes on the role of an undergraduate student with a problem and visits the TA (examinee) at his/her office hours; and 4. a listening comprehension
task in which the examinee answers questions based on viewing a video. The researchers both quantitatively measured and qualitatively evaluated interactional/social competence (in parts 1 and 3), transactional/expository competence (in part 2), listening comprehension (in part 1, 3, and 4), vocabulary, grammar, and pronunciation, “which includes fluency and intelligibility (e.g. pausing/hesitation, prosody, articulation, and voice projection)” p. 240. The qualitative and quantitative results agreed that listening comprehension and pronunciation were predictive in determining whether examinees were approved to receive TA assignments. Prosody measures are conflated with pronunciation and fluency, and the authors do not attempt to rank importance among them. This conflation, however, may be a minor drawback when considered in light of a study by Anderson-Hsieh et al. (1992) that demonstrated a strong correlation (r = 0.90) between human judgment of pronunciation of segments and human judgments of prosody. The Plough et al. (2010) study does offer some evidence of the primacy that pronunciation and its components have over other oral proficiency factors of grammar, vocabulary, interactional/social competence, and transactional/expository competence in an academic context.

Studies of ITAs’ prosody have tended to focus on higher levels of discourse structure, which is logical considering the speech acts that are typical of classrooms. But as we will see in the next chapter in a discussion of the role of prosody in comprehending speech, prosody impacts language processing at all levels of discourse, from paragraph level utterances right down to word recognition.
CHAPTER 3. AN INFORMAL DEFINITION OF PROSODY

Until now, I’ve relied on lay terms to describe prosody, which can seem very familiar, as when discussing intonation and word stress. But prosody is multidimensional and complex (it is adaptive to its context), and therefore, it is in fact, very complicated. So I’ll start this definition of prosody with a simple analogy.

We can informally describe prosody with a noisy apartment building (Figure 7). The sound of neighbors’ voices travels through the walls, but not all of the sound frequencies do so. The walls of the apartment act as a filter that block the high frequency sounds in consonants and sonorant formants, but the walls do transmit lower frequency components of vowels.

Figure 7: The Apartment Building Effect

In fact, perceptual studies (Cunningham, 2012; Derwing & Munro, 1997; Nazzi et al., 1998; Nazzi & Ramus, 2003) have attempted to isolate prosody from the influence of other
speech sounds by using a low pass frequency filter similar to this apartment building analogy. The listener receives information such as low frequency pitch variations (which includes intonational contours that may convey speaker intent, attitude and emotion), number of pitch peaks per phrase (which may indicate whether a language is tonal or intonational), location of pitch peaks within a phrase (which indicates whether the language is left or right branching), and relative durational differences between adjacent vowels (which describe the rhythmic class of the language). However, that unintelligible filtered speech from the next apartment is only an analogy, so a more detailed definition of prosody is called for.

**Definitions of prosody**

Prosody is a difficult phenomenon to define partly because a working definition depends on the approach taken to examine prosody. If prosody is defined from a perceptual point of view, then prosody is a crucial component of the speech signal beyond segmental information that promotes word recognition, parses phrases, and co-occurs with syntax and discourse. But if prosody is defined from an acoustic point of view, then prosody is a gestalt of acoustic cues such as pitch contours, vowel duration, vowel intensity and vowel quality that are integrated as part of the speech signal that simultaneously work together to produce multiple perceptual purposes. Yet a third point of view is an abstracted phonological representation of prosody that attempts to describe systematic meaning that connects the perceptual and physical dimensions of prosody. Finally, all three of these descriptions of prosody interact with the framework of syntactic structure. These multiple viewpoints result in one phenomenon (e.g., lexical word stress) having multiple corollaries (prosodic word and lexical word) from the perspective of other viewpoints.

Table 6 offers the example sentence, *The sluggers boxed in the crowd* (Selkirk, 1981), to provide some disambiguation to these related but distinct viewpoints of prosody. The four
approaches displayed in the table are perceptual, acoustic, phonological and syntactic. Perceptual elements that listeners attend to such as lexical word stress, phrasal stress, and intonational pitch contours are presented in the first section of Table 6 labeled, “1. Perception of prosody” LARGE CAPITAL TEXT denotes stressed syllables, and lowercase text denotes no stress. Stylized curved lines above the sentence denote intonational pitch contours belonging to the intonational phrase. Lexical stress applies to lexical words such as nouns, verbs, and adjectives while function words such as articles, pronouns, and prepositions receive no stress.

Next we move from perception to the physical dimensions of prosody. Physically measureable acoustic cues that speakers produce such as fundamental frequency (the physical measurement of pitch), vowel duration (measurement of segmental length), and intensity (measurement of loudness) are provided in the section of Table 6 labeled, “2. Acoustic production of prosody”. The blue and green lines represent measurements of recorded speech, fundamental frequency and intensity, respectively. Vertical lines indicate boundaries between consonants, consonant clusters and vowels. Fundamental frequency (F0) is the rate of vibrations of the vocal folds per second, and is measured in Hz. Duration is the time measurement of a vowel, and is measured in milliseconds. Intensity is a measurement of air pressure created by the sound of the voice, and is measured in decibels. Distinguishing between production and perception of the same acoustic cues is important because there remain questions as to which acoustic cues or combinations of cues trigger which listener perceptions, and how these cues interact with information structure. These distinctions will be discussed further in this chapter.

Drawing on information from perceptual and acoustic prosody, prosodic phonology attempts to construct a mental model of meaningful speech sounds. A phonological representation of prosody arranges prosodic elements within a hierarchy (Nespor & Vogel, 1986;
Selkirk, 1981), which includes the prosodic word, prosodic phrase, and is dominated in this example sentence by the intonational phrase (IP). Starting from the bottom of this hierarchy\(^5\), prosodic words (which align with lexical words) are contained within prosodic phrases (which group words into meaningful units, and typically align with syntactic phrasal units such as noun, verb, and prepositional phrases), and prosodic phrases are contained within phonological intonational phrases (which typically correspond with clauses). This phonological analysis is presented in the section of Table 6 labeled, “3. Phonological prosody”.

In the section of Table 6 labeled “4. Syntactic structure”, syntactic elements, such as nouns (N) and verbs (V), and their corresponding noun phrases (NP) and verb phrases (VP), which together make up the sentence (S), are presented.

\(^5\) The hierarchy used in Table 6 includes only the phonological prosodic levels that this study is concerned with, and omits commonly described phonological prosodic levels (e.g., prosodic foot).
Table 6: Disambiguation of relationship among perceptual, acoustic, and phonological viewpoints towards prosody

1. Perception of prosody

LARGE CAPITAL TEXT denotes stressed syllables, and lowercase text denotes no stress.

<table>
<thead>
<tr>
<th>Intonational pitch contour</th>
<th>Phrasal stress</th>
<th>Lexical word stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>[The SLUGgers BOXed in the CROWD]</td>
<td>[The SLUGgers] [BOXed] [in the CROWD]</td>
<td>[SLUGgers] [BOXed] [CROWD]</td>
</tr>
</tbody>
</table>

2. Acoustic production of prosody

Fundamental frequency (F0) (indicated by a blue line, which ranges from approximately 84 Hz to 126 Hz),
Duration (bounded by dotted vertical lines that indicate relative temporal differences among segments),
Intensity (green line, which ranges from approximately 22 dB to 58 dB)

<table>
<thead>
<tr>
<th>F0 contour</th>
<th>Phrasal stress measurements of F0, vowel duration, and intensity</th>
<th>Lexical word stress measurements of F0, vowel duration, and intensity</th>
</tr>
</thead>
</table>

The sluggers boxed in the crowd

3. Phonological prosody

Phonological intonational phrase (IP)
Phonological prosodic phrases (ɸ)
Phonological prosodic words (ω)

[The sluggers boxed in the crowd]_IP  [The sluggers]_ɸ [boxed]_ɸ [in the crowd]_ɸ  [sluggers]_ω [boxed]_ω [crowd]_ω

4. Syntactic structure

Sentence (S)
Noun (NP), verb (VP), prepositional phrases (PP)
Lexical words (N, V)

[The sluggers boxed in the crowd]_S  [The sluggers]_NP [boxed]_VP [in [the crowd]_NP]_PP  [sluggers]_N [boxed]_V [crowd]_N
Definitions of prosodic terminology

Prosodic terminology is notoriously fraught with complications and involve debates over which terms cover which concepts and phenomena. For example, within the prosodic phonological hierarchy of the intonational phrase, prosodic phrase, prosodic word, and syllable, the most prominent pitch event within an intonational phrase “anchors on” the most prominent syllable in a prosodic word. So, pitch events and sentence stress in an intonational phrase at the sentence level travel through the hierarchy, through the prosodic phrase, through the prosodic word, and right down to the ground floor to the syllable. The complicated nature of stress led Hayes (1995) to describe stress as, “parasitic, in the sense that it invokes phonetic resources that serve other phonological ends” (Hayes, 1995, p. 7). To avoid potential confusion, the following definitions of prosodic terminology are intended as a guide for how terms will be used in this study and not as an argument for or against terms and the semantic territory they include or exclude.

Lexical stress/prosodic word stress

Lexical stress denotes the most prominent syllable(s) in a lexical word. Lexical words obligatorily receive lexical stress. Single syllable lexical words have stress on its only syllable, but in multisyllabic lexical words, at least three levels of lexical stress are possible, which are evident in examples like PUGILISM and SUCKER-PUNCH (LARGE CAPITAL TEXT denotes primary word stress, SMALL CAPITAL TEXT denotes secondary stress, and lowercase text denotes

---

6 Lexical word stress is connected to phonology by prosodic words, to syntax by lexical words, to semantics by content words, and to acoustics by the physical cues of duration, F0, intensity, and vowel quality.
no stress). Two syllable words may include one primary stress and one unstressed syllable (e.g., RABbit) or one primary and one secondary stress (e.g., SOUTHpaw). The presence of lexical stress is perceived via contrasts in duration, loudness, pitch, vowel quality and other perceptual cues. Perceptual cues (which are processed in the auditory system) are indirectly measured by the physical acoustic cues of duration, intensity, F0, and vowel formants, respectively.

**Phrasal stress/prosodic phrasal stress**

Phrasal stress (also called “nuclear stress”) also denotes the most prominent syllable within a phonological prosodic phrase, which typically aligns with syntactic constituent phrases such as noun and verb phrases. The head of a prosodic phrase, a noun (SLUGgers) in this example, or a verb, (BOXed), receives primary word stress while function words (e.g., The, in, the) receive no stress. Prosodic phrases that contain more than one lexical word may include secondary stress on the non-head lexical words (e.g., [The mighty SLUGgers]).

**Sentence stress/intonational phrase**

Sentence stress and primary stress have often been used interchangeably, which unfortunately creates some confusion. The term, sentence stress, is straightforward and analogous to word stress and phrasal stress: sentence stress (“pitch accent” is also used) refers to

---

7 Vowel quality is measured in Hertz via vowel formants. Vowel formants give vowels their distinct and recognizable character. The examples in Chapter 1 of vowel quality in the pronunciation of “we’ll” as [wiˈl], [wɪl], and [wəl] can be expressed in acoustic measurements of vowel formants. The first two vowel formants for [i] are 280 Hz and 2250 Hz. The first two vowel formants for [ɪ] are 400 Hz and 1920 Hz (Ladefoged & Johnson, 2010). As a speaker shifts vowel quality from [ɪ] to [ɬ] to [ə], the first formant rises in frequency while the second formant declines in frequency.
the most prominent syllable in a sentence (Pierrehumbert, 1980; Welby, 2003). The domain of
the intonational phrase is the sentence, and as such, sentence stress usually aligns with the most
prominent stress or pitch event in an intonational phrase as well as longer relative duration. The
conflation of these two names for one phenomenon has a logical source. Primary word stress
marks the most prominent syllable in a lexical word, and also in a prosodic phrase. The most
prominent stress in a sentence is elevated beyond lexical or phrasal stress, although sentence
stress “anchors on” the lexical head of one of the prosodic phrases in the sentence. If a sentence
has only one prosodic phrase, or one lexical word, then sentence stress may be conflated with
lower levels of the prosodic hierarchy (Beckman & Edwards, 1994). We will see examples of
this conflation below.

However, when measuring stress relationships in a sentence, it is important not to
conflate stress relationships between lexical words and sentences. Lexical stress is, of course,
lexically determined; that is, lexical stress placement is not optional. Sentence stress (of the non-
contrastive and non-focused variety), on the other hand, may occur on the head of any prosodic
phrase and occur as the most prominent stress within a sentence. So, in the absence of context for
our example sentence from Table 6, “SLUGgers,” “BOXed,” and “CROWD” are all eligible to
receive sentence stress.

Focus

Prosodic focus is a discourse level phenomenon that relies on context for its deployment.
Focus may be used to mark new information, contrastive information, or comparative
information. Focus may be applied to a lexical word head of a prosodic phrase (called narrow
focus), an entire prosodic phrase, or an entire sentence (called broad focus). Again, focus is the
most prominent prosodic event at a discourse level, and focus is more perceptually prominent than nearby sentence stress events, and may be acoustically marked by higher measurements of all three traditional acoustic cues of duration, F0 and intensity (Breen, Fedorenko, Wagner, & Gibson, 2010), or acoustically marked by low nuclear pitch accent or pre-nuclear pitch accent (Pierrehumbert, 1980). A hierarchical schematic in the tradition of metrical stress theory (Hayes, 1995) is provided in Figure 8 shows a conceptual phonological representation of stress levels, which is unambiguous when compared to the tangle of physical acoustic correlates of stress in the example sentence, “The sluggers boxed in the crowd” (see Table 6, section 2. Acoustic production of prosody). It is not obvious which lexical word (sluggers vs. boxed vs. crowd) carries sentence stress when looking the acoustic measurements. For that reason, phonological representations of stress and intonation have taken many visual forms to clearly represent attested and meaningful prosodic components (Beckman, Hirschberg, & Shattuck-Hufnagel, 2006; Bolinger, 1986; Cruttenden, 1997; Hayes, 1995; Ladd, 1996; Pierrehumbert, 1980; Schmerling, 1976).

Figure 8: A conceptual schematic of prosodic levels
Prosodic prominence

The basic layout of stress levels have been defined, but these levels of stress only apply to understanding the literature review that follows. The present study will not use any of these definitions of stress in the data that it measures. Following previous research (Aylett & Turk, 2004; Calhoun, 2010; Cole, Mo, & Hasegawa-Johnson, 2010; Shattuck-Hufnagel & Turk, 1996), this study will use the term “prosodic prominence” to refer to an acoustically measurable prosodic distinction between adjacent function and content words but without reference to a specific phonological domain. The prosodic phonological relationship between a lexical content word and function word is expected to be lexically stressed and not stressed, respectively. The stress relationship between adjacent function and content words has typically been associated with phrasal stress at the level of the prosodic phrase. However, “prosodic prominence” will be used in favor of “phrasal stress” for several reasons. First, because the term, “stress” typically excludes focus, and although the study does not eliminate all instances of focus, the study does eliminate the most likely position for prosodic focus⁸, the end of the sentence (Calhoun, 2010; Pierrehumbert, 1980). Second, this study is limited to collecting two-word tokens, which means that any prosodic phrases that are three words or greater in length would be omitted from the data. Therefore, the term “prosodic phrasal stress” would be an inaccurate description of the data. Third, this study allows data to be collected from both inter- and intra-prosodic phrase function-content word pairs, which would make “prosodic phrasal stress” an inaccurate description of the stress that occurs within a prosodic phrase. Fourth, because of the ambiguity of the prosodic nature of the data that will be measured, the superordinate term “prosody” rather than “stress” will be used.

⁸ The end of the sentence as the most likely position for prosodic focus will be discussed in Chapter 4.
Prosody refers to suprasegmental features of speech in any language, but as this study measures the prosody of L1 Chinese NNESs who are examinees in an oral proficiency test, the literature review will focus on English prosody from both non-native and native speakers of English. Although the reader now has a general understanding of prosody, there remains the question of why we should be concerned with prosody. And the best way to answer that question lies within the perceptual approach to prosody.

The Role of Prosody in Comprehending Speech: A perceptual description of prosody

Prosody contributes a fundamental component of language that enables efficient comprehension of English speech (Birch & Clifton, 1995; Bock & Mazzella, 1983; Cole, 2015; Cutler, Dahan, & van Donselaar, 1997; Frazier, Carlson, & Clifton, 2006; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Tyler & Warren, 1987; Wagner & Watson, 2010; Watson & Gibson, 2005). At the level of discourse, prosody adds structure and meaning to spoken text; it works as structural and semantic glue that establishes relationships among words within a phrase, among phrases within a sentence, and among sentences within a larger discourse (Eskenazi, 1999). Prosody creates acoustic distinctions across multi-word constituents within a sentence (Price et al., 1991) and signals relationships among the lexical elements within the boundaries of those constituents (Beckman & Edwards, 1994). Prosody signals new versus given information (Bock & Mazzella, 1983; Nooteboom & Kruyt, 1987) and indicates speaker attitude and emotion (Liberman, 1979). Conversely, if a speaker does not provide expected prosodic patterns, the time that a listener requires to process language increases, and the comprehensibility of spoken English decreases (Birch & Clifton, 1995).
To reach many of these conclusions, word detection tasks are commonly used in speech perception experiments, and particular in studies of the effects of prosody on speech perception. Word detection tasks measure the time a listener requires to recognize a target word (or other target text) under the effects of independent variables, such as prosody that has been digitally manipulated to remain flat and emphasize no syllable above other syllables. These word detection latencies provide information about how listeners perceive and process language.

**Word detection within a lexical word**

Cutler and Norris (1988) measured listeners’ word detection rates while systematically varying lexical stress assignment. Lexical stress or word stress is simply the prominence of a syllable (or syllables) over other syllables within a word. Lexical words, whether monosyllabic or multisyllabic, obligatorily carry stress. Working from these assumptions, Cutler and Norris measured word detection rates of monosyllabic English words embedded within nonce disyllabic words that stressed either the target word that participants were listening for, or the nonce ending of the disyllabic word, or both syllables. For example, acoustically equal realizations of the target word *mint* were embedded within differing lexical stress environments such as MINTesh, minTAYVE, and MINTAYF (CAPITAL TEXT denotes a stressed syllable). They found that 1) a stressed syllable signals lexical word segmentation, 2) that word detection rates were significantly faster when the target word (MINTesh) was stressed than when the nonce portion (minTAYVE) was stressed, and 3) that word detection rates were also significantly faster when the target word (MINTesh) was stressed than when both target word and nonce portion were stressed (MINTAYF). The authors argue that “weak-strong” and “strong-strong” prosody initially leads listeners to erroneously assemble a segmental model that spans both syllables.
(min-TAYVE and MIN-TAYF), which delays recognition of the target word, as opposed to a model where the stressed syllable and target word align (MINT-esh) with expected English stress patterns, which facilitates recognition of the target word.

**Word detection in the prosodic phrase**

Tyler and Warren (1987) tested to what degree listeners use prosodic structure versus syntactic structure to develop a mental representation of spoken sentences. One of the experiments focused on the prosodic phonological phrase and syntactic organization. The authors constructed sentences with syntactically normal prose (A) as well as syntactically scrambled prose (B). The scrambled sentences were syntactically normal at the prosodic phrase level, but nonsensical at the sentence level.

A. Normal prose: The maid / was carefully peeling / the potatoes / in the garden / because during the summer / A HOT KITCHEN / is unbearable / to work in.

B. Scrambled prose: Because during the summer / in the garden / was carefully peeling / the potatoes / the maid / A HOT KITCHEN / is unbearable / to work in.

The pairs of sentences were also presented under one of two prosodic conditions: with normal, expected prosody between prosodic phrases and with digitally disrupted prosody between prosodic phrases. These four conditions (1. normal prose + normal prosody, 2. normal prose + disrupted prosody, 3. scrambled prose + normal prosody, and 4. scrambled prose + disrupted prosody) were presented to listeners in a word recognition task (the target word was “kitchen” in these examples). The authors found that word recognition was significantly faster in the semantically normal prose condition than in the semantically nonsensical scrambled
condition when both were spoken with normal prosody. More notably, when spoken with disrupted prosody, word recognition times were equally slow for the two semantic conditions.

The results of the experiment support the importance of the role that prosody plays when listeners process spoken text. The authors concluded that “listeners always use prosodic structure as they attempt to interpret a sentence, and [prosodic structure] is not merely used as a fallback when a phrase cannot be interpreted semantically” (Tyler & Warren, 1987).

**Prosody and the given-new contract**

In a study that examined how participants were affected by placement or non-placement of prosodic focus-marking on given and new information, Terken and Nooteboom (1987) found that participant reaction times are faster when given information is unfocused (versus focused) and when new information is prosodically focused (versus unfocused). The authors argued that delayed processing times result from an informational-prosodic mismatch – that, for example, upon hearing an unfocused noun, a participant would first mentally search the list of nouns that have been established in the present discourse context, and, upon finding no matches, create a new mental representation within the discourse. This situation involves more cognitive steps and thus a longer reaction time than if a participant hears a focused noun and creates a new mental representation in the discourse. They argued that the reverse was also true – that prosodically focused given information leads listeners to reflexively begin to create a new mental representation before backtracking and retrieving the previously given information.

Birch and Clifton (1995) supported Terken and Nooteboom’s findings that discourse expectations of the given-new contract are strongly supported by prosodic focus-marking. They showed that listeners overall judged sentences to be more appropriate, and response times were
shorter when new information was prosodically focused when given information was also
focused, which violates the expected prosodic pattern for given information. Birch and Clifton
found that “listeners showed less processing difficulty for dialogues in which a given argument
NP was accented if the new verb was accented, compared to when [a given argument NP was
accented, and] the new verb was not accented” (Birch & Clifton, 1995). (See Table 7.)

Table 7: Birch and Clifton’s (1995) Given-New Contract and Processing Difficulty

<table>
<thead>
<tr>
<th>Accented</th>
<th>Accented</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given NP</td>
<td>New VP</td>
<td>= Less processing difficulty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accented</th>
<th>Unaccented</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given NP</td>
<td>New VP</td>
<td>= More processing difficulty</td>
</tr>
</tbody>
</table>

These preferences might be explained in an optimality theory paradigm in which pairing
prosodic focus with new information and no focus with given information are phonological
constraints that may be violated, but, the more violations that occur, the less acceptable the
utterance becomes.

Nootboom and Kruyt (1987) also found general listener preference for prosodic focus
marking on new information and lack of focus on given information, but, unsurprisingly, the
relationship between information status and focus is non-binary and complicated. For example,
they found that a multi-word constituent, such as “the mayor of OXFORD”, was given high
acceptability ratings where the entire phrase is presented as new information, but only OXFORD
is focused. So, it seems that an argument, if focused, can carry focus to the unfocused head of the
phrase.
Word detection in the intonational phrase

In a study conducted in Dutch, which, like English, is both stress-timed and an intonation language, Braun, Dainora, and Ernestus (2011) presented listeners with sentences that had either a normal intonational contour produced by a speaker of Standard Dutch, or a digitally manipulated intonational contour. The unfamiliar intonational contour traced its fundamental frequency values along the curve of one period of a sine wave across the length of a sentence, which, the authors explain “is articulatorily possible, but - to our knowledge - unattested in any language” (p.353). Three word monitoring tasks that measured participant reaction times were conducted under these two intonation conditions, and, in all three experiments, reaction times were significantly slowed under the unfamiliar intonation condition.

In the first experiment, participants monitored content and function words, and, while the unfamiliar intonation delayed identification times for all words, function words were identified much more slowly - about 139ms slower than content words on average. This reaction time difference is likely attributable to the lack of prosodic distinction between content and function words as, in English, listeners both acoustically create and also auditorily perceive prosodic distinctions between content and function words (Cole, Mo, & Hasegawa-Johnson, 2010). In addition to slower reaction times for function words, reaction times increased as the experiment progressed, which suggests that these delaying effects are cumulative.

The second and third experiments used the normal and unfamiliar intonational conditions in cross-modal identity priming tasks, which involve alternating methods of communicating priming words and their targets. For example, a priming word might be printed text while the target word is auditory. In both experiments, the unfamiliar intonation condition slowed reaction times. Braun, Dainora, and Ernestus argue that, more importantly, based on the links between
conceptual representation of a word and varying modes of references to that word (Norris, Cutler, McQueen, & Butterfield, 2006), these two experiments show “that intonation has a direct effect on lexical access…[and t]he effect of an unfamiliar intonation contour is then robust and long lasting” (Braun et al., 2011, p. 363).

**Summary of the perceptual importance of prosody**

These individual perceptual studies provide strong evidence that prosody plays a central role in processing spoken language in a variety of experimental conditions. If we consider these studies as a group, they show that multiple layers within the prosodic hierarchy (i.e. lexical stress, the prosodic phrase, the intonational phrase, and discourse level prosody that is represented by focus, which signals information status) are all involved in providing structure to speech and aiding listener comprehension.

Moreover, these perceptual studies demonstrate that listeners have expectations of prosodic structure. Listeners expect lexical words to be more prosodically prominent than non-lexical words, listeners expect new information to be more prosodically prominent than given information, and listeners expect intonational information to serve as a guide to information structure. Without receiving this expected information, listeners’ reaction times are slower and/or processing difficulty increases. Now that the perceptual importance of prosody has been established, we move to the question of how the speaker acoustically communicates that prosodic information to the hearer.
An Acoustic Description of Prosody

Problems in acoustic studies of prosody

“Prosody” is a superordinate term for all types of suprasegmental phenomena, which is helpful when speaking in general terms. But when we start to quantitatively describe common prosodic phonological features such as lexical stress, or focus, or nuclear pitch accent, it is helpful to associate those particular prosodic features with measurable acoustic cues. Toward that end, it would be more accurate to say that there have been themes on how prosodic features may be physically realized rather than a consensus among quantitative researchers. In recent years, prosodic research has often targeted a specific prosodic-syntactic context such as *nuclear accented lexical stress, prosodic focus, or prosodic boundary effect* in an effort to disambiguate in which contexts these acoustic cues typically occur. Despite efforts by researchers to disambiguate prosodic-syntactic contexts, a confusion of terminology persists. For example, prosodic focus, sentence stress, phrasal stress, phrasal accent, and phrasal prominence may all refer to the same phenomenon that perceptually elevates one word (or group of words) above all of the other stressed syllables within the boundaries of an intonational phrase (Turk, 2014). Beckman and Edwards (1994) broadly critiqued studies of acoustic cues of stress that have produced conflicting results over the years because they have been confounding multiple levels of prosody while claiming to measure lexical stress. They argue that previous lexical stress studies had not controlled confounding variables sufficiently to distinguish which prosodic cues operate in which prosodic contexts. Presciently, in one of the first quantitative studies on lexical stress, Fry (1958) warned against this conflation of interconnected prosodic systems.

Despite warnings from prominent researchers in the field, this problem has continued to appear in recent studies. For example, Kondo (2009) claimed to measure lexical stress while
actually measuring prosodic focus in an intonational phrase, and Sereno and Jongman (1995) claimed to measure lexical stress but used a citation form that includes a complete intonational phrase. In contrast, de Jong (2004) avoids conflating focus with lexical stress by eliciting lexical stress (along with boundary effects) with carrier sentences such as “HE said X, not her,” where X is a stressed but unfocused target word.

Concurrently during this history of conflation of prosodic systems, the traditionally measured acoustic cues of English prosody have been vowel duration, vowel intensity, vowel quality, and fundamental frequency. Other acoustic measurements such as spectral balance (Sluijter & van Heuven, 1996b; Sluijter, van Heuven, & Pacilly, 1997), pitch slope (Plag, Kunter, & Schramm, 2011), and peak pitch location (Zhang, Nissen, & Francis, 2008) have also been reported as influential in the acoustic realization and perception of stress. What follows is a description of acoustic cues and the prosodic contexts in which they occur.

**Acoustic measurements of lexical stress**

Fry (1955) measured duration and intensity in five stress-based minimal pairs (*contract*, *digest*, *object*, *subject* and *permit*) and found that both acoustic cues of duration and intensity were correlated with stress judgments, but that duration was a much stronger indicator of stress. Beckman (1986) measured duration, intensity, and F0 for the same five stress-based minimal pairs and also found duration and intensity to be physical cues of lexical stress production. F0 levels showed no statistical difference between stressed and unstressed syllables.

In 1996, two articles appeared that questioned assumptions of simplicity in the relationship between lexical stress perception and acoustic cues. Turk and Sawusch (1996) designed experiments that used a flat F0 contour, which eliminated F0 from consideration as a
cue to stress, while at the same time they digitally manipulated values of duration and intensity. They found that duration and intensity were integrated with one another and were not processed as separate percepts as previous work had suggested. The results showed that the duration of a syllable affected the perception of intensity and vice versa. For example, while holding intensity values constant but varying durational values, test subjects reported changes in perceived loudness although intensity had remained constant. Turk and Sawusch also found that although these two acoustic cues were perceived in combination with one another, the strength of duration as a perceptual and acoustic cue to stress was much stronger than intensity. That is, changes in duration affect stress perception more effectively and reliably than changes in intensity.

Sluijter and van Heuven (1996b) were skeptical of prior research (particularly Beckman, 1986) that employed unrealistic experimental manipulations of intensity by raising or lowering intensity uniformly across the speech frequency spectrum. They argued that when humans increase glottal effort, the effect on intensity production is not uniform across all frequencies of human speech. Sluijter and van Heuven took a more nuanced approach to intensity by including an acoustic cue they term spectral balance, which is akin to the amount of effort a speaker is making on a given syllable. Spectral balance refers to where within the speech frequency range intensity is concentrated, for example, at lower (100 to 500 Hz) or higher frequencies. Using read-aloud data, the researchers measured syllable duration and contrasted two experimental conditions of intensity measurement. The first was a more realistic raising of intensity at frequency levels above 500 Hz and the second was an unrealistic uniform manipulation of intensity. They found that in the first spectral balance condition, duration and realistically manipulated intensity were nearly equal in terms of predicting lexical stress but, in the second condition, duration carried about 90% of the predictive power, leaving uniform raising of
intensity as a very weak predictor of stress. This study along with Turk and Sawusch (1996) demonstrate that while duration and intensity are indeed instrumental in the physical realization of stress, quantitative researchers should also recognize a complex relationship between duration and intensity.

**Measurements of primary lexical stress, secondary lexical stress, and no stress**

De Jong (2004) found very large durational differences of syllables in both monosyllabic and disyllabic words depending on the type of stress effects that they receive. He varied elicitation sequences that produced primary lexical stress that excluded prosodic focus (i.e., a nuclear accented vowel in an intonational phrase such as the /æ/ in bat in the sentence, \textit{HE said bat, not her}), secondary lexical stress (compounded with another word that receives primary lexical stress (e.g., /ɛ/ in \textit{sports bet})), and an unstressed condition (an unstressed vowel in a disyllabic word (e.g., /i/ or /ə/ in \textit{rabid}). De Jong found that the syllables with primary lexical stress were more than double the duration of unstressed syllables, and syllables with secondary stress were about 75% longer than unstressed syllables. De Jong also considered consonant voicing effects on the duration of a preceding vowel (e.g., rabid vs. rabbit), and found that the unstressed syllables showed no significant durational effects with regard to voicing. This study demonstrates a reliable durational difference among three categories of stress.

Aylett and Turk (2004) used a corpus of 15-hours of spontaneously spoken Glaswegian English compiled from 64 speakers to study durational effects of lexical type, syntactic context, and informational redundancy. After normalizing durational differences among the 64 speakers and eliminating variables of syntactic position and informational redundancy, the authors found that syllables that are ineligible to receive lexical stress (i.e., function words and unstressed
syllables in lexical words) were over three-quarters of a standard deviation shorter in duration than syllables in lexical words that are eligible to receive either primary or secondary stress. This finding indicated that stressed syllables are durationally much longer than unstressed syllables.

Shi, Gick, Kanwischer, and Wilson (2005) measured duration and intensity of 90 read-aloud tokens of function (should, could and must) and content (good and best) words. As function words are phonologically described as unrecognized by stress, they expected that the three function words would be quantitatively reduced when compared to the content words. They also asked the question of whether high-frequency words (should, could and good), which tend to be less carefully articulated as their frequency rises, are also quantitatively reduced when compared to low frequency words (must and best). They found that the answer to both of these questions was yes. Function words had statistically different lower values than content words for both duration and intensity, and high frequency words had statistically different lower values than low frequency words. This study along with de Jong (2004), Aylett and Turk (2004), and Shi (1995) (reviewed in Chapter 1) demonstrate that durational differences play a central role, and intensity to a lesser extent, in stress distinctions among all types of stress.

**Measurements of sentence stress and prosodic prominence**

Adams and Munro (1978) attempted to answer the question “…in connected [speech] what does a speaker do that causes the listener to receive an impression of stress?” (p. 127). In their study, 10 native English speaking non-linguist raters judged sentence stress in recordings of 8 native speakers of English and 8 high proficiency non-native speakers of English (from a variety of L1s) who read prepared sentences aloud. Those perceptual ratings of stressed and unstressed syllables were then compared with acoustic measurements of duration, F0, and
intensity. They found that duration was significantly associated with sentence stress for 13 of the 16 subjects (7 native English speakers and 6 NNESs) and that intensity was statistically unassociated with stress. They found that F0 was associated with stress for 7 individuals (2 native English speakers and 5 NNESs) but not for the other nine, which agrees with later findings that F0 can be used as an indicator of stress but is unreliable across speakers (Sluijter & van Heuven, 1996a).

Working with the knowledge of past discrepancies in measurements of stress and prosodic prominence, Kochanski, Grabe, Coleman, and Rosner (2005) set out to describe acoustic cues of “prosodic prominence” in a large corpus of a variety of styles of speech rather than just from strictly controlled scripted speech. They obtained their data from the Intonational Variation in English (IViE) Corpus (Grabe, Post, & Nolan, 2001), which, in Kochanski et al.’s sample, included 240 minutes of audio data from 6 speakers of each of 7 dialects from England and Ireland. About half of the data came from spontaneous speech and the other half from a variety of read-aloud scripts.

Kochanski et al. (2005) used the annotations of prosodic prominence that are included as part of the IViE Corpus (Grabe et al., 2001). The annotations of prosodic prominence were impressionistically judged by two of the authors of the IViE Corpus, who are phoneticians at Cambridge University. The term, “prosodically prominent” was a binary question: either prosodically prominent or not. Kochanski et al. found that among duration, intensity, and F0, intensity was the best predictor of prosodic prominence, which was followed by duration. They determined that pitch and spectral slope (which is similar to spectral balance) were very weak predictors of prominence.

Cole, Mo, and Hasegawa-Johnson (2010) used unscripted conversational speech from the
Buckeye Corpus to measure “prosodic prominence,” which they defined to untrained raters as: words that are more prominent than surrounding words in a phrase. With this definition of prosodic prominence, the study likely measured both prosodic focus and primary unfocused stress [as we saw in de Jong (2004) above]. The study employed 91 raters to make this determination on two short audio excerpts from 36 speakers and one longer excerpt from 18 of the 36 speakers. Duration, F0, intensity, and spectral emphasis were measured for the syllables that raters agreed were prosodically prominent. Prosodically prominent syllables showed significant positive correlation with both duration and overall intensity, although duration was the stronger predictor. Spectral emphasis was mixed. For 5 English vowels, spectral emphasis above 500 Hz was more highly correlated with prosodic prominence than either duration or overall intensity. F0 had either weak or no correlation with prosodic prominence. As other studies have shown, intensity and spectral emphasis may play a role in prosodic prominence, but duration is stronger than any of the other predictors.

**Summary of description of acoustic prosody**

Figure 9 presents the findings from the acoustic prosody literature review. All of these studies demonstrate that a combination of elevated levels of duration and intensity are associated with stress, which includes primary and secondary lexical stress, as well as “prosodic prominence,” where binary, stressed vs. unstressed evaluations are made. Acoustic cues of stress are primarily duration and secondarily measurements of intensity, whether it is average intensity of a vowel or intensity that is concentrated higher in the speech frequency spectrum. Recall from Chapter 1, prosodic focus and sentence stress are associated with elevated F0 values, as well as with longer duration and less reliably with higher intensity.
We’ve seen evidence for how listeners mentally perceive prosody and we’ve seen acoustic measurements that support listeners’ perception of prosody. We now turn our attention to how these prosodic features have been phonologically described.

A Phonological Description of Prosody

As acoustic descriptions of prosody have evolved, so too have phonological descriptions of prosody. Phonological prosody attempts to describe underlying systems of observable acoustic prosody and associate prosody with linguistic meaning and purpose. Prosodic descriptions have taken the form of discrete phonological features and linear phonological rules in a generative grammar framework (Chomsky & Halle, 1968) or as phonological constraints in an optimality theory framework (Shattuck-Hufnagel & Turk, 1996). Individual prosodic components may also
be described separately from one another, as in the ToBI system (Tones and Break Indices) to describe intonational contours and boundaries (Beckman et al., 2006).

One of the challenges of describing phonological prosody is to distinguish observably different prosody from meaningfully different prosody. Establishing a “meaningful difference” is both similar to and different from the practice in segmental phonology, which has been much more straightforward than prosodic phonology. Segmental phonology describes phonemically distinct sounds that change the meaning of words, and also describes allophonic segmental sound changes that do not change the meaning of a word. In segmental phonology, questions generally have binary answers like, “does this sound change produce a change in meaning?” Or, “does this sound change produce a word that exists in this language?” In perceptual studies on vowel segments, there tends to be a sharp division as vowel formant frequencies change. So as a vowel is synthetically shifted at equidistant intervals from [æ] to [ɛ] in the carrier words bat and bet, respondents tend to choose one vowel or the other and not vacillate between the two choices.

Prosodic phonology is occasionally phonemic also, but more frequently the answers to questions are more complicated and complex than in segmental phonology. Prosodic phonology may include questions like the one above, “Does a stress shift in this word produce a change in meaning in your dialect?” In the case of a stress based minimal pair like CONverse and conVERSE, the answer is yes in many English dialects. But sometimes making a binary judgment is difficult. “Does a stress shift in this word produce a change in meaning in your dialect or an

---

9 One metric of straightforwardness is the difference in scholarly response to Chomsky and Halle’s (1968) treatment of segmental phonology compared to the response to prosodic phonology. Chomsky and Halle’s treatment of segmental phonology became mainstream in the years following The Sound Pattern of English (despite criticism from advocates of optimality theory) while at the same time, Chomsky and Halle’s use of the same paradigm in prosodic phonology was widely critiqued and revised (see Jackendoff, 1972; Liberman & Prince, 1977; Selkirk, 1980).
acceptable pronunciation in your dialect?” With example data such as Invoice versus inVOICE, or COMputer or keyBOARD, these answers are less clear and the answers depend on how strictly we adhere to lexical stress conventions.¹⁰

Intonational contours may initially appear uncomplicated. For example, a falling intonational boundary produces a statement in, “George and Mary give blood.” Whereas a rising intonational boundary produces a question in, “George and Mary give blood?” But intonational phonology simultaneously includes common questions like, “Which constituent is receiving the greatest prosodic prominence?” and, “How is prosody providing a phonological map of information structure?” There are also questions that necessitate separating the signal from the noise, such as, “What emotional information is the speaker conveying beyond indicating whether this is a statement or question?” And, “What prosodic variation is attributable to this individual speaker and unrelated to the questions above?”

Information structure, function and content words

It has been stated above that this study will measure acoustic cues of pairs of adjacent content and function words as a proxy for prosody. Pairs of content and function have been selected because the contrast directly reflects word class and also, by extension, information structure. Information structure is typically concerned with acoustically prominent features such as focus, givenness, and topic (Breen et al., 2010). As these features of information structure are acoustically connected to content words (or lexical words or prosodic words), this study is concerned with a prosodic phenomenon of information structure that acoustically elevates relatively meaningful content words above prosodically unrecognized grammatical words. But

¹⁰ Lexical stress figures into the limits that are imposed on data collection in Chapter 4.
before discussing the phonological relationship between function and content words, reviewing a bit of history is helpful.

**Prosodic phonology in Generative Grammar**

Early in the modern phonological investigation of stress and prosody, Chomsky & Halle (1968) described prosodic systems within a generative grammar framework, which used conventions like [±stress]. Lexical morphology and syntax determined the location and the strength of stressed syllables relative to neighboring syllables, and changes in morphology and in constituent structure triggered transformational rules within a generative grammar framework. For example, “black” and “board” both receive primary lexical stress (or in their terms, “main stress”) as they are content words, but how they are combined in a phrase determines their stress relationship. So, despite identical surface forms, [[black ADJ [board] N] NP has primary stress on “board” while [blackboard] N has primary stress on “black” because of underlying syntactic structure (Chomsky & Halle, 1968, p. 16). Liberman and Prince (1977) responded to these proposals on stress by further developing a representation of stress as a hierarchical metrical grid that includes “transderivational” stress changes. Their metrical stress theory was designed to maximize its generalizability to English stress. The metrical stress theory accommodated a single lexical word with one morpheme, its transderivations of stress placement through addition of morphological derivations, and any number of words within an entire sentence. More importantly, Liberman and Prince (1977) argued

[T]hat relative prominence is defined between phonological constituents, rather than on individual segments. Prominence, so defined, is projected onto syllables by associating them with a "metrical grid", which can be thought of as a hierarchy
of intersecting periodicities (rather than constituents), the structure of which is relevant to phenomena of rhythm and timing” (p. 333).

Working from this notion of a metrical grid, Selkirk (1980, 1984) developed a description of phonologically interactive layers of prosody called a Strict Layer Hypothesis (SLH). The SLH first borrows a bottom-up description of prosody that begins from the metrical grid, which combines strong and weak syllables to build prosodic feet, prosodic feet that build prosodic words, prosodic words that build phonological phrases, and phonological phrases that build intonational phrases. The SLH adds to the metrical grid a top-down hierarchical description of dominance. That is, the intonational phrase dominates the phonological phrase, and the phonological phrase dominates the prosodic word, and so on down to the syllable level. This very phenomenon affected a study mentioned in the above section, “Problems of Acoustic Studies of Prosody,” where the researchers intended to measure acoustic correlates of lexical stress but prominence associated with the phonological phrase dominated the acoustic data that was collected (J. A. Sereno & Jongman, 1995). This provides one example of how prosodic phonological theory has a direct impact on experimental design of studies of the acoustics of prosody.

**Prosodic phonology in Optimality Theory**

Selkirk’s later work (2004) describes the prosodic relationship between content words and function words, and translates the relationships of the SLH into phonological constraints in an Optimality Theory (OT) framework. Using OT to describe prosodic phonology allows prosodic structure to emerge from among prosodic constraints. This framework is preferable
since there is a potential for metrical stress theory to overdescribe lexical stress placement to a degree that it rules out attested physical realizations of stress.

One of Selkirk’s constraints creates a hierarchy of layers, as we saw in the SLH. Another constraint, which is central to the theoretical framework for this study, deals with prosodic treatment of content words. Recall that content words, lexical words, and prosodic words all refer to the same word from different viewpoints of analysis. Starting with the phonological prosodic phrase, we see the relationship between prosodic rules and syntax in Selkirk’s (2004) Word Alignment Constraint (1) that implicitly states that for phonological prosodic phrases, only lexical words are prosodically recognized in the domain of a prosodic phrase while non-lexical function words remain unrecognized and therefore unstressed. “[P]rosodic structure makes no reference to functional categories at all” (Selkirk, 2004, p. 468). The second aspect of (1) shows that a given language assigns stress to either the left or rightmost position within the prosodic phrase.

(1) The Word Alignment Constraints (WdCon)

i. Align (Lexical word, Left; Prosodic word, Left) (= WdConL)

ii. Align (Lexical word, Right; Prosodic word, Right) (= WdConR)

The examples for English in (2), (3) and (4) below show that prosodic words (w) align on the right edge of a prosodic phrase (φ) according to constraint (1) above. The verb “boxed” in (2) is analyzed as a prosodic word while the preposition “in” remains prosodically unrecognized in the prosodic phrase “in the crowd”. Meanwhile, the separable phrasal verb “boxed in” in (3) and (4) is analyzed as two prosodic words. Analyzing “boxed” and “in” in (3) and (4) as two separate prosodic words correctly allows for both phrasal verb separability and main stress to the rightmost prosodic word in each prosodic phrase in both (3) and (4). The right edge of the
prosodic phrases are stressed, which correctly assigns no stress to “in” in (2) but main stress to the phrasal verb particle “in” in both (3) and (4) (Nespor & Vogel, 1986, p. 179). Examples of acoustic measurements are presented for (2) in Figure 10 and (3) in Figure 11 (Blue contours indicate F0 and green contours indicate intensity).


Figure 10: Acoustic measurements for (2)…[[BOXed]ω]φ [in the [CROWD]ω]φ
This stress rule (1) involving adjacent function and prosodic words is key to the design of this study for two reasons: First, it identifies the prosodic expectation applied to the data that will be measured; function words should remain prosodically unrecognized and therefore prosodically less prominent than neighboring stressed prosodic words. Second, studying two syllables on either side of the prosodic boundary between a prosodic word and a neighboring function word avoids conflating the stress relationship of this study with the stress relationship on two syllables within the boundaries of a multisyllabic prosodic word. This design distinguishes between those two stress issues.

In my anecdotal classroom observations, non-native English speakers are quite capable of producing lexical word stress on a multisyllabic word, while at the same time, they are less capable of assigning expected stress within a prosodic phrase. This made the prospect of measuring lexical word stress uninteresting, but made the prospect of measuring the prosodic prominence between neighboring words very interesting. A speaker’s ability to make prosodic distinctions between the two categories of function and content words will be used as a proxy for the speaker’s prosodic performance in English.
Selkirk’s (2004) description of function words as prosodically unrecognized in English comports with findings from multiple acoustic studies. As described above, Shi (1995) and Shi, Gick, Kanwischer, and Wilson (2005) specifically measured acoustic cues on function words. Both studies found that duration and intensity were used to elevate acoustic cues of content words above those of function words. Shi (1995) found that this was true for both native English speakers speaking English and native Mandarin Chinese speakers speaking Chinese. Shi, Gick, Kanwischer, and Wilson (2005) found that word frequency was also reflected in the same measurements such that low frequency words (i.e., lexical words with high information content) were longer and louder than high frequency words (i.e., lexical words that have been repeated and have lower information content). These studies along with Aylett and Turk (2004) and Cole, Mo, & Hasegawa-Johnson (2010) demonstrate that measurements of unstressed syllables (which include function words) are consistently shorter and quieter than stressed syllables (which include content words). All of these acoustic studies support the expectation based on Selkirk (2004) that content words are canonically more acoustically prominent than function words.

**Summary of relevant findings from perceptual, acoustic and phonological descriptions of prosody**

We have seen above that listeners depend on prosody to distinguish word class (function and content words) and information status (given, new, and contrastive information). Perceptually, prosodic prominence draws listener attention toward the stressed text, which facilitates lexical word recognition and information processing. But if information status or word class misalign with listeners’ prosodic expectations, then prosodic misalignment delays word
recognition and increases listener effort. In addition, negative effects of prosodic misalignment are cumulative.

Acoustically, lexical stress is reliably indicated with longer duration, and may also be indicated although less reliably with higher intensity values. Prosodic prominence, which this study measures, may include higher pitch values in addition to longer duration and higher intensity.

Findings from perceptual and acoustic data lead to conclusions that prosodic phonology references word class. Function words remain phonologically unstressed, and content words minimally receive primary stress. In light of these findings, and the international academic context in which this study originated, the research questions are restated:

• Do L1 Chinese speakers of non-native English acoustically distinguish content and function words with prosodic features?
• Which acoustic measurements distinguish content and function words in the speech of L1 Chinese speakers of non-native English?
• How are acoustic measurements of prosodic prominence associated with measurements of English language proficiency of L1 Chinese speakers of non-native English?
CHAPTER 4. METHOD

Participants

The data from OEPT recordings of 10 IGS participants from each of three OEPT levels (35, 45, and 55) were obtained from the OEPP, which yielded a total of thirty participants. The three OEPT levels were chosen because they are non-adjacent in the OEPT scale, and would therefore be more likely reveal differences in treatment of a subtle aspect of English prosody. The participants identified Chinese (either Mandarin or a “dialect” of Mandarin Chinese) as their first language. The participants were graduate students enrolled in the following graduate programs at Purdue: Aeronautics, Astronautics and Engineering; Animal Sciences; Biomedical Engineering; Chemistry; Computer Graphics Technology; Computer Science; Consumer Sciences and Retailing; Earth, Atmospheric, and Planetary Sciences; Economics; Educational Studies; Electrical and Computer Engineering; Electrical Engineering; Industrial Engineering; Interdisciplinary Life Science Program; Interdisciplinary Studies; Management; Mechanical Engineering; Political Science; and Technology.

Syntactic constraints on data collection

Based on Selkirk’s (2004) analysis of the phonological prosody of function and content words, two-word tokens were collected from the data set. As some of the above perceptual, acoustic and phonological studies have pointed out, many factors can influence prosody at a lexical (content) word level. Therefore, it was necessary to establish some constraints on the data that were collected. This chapter describes the restrictions placed on data collection and to acoustic measurements of the collected data.
**Sentence initial constituents**

The beginning of a new sentence begins a new intonational phrase, which typically starts with a high fundamental frequency relative to the middle of the sentence (Wennerstrom, 2001). English sentences usually begin with a determiner phrase, and a common syntactic construction of a noun phrase is a DETERMINER + NOUN, so the function word in the subject position will likely be affected by the onset of an intonational phrase. Therefore, sentence-initial Function-Content tokens were eliminated from consideration as containing potential data, following the methodology in Wennerstrom (1998). Any new occurrences of subject constituents created by a complementizer were also eliminated from consideration.

**Paratone**

In addition to subject noun phrases, I eliminated from consideration introductory phrases before the subject because they typically include a high paratone, (in acoustic terms) an elevated and sustained F0, or (in perceptual terms) an intonational emphasis that acoustically distinguishes the phrase from the rest of the sentence (Wennerstrom, 2001). Figure 12 shows the changing pitch of a speaker who uses paratone during an academic lecture in which she transitions between two plots drawn on a blackboard. Her F0 measurements steeply rise to above 400 Hz at the arrow (†) when she begins her sentence with “Let’s skip for right now…” An excerpt of the text of the lecture follows Figure 12.
Figure 12: Acoustic example of high paratone, which begins with "Let's skip..."
(Wennerstrom, 2001, p. 25)

“DUCKS that had a LÓWER PLÚMAGE RÁTING TÉNDED ÁLSO TO HAVE A LÓWER BEHÁVIORAL RÁTING. (2.1 sec.)

↑ Let's SKIP FOR RIGHT NOW onto the NEXT PLOT . . . ” (Wennerstrom, 2001, p. 24, emphasis in the original)

This type of utterance level intonation may result in acoustically distinct F0 relationships when compared to Function-Content tokens in a sentence-medial position. As paratone is a prosodic domain that may present exactly the type of problems of measuring multiple layers of prosody that were outlined in Beckman and Edwards (1994), it has been eliminated from consideration.

**Sentence-final constituents**

English aligns lexical words on the right edge of a prosodic phrase, so most information-rich words (content words) occur at the right edge of a verb phrase or noun phrase, as we saw in
Selkirk (2004) above. Related to this feature of English syntactic structure are questions about whether English prosodic structure favors the placement of pitch accents (judged by human annotators) at the end of a sentence, and whether English discourse structure acoustically highlights new information with prosodic prominence (measured acoustically).

Calhoun (2010) tested these questions with a large corpus of spoken data: the Switchboard Telephone Speech Corpus, which was collected in 1990 with U.S Government sponsorship, and contains about 2,400 two-person telephone conversations on a wide variety of topics. The corpus was later transcribed and annotated for syntactic and discourse information (Godfrey & Holliman, 1993). Calhoun selected 18 conversations that included 33 speakers, about 2.5 hours of speech, and 10,143 individual words. Calhoun developed a model for prediction of the location of nuclear pitch accents in English. Two annotators marked the “most important” accent that they heard in each fluent phrase; that is, a phrase without pauses. The results of Calhoun’s model showed that among 22 features measured, “[t]he most important feature [to predict the location of a nuclear pitch accent] was relative position in the phrase, with nuclear accents more likely later in the phrase” (p. 1125). Sixty-six percent of nuclear pitch accents occurred on the last word in the phrase, and acoustic measures of prosodic prominence agreed with annotators’ judgments of “most important” accent locations. However, the connection between acoustic measures of prosodic prominence and information status, whether information was new or given, was not significant in this model. Based on Calhoun’s findings, sentence-final content words were eliminated from data collection. In addition, information status of content words was not considered in data collection.

Calhoun also measured prenuclear accents in the Switchboard dataset, that is, lexical stress that precedes nuclear pitch accent but occurs within the same prosodic phase. For example,
the prenuclear accent would refer to an adjective in a noun phrase with the construction

\[[[\text{DET}] + [\text{ADJ}] + [\text{NOUN}]]_{\text{NP}}\]. She found significant ($p = <.01$) values for prenuclear content words for duration, intensity and one type of measurement for F0. In other words, speakers acoustically highlight the pitch accented word, but also content words that precede the pitch accented word. Therefore, in sentence-final noun phrases, all content words were eliminated from data collection.

**Pausing**

Potential data included only tokens that had no pauses greater than 150 ms either before or after each token, nor any pause between the two words. Anticipated pauses create preboundary lengthening, which increases vowel duration on the final stressed syllable of a phrase (Aylett & Turk, 2004; Cole, Mo, & Baek, 2010; Oller, 1973; Price et al., 1991; Umeda, 1975; Veilleux, Shattuck-Hufnagel, & Brugos, 2006; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992; Yoon, Cole, & Hasegawa-Johnson, 2007). Phrase final durational lengthening has also been observed for Mandarin Chinese speakers speaking in English (Rosenberg & Hirschberg, 2010), and also for Mandarin Chinese speakers speaking Chinese (Shen, 1990). This restriction on anticipated pauses also eliminates sentence-final constituents.

Unanticipated pauses likewise cause prosodic phenomena similar to anticipated pauses that cause pitch resets in a new sentence. De Pijper and Sanderman (1994) also found strong relationships among syntactic and prosodic boundaries, measurable acoustic cues, and perceptual judgments of prosodic boundaries. They found that pitch resets accompanied both anticipated and unanticipated preceding pauses, and these pitch resets occurred throughout the sentence regardless of syntax.
Both phenomena of unanticipated pauses causing preboundary lengthening (on duration) and pitch resets (on F0) were observed during data collection for the study. Part of the sentence, “It’s within the university’s right to monitor what their students are doing on their network,” is shown in Figure 13, which was excerpted from the audio data used in this study. The speaker pauses before she completes her sentence: “…are doing on that… on their… on their network”. When we look at the two instances of “their” in this example, the length of the vowel before the pause (marked in tier 1 as “1,447”) is three times longer (105 ms) than the vowel in “their” with no following pause (36 ms).

![Figure 13: The effect of pausing on F0 and duration](image)

The three instances of the function word “on” illustrate how pausing resets a new intonational phrase. The first pitch measurement (occurring in continuous speech with no pause) on the function word “on” is 166.58 Hz, the second pitch measurement (preceded by a pause) is
189.12 Hz, and, preceded by a longer pause, the third measurement is 190.76 Hz.

As both the literature and this example from data collection show, pausing causes a preceding word to lengthen and also resets the pitch to a new intonational phrase. Therefore, function or prosodic words on either side of an unanticipated pause were excluded.

**Single-syllable words**

The data will be limited to single-syllable content words and single-syllable function words for reasons I discuss below.

**Trading relationships**

Plag et al. (2011) studied acoustic cues in multisyllabic words that included three levels of lexical stress: primary, secondary, and no stress. In particular, they were interested in describing the acoustic effects of the location of primary and secondary stress in the target words (e.g., INdicate and CALCulate vs. indiCAtion and calcuLAtion). They also studied the effects of the presence or absence of pitch accent. They found that F0, intensity, and spectral balance measurements were affected by whether the stress was left or right prominent in a multisyllabic word. That is, a cue might be involved in producing left-prominent stress but not right-prominent stress. They found that duration was involved in stress production regardless of stress position in the word.

Likewise, Umeda (1975) measured vowels from 20 minutes of read-aloud speech from three speakers and found that the average duration of stressed vowels in polysyllabic content words were shorter than the duration of vowels in monosyllabic content words. Moreover, the
duration of vowels in function words was shorter than the stressed vowels in either polysyllabic or monosyllabic content words.

Reducing the eligible data to single syllable content and function words increases the possibility that all of the cues that are associated with stress will occur on the single syllable. Limiting tokens to single-syllable words eliminates any chance of trading relationships in multisyllabic words that were observed in the above study.

**Effect of adjacent stress on function words**

Although it is widely assumed that most function words categorically tend to be reduced, Lavoie (2000) found that for some speakers of English, function word reduction depended on whether the adjacent syllable was stressed or unstressed. In a study of 5 Native English Speakers, three participants categorically reduced the function word “for.” However, two speakers reduced “for” when it occurred adjacent to a stressed syllable, as in the prompt, “SAY bouTIQUE for me.” However, “for” was stressed when it occurred adjacent to an unstressed syllable, as in “SAY BOOty FOR me.” In the prompt “SAY BOOty for me,” “SAY” and the first syllable of “BOOty” obligatorily receive stress. The second syllable of “BOOty” is obligatorily unstressed. Stressing “for” in this phrase is optional and does not change the meaning of the sentence. Lavoie’s study suggests that, when phonologically and semantically allowable, some native English speakers may prefer to use alternating strong-weak-strong-weak stress patterns rather than use multiple adjacent weak syllables (i.e., strong-weak-weak-weak). Again, in the context of the present study, a two-syllable prosodic word that allows the option of a stressed function word complicates the relationship between the function and content words. Considering only single-syllable words eliminates this issue.
Summary of constraints on data

Table 8 illustrates how all of these constraints determine what type of text is included in and eliminated from potential data analysis. The sentence, “A lot of times, the teacher just speaking in front of the class, and the students are sitting in the lecture hall,” yields the text, “just speaking in front of,” from which to extract potential tokens, as shown in Table 8. In this case, one of two tokens is potential data: “in front” or “front of.” The question of which prosodic token to choose will be explained in the next section.

Table 8: Eligibility of data

<table>
<thead>
<tr>
<th>Paratone (eliminated)</th>
<th>IP-initial constituent (eliminated)</th>
<th>Potential data</th>
<th>IP-final constituent (eliminated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A lot of times,</td>
<td>the teacher</td>
<td>just speaking in front of</td>
<td>the class,</td>
</tr>
<tr>
<td>and the students</td>
<td>are sitting in</td>
<td></td>
<td>the lecture hall.</td>
</tr>
</tbody>
</table>

Data collection

For each of the seven test items, the sound file was searched in Praat (Boersma & Weenink, 2014) to locate prosodic tokens. The location of tokens was marked in a TextGrid in Praat. In cases where the transition between consonants and vowels was clearly visible in the waveform, the boundaries of vowels were marked at the onset and offset of periodic wave activity. In cases where the transition between consonants and vowels was less clear, additional cues were used, such as appearance of formants or antiformants, transitions between formants (for example, from a liquid to vowel), and abrupt intensity changes due to occlusion.

An automated Praat script was used to measure and record duration, peak F0, and mean
intensity of the two vowels. Per Praat’s recommended pitch settings, male subjects’ pitch range was set from 75 to 300 Hz and female subjects’ pitch range was set from 100 to 500 Hz. Approximately 70 potential tokens were eliminated because: 1. Praat was unable to detect a pitch value, 2. measured F0 values were unrealistically low due to creaky voice or, 3. measured F0 values were unrealistically high due to aspiration from an adjacent segment, which were frequently preceding fricatives.

During the token location phase, four permutations of word sequences emerged that contained potential tokens. The most common sequences of words in a prosodic token were Function-Content (e.g., “the class”) and Content-Function (e.g., “get the”). Of these two token word sequences, 162 Function-Content tokens and 408 Content-Function tokens were identified. The remaining potential two-word tokens had to be extracted from a 3-word phrase with the sequence of Function-Content-Function (e.g., “in front of”) or Content-Function-Content. So, to more evenly balance the frequencies of the two sequences, the 3-word phrases were converted into Function-Content sequences. Ultimately, from the 210 OEPT test items that were examined, a total of 752 tokens were located and measured (see Table 9).

### Table 9: Token type and total

<table>
<thead>
<tr>
<th>Token type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function-Content</td>
<td>344</td>
</tr>
<tr>
<td>Content-Function</td>
<td>408</td>
</tr>
<tr>
<td>Total tokens</td>
<td>752</td>
</tr>
</tbody>
</table>

Table 10 shows the number of tokens collected for each examinee. Examinee
identification numbers are the examinee’s OEPT score followed by examinee number. So, examinee 45-2 scored 45 on the OEPT and is the second examinee of 10. A few individuals who scored 35 on the OEPT had a low number of viable tokens. This was because of frequent pausing, which disqualifies any adjacent word from inclusion as a potential token. The 752 tokens are more concentrated toward the middle and top of the OEPT scale.

Table 10: Frequency counts by OEPT level and examinee

<table>
<thead>
<tr>
<th>OEPT level 35</th>
<th>OEPT level 45</th>
<th>OEPT level 55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examinee ID</td>
<td>N of tokens</td>
<td>Examinee ID</td>
</tr>
<tr>
<td>35-1</td>
<td>15</td>
<td>45-1</td>
</tr>
<tr>
<td>35-2</td>
<td>39</td>
<td>45-2</td>
</tr>
<tr>
<td>35-3</td>
<td>24</td>
<td>45-3</td>
</tr>
<tr>
<td>35-4</td>
<td>12</td>
<td>45-4</td>
</tr>
<tr>
<td>35-5</td>
<td>6</td>
<td>45-5</td>
</tr>
<tr>
<td>35-6</td>
<td>11</td>
<td>45-6</td>
</tr>
<tr>
<td>35-7</td>
<td>8</td>
<td>45-7</td>
</tr>
<tr>
<td>35-8</td>
<td>34</td>
<td>45-8</td>
</tr>
<tr>
<td>35-9</td>
<td>12</td>
<td>45-9</td>
</tr>
<tr>
<td>35-10</td>
<td>26</td>
<td>45-10</td>
</tr>
<tr>
<td>Total</td>
<td>187</td>
<td>Total</td>
</tr>
</tbody>
</table>

For each word, raw values of duration, peak F0, and mean intensity were converted into a durational ratio (duration of Content / duration of Function), F0 difference (F0 of Content – F0
of Function), and intensity difference (intensity of Content – intensity of Function). All raw measurements were then normalized as z-scores. Therefore, associated with each token are 3 quantitative values that contribute to the prosodic distinction between the two components of that token. Measurements of vowel quality, analysis of inherent vowel length, and the use of human raters were all excluded from the study, as will be explained below.

**Non-native English speakers’ use of vowel quality**

Vowel quality can affect the perception of lexical stress, (Fear, Cutler, & Butterfield, 1995), but vowel quality measurements were not considered in this data analysis. While L1 speakers of Mandarin Chinese do perceive vowel quality as a cue of English lexical stress (Zhang & Francis, 2010), they do not reliably employ vowel quality to signal lexical stress. Zhang, Nissen, and Francis (2008) found that L1 speakers of Mandarin Chinese consistently used duration, intensity, and F0 to signal English lexical stress, but they found that vowel reduction was inconsistently used as a cue of lexical stress. Zhang et al. compared 7 pairs of stress-based minimal pairs (e.g., OBject and obJECT) that were spoken by 10 native English speakers and 10 native speakers of Mandarin Chinese. The 10 native speakers of Mandarin Chinese had spent 3 to 4 years in the United States prior to participating in the study. Acoustic measurements showed that native speakers of Mandarin Chinese did not distinguish stress with vowel reduction in three of the seven minimal pairs and “[t]he only syllable [out of 14] in which both American and Mandarin speakers appear to show a similar degree and quality of vowel reduction is the syllable -ject [in] subject,” although all tokens in the study were deemed to have expected lexical stress (Zhang et al., 2008, p. 4507). The study method focuses participants’ attention on correct stress production on minimal pairs in a short carrier sentence, but Mandarin
speakers did not use vowel quality as a marker of lexical stress. However, American and Mandarin speakers did use the acoustic cues of duration, F0, and intensity in a similar way. We can therefore infer that, in the present study, vowel quality would be an unreliable indicator of stress in natural extemporaneous speech among L1 Chinese speakers.

Flege, Bohn, and Jang (1997) investigated the relationship between second language (L2) English speakers’ production and intelligibility of four English vowels and subjects’ familiarity with English, which ranged from about 5 1/2 years to one year of residency in the US. Accuracy of vowel production was determined by perceptual judgments by human raters and also by measuring formants for the English vowels /i, ɪ, ɛ, and æ/ that were produced by twenty L1 speakers of Mandarin Chinese. While Mandarin and English share similar high-front vowels, /i/, the remaining three vowels, /ɪ, ɛ, and æ/, are not found in standard Mandarin. As expected, the results showed that correct vowel perception of human raters differed according to subjects’ experience with English. Low English-familiarity speakers produced less accurately identified vowels than higher familiarity speakers; /i/: 80% vs. 84%, /ɪ/: 83% vs. 90%, /ɛ/: 60% vs. 63%, and /æ/: 58% vs. 77%. Acoustically measured vowel formant values confirmed those perceptual differences. Likewise, high English-familiarity speakers distinguished vowel formants between /i/ and /ɪ/ and between /ɛ/ and /æ/ while the low familiarity speakers did not. Flege et al.’s study demonstrates that L1 Chinese speakers’ English proficiency plays a role in their ability to control formants of some English vowels. Since this study intends to measure prosodic prominence across a range of oral English proficiency levels, and includes low proficiency L2 English speakers, vowel quality would likely be an unreliable measurement.

Chen, Robb, Gilbert, and Lerman (2001) showed similar results in terms of NNESs’ difficulty in differentiating vowel formants. Chen et al. measured vowel formants of eleven
English monophthongs produced by 40 native English speakers and 40 NNESs who spoke Mandarin Chinese as a native language. The authors found that L1 speakers of Mandarin used a significantly smaller vowel space than American English speakers, and that vowel formants of L1 speakers of Mandarin were produced with less acoustic diversity, which meant that vowels became physically crowded and resulted in more acoustic overlap between neighboring vowels. Findings from Zhang et al. (2008) support findings from both Flege et al. (1997) and Chen et al. (2001). Zhang et al. (2008) concluded that the Mandarin speakers’ difficulty to reduce vowel quality in unstressed syllables seemed to be related to their difficulty to produce the same full-quality vowel in stressed contexts.

An acoustic study by Kondo (2009) measured English lexical stress production by L1 Japanese speakers. She looked at minimal pairs that are differentiated by stress assignment (e.g., DEcrease and deCREASE) and measured the acoustic cues that the speakers used. The experiment used a carrier sentence (“I’ll say ‘target word’, now”) that elicited prosodic focus on the target word. Kondo found that L1 Japanese speakers use duration, F0 and intensity to differentiate stressed and unstressed vowels in English. However, despite the target word occurring in prosodic focus, participants did not use vowel quality to differentiate between stressed or unstressed vowels.

Finally, lexical acquisition may play a role in NNES’s production of vowel quality. For example, producing the full versus reduced quality of the initial vowel in “able” and “ability” appears to be a question of whether a speaker has achieved sufficient segmental control to reduce /eɪ/ to /ə/, but this assumption may not be accurate. Flege and Bohn (1989) measured four pairs of morphologically related words (e.g., able and ability, apply and application) that included a vowel that was stressed in one word and unstressed in the other. Flege and Bohn gathered data
(perceptual, acoustic and physical-articulatory) from 6 native Spanish speakers who spoke English as a second language. The researchers concluded that the model that NNESs use for vowel reduction in lexical stress may not be a morphological and phonological process that applies generally to multisyllabic words that undergo a stress shift. They hypothesized, rather, that NNESs learn individual pronunciations of related words like “able” and “ability” rather than applying a morpho-phonological rule that governs vowel reduction that accompanies a suffix like “–ity” that produces a stress shift. Zhang, et al.’s (2008) findings that vowel reduction was not systematic and seemed to vary with vowel category or the lexical item, supported Flege and Bohn’s hypothesis. If this model applies to L2 English learners generally, then conducting a study that is specifically designed to determine how L1 Chinese speakers use vowel quality (and other acoustic cues) to represent stress would be appropriate. The present study, however, is attempting to describe whether L1 Chinese speakers use unambiguous acoustic cues to represent prosodic prominence. And based on this literature of L2 speakers’ use of vowel quality in English, this study will not include vowel quality as a cue to prosodic prominence.

**Inherent vowel length**

Crosslinguistically, because of the mechanics of the articulatory system, it has been shown that low vowels are inherently longer than high vowels (Beckman, 1986). And in English, it has been shown that tense vowels are inherently longer than lax vowels ([i] vs. [ɪ] and [e] vs. [ɛ]) (Ladefoged & Johnson, 2010). For the same reasons that were listed above, at all L2 proficiency levels and especially at lower and intermediate levels, examinees do not reliably produce a distinction between lax and tense vowels. For example, according to any standard American dictionary, *good, get, tell,* and *give* should be pronounced with lax vowels. And
referring to both the literature reviewed above and also specifically to this data set, these words are very frequently pronounced as /gu[w]/, /ge[t]/, /te[l]/, and /giv/, respectively, as tense vowels. Therefore, no analysis will be made to contrast tense and lax vowels in this data set.

**Exclusion of human rating of prosodic prominence**

There are two common ways to measure stress and prosody. The more common of the two is to delineate vowels of the speech to be measured and collect discrete acoustic data such as duration, F0, intensity, vowel quality, etc. Another method uses human raters to judge prosody, as we have seen above in perceptual prominence studies. This second model may be considered advantageous as human raters simultaneously consider all acoustic cues, but this model includes the drawback of human subjectivity and necessitates multiple raters to establish reliability. In addition, the design of the study itself presents additional issues because of the choice to measure two-word tokens.

In the context of this study, there are two options for human raters to judge prosody. One method to rate prosody is to listen to a full sentence in which a token occurs and ask raters to judge the prosody on the token. An example of this task from the dataset might be to listen to an audio clip of the following sentence, “So, I will list my office hours,” rate only the prosody on “will list,” and ignore the prosody in the surrounding sentence. But, asking raters to mentally separate the prosody on the token from the surrounding prosody of the sentence is unrealistic since prosodic appropriateness necessarily depends on the surrounding context.

Another method might be to digitally excise the audio “will list” from the sentence and ask raters to assign a score. The duration of this particular example token, “will list,” is shorter than 1/2 second, which means that the rater would have to judge prosodic prominence from a
very short audio clip that lacks prosodic continuity and meaningful context. Recall that this study measures prosody at a phonological phrase level where the stress differences between function and content words are likely subconscious in nature, which contrasts with a learner’s relationship with lexical stress. English language learners typically learn lexical stress as a concomitant part of the correct pronunciation of the word (e.g., SYL.la.ble rather than syl.LA.ble).

Although content words tend to be more prominent than function words, that relationship may change depending on information structure and syntactic position. Umeda (1975) commented that durational differences between function and content words are non-binary, and “the change from extremely functional to very contentful seems to be continuous” (p. 437). This is unlike judging lexical stress, as we have seen with SUBject and subJECT, which is lexically determined and more easily identifiable as correct or incorrect.

Segmental approximations, grammar irregularities, and other features that may distract from the task of judging prosody are likely to accompany prosodic prominence in this dataset, which would make an already challenging task of identifying a prosodic relationship between function and content words more difficult. Studies that have used human raters to judge prosody (excluding those measuring lexical stress) have generally taken a holistic approach that includes paragraph-level information. Cunningham (2012) describes the problem of attempting to separate prosody from concomitant features of language in the minds of human raters:

Another reason that prosody has often been avoided in terms of direct measurement for assessment purposes is because it is often so entangled with other segmental features and syntactic information that it is hard to isolate. As prosodic features often act as the scaffolding for the more salient and individual segments related to phonological and syntactical systems, it is a challenge to
separate the scaffolding from the building it is holding up. Prosodic features do not reveal their true properties and functions easily when they are taken apart and broken down into quasi-artificial components such as “rhythm”, “intonation”, [or] “stress” (p. 4).

As the duration of the tokens of prosodic information is unreasonably short to judge in isolation, and as the appropriateness of the prosodic prominence on the tokens relies on surrounding syntactic and prosodic performance (the influence of which influences judgment), the judgment of human raters was excluded from consideration for this study.
CHAPTER 5. RESULTS

Boxplots for the independent variables of duration, F0, and intensity are shown in Figure 14. The observations of the three OEPT groups appear to overlap nearly completely for F0 and intensity, but duration appears to show some difference in median values despite overlapping in the second and third quartiles. This same trend is shown in the density plots in Figure 15 (Red = OEPT score 35, yellow = OEPT score 45, and green = OEPT score 55 for all plots and histograms). Recall that the OEPT uses a 6-point scale, and the three OEPT groups in the present study represent non-adjacent English proficiency groups on that scale. Those who score 55 on the OEPT have relatively high proficiency and are certified to serve as ITAs. Those who score 45 have borderline proficiency and are allowed to teach provided they are concurrently enrolled in the OEPP course (ENGLISH 620). And those who score 35 have restricted proficiency and typically need more than one semester of the OEPP course until they are certified to serve as an ITA.
Figure 14: Boxplots for Duration, F0, and Intensity

Figure 15: Density plots for duration, F0, and intensity. Red = OEPT 35, Yellow = OEPT 45, Green = OEPT 55.
QQ plots for duration and F0 indicated that the data may not be normal, as shown in Figure 16.

![QQ Plot for Duration z-scores](image1)

![QQ Plot for F0 z-scores](image2)

![QQ Plot for Intensity z-scores](image3)

**Figure 16: QQ-plots for duration, F0, and intensity.**

Shapiro-Wilk tests were performed on the data for each of the independent variables to test for normality. The tests indicate that the data are not normally distributed for any of the independent variables: Duration (W = 0.84733, p-value < 2.2e-16), fundamental frequency (W = 0.74951, p-value < 2.2e-16), and intensity (W = 0.99151, p-value = 0.00026). Based on the non-parametric nature of the data, an ordered logistic regression was performed, and duration was the only significant variable in the model (z = 3.74, df = 751, p = 0.00184) (See Table 11).
Table 11: Ordered Logistic Regression

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>z value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.14601</td>
<td>0.08767</td>
<td>13.072</td>
<td>2.00e-16***</td>
</tr>
<tr>
<td>Duration</td>
<td>0.39805</td>
<td>0.10644</td>
<td>3.74</td>
<td>0.000184***</td>
</tr>
<tr>
<td>F0</td>
<td>-0.03713</td>
<td>0.08874</td>
<td>-0.418</td>
<td>0.675625</td>
</tr>
<tr>
<td>Intensity</td>
<td>0.03954</td>
<td>0.08818</td>
<td>0.448</td>
<td>0.653913</td>
</tr>
</tbody>
</table>

The large p-values for F0 (p = 0.676) and intensity (p = 0.654) in Table 11 are expected based on the shape of the data in the boxplots in Figure 14 and density plots in Figure 15. There appears to be no difference in how the three groups used F0 or intensity despite large differences in English proficiency. The finding for F0 may indicate that this study’s methodological strategy to heavily restrict eligible data was successful in filtering out new information that would be expected to be prosodically marked with pitch accent and therefore with higher F0 values, as we saw in Wennerstrom (1998). This strategy was adopted in order to target the measurement of a function-content word prosodic relationship and limit the influence of discourse prosody such as the given-new contract.

According to the ordered regression, the three groups are indeed using duration differently, so a Spearman's rank correlation was performed to determine the association between OEPT scores and durational differences between content and function words (S = 54437000, p-value = 1.212e-10, r = 0.23). Again, based on the overlap of the data from the boxplots and density plots for duration in Figure 14 and Figure 15, a weak correlation of 0.23 between duration and OEPT scores is unsurprising.
The answer to the first research question, “Do L1 Chinese NNESs acoustically distinguish content and function words?” is yes, but with some qualification up until this point. The data of the independent variables are distributed on a normalized scale, which tell us the relative distance among observations, but do not describe the difference in physical acoustic terms. To help answer this question in more meaningful measurements, we need to do two things: the first is to reconnect z-scores to acoustic reference points to know how frequently participants are using durational cues to distinguish content and function words. Second, with regards to durational ratio, which was the duration variable measured in this study, we need to know specifically how duration distinguishes stressed and unstressed syllables. Sluijter (1995) found that for native speakers of American English, the average ratio of durational contrast between adjacent unstressed and stressed syllables was 1:1.2.

Setting that threshold ratio of 1:1.2 for the durational data in the present study produces tokens that are above and below that threshold. Figure 17 shows that the three OEPT groups are clearly stratified by percentages of function-content word tokens that meet or exceed that 1:1.2 native-speaker threshold.
Furthermore, if we consider the average durational stress ratio of native English speakers is unaccompanied by a variety of distracting speech variables, then 1:1.2 is sufficiently accessible for a listener to distinguish stress differences. However, in the context of multiple linguistic factors of NNES production that account for a proficiency score (intelligibility, coherence, comprehensibility, correctness of grammar, appropriate complexity of syntax, appropriateness and breadth of vocabulary usage, segmental pronunciation, fluency measures such as unexpected pausing and speech rate, and prosody), it’s reasonable to suspect that a rater might not perceive a function-content word token as acoustically distinct with a durational ratio of 1:1.2 in the presence of many other perceptual variables. If we, therefore, reset the ratio at a higher durational contrast at 1:1.5, which is more prosodically prominent than 1:1.2 and more
likely to withstand the effects of the list of distracters above, we see in Figure 18 that the previous trend holds.

![Percentage of Prosodic Contrast Caused by Duration (ratio = 1:1.5)](image)

Figure 18: Percentage of Acoustically Distinct Tokens that Meet or Exceed a 1:1.5 Ratio

Based on the percentages in Figure 18, we can say that all three groups of participants produce prosodic prominence of content over function words, although clearly they do so at different rates, which leads to the second research question: “How are acoustic measurements of prosodic prominence associated with overall English language proficiency of L1 Chinese NNESs?” Regardless of which durational ratio we use as “prosodically prominent,” higher proficiency speakers (55) distinguish function-content tokens more often than intermediate proficiency speakers (45), who do so more often than low proficiency speakers (35).
This study contrasts with previous studies that have measured function and content words in non-native speech in terms of the prosodic territory that it targets. If we consider the prosodic hierarchy that includes prosodic focus, sentence stress in the intonational phrase, prosodic word stress, secondary word stress and prosodically unrecognized function words, previous studies have targeted prosody at the top and bottom of this hierarchy. Wennerstrom (1998, 2000) investigated how oral English proficiency levels were associated with the differences between higher-level discourse prosody (i.e. prosodic focus and nuclear pitch accents) on new and contrasting information versus given information and function words. In her (1998) study, Wennerstrom found that all participants used F0 to prosodically contrast new information content words from function words, but she found no difference in how participants of different proficiency levels did so. In her (2000) study, she found that higher proficiency L2 English speakers used higher F0 on new and contrasting information compared with given information and function words, while the lower proficiency L2 English speakers did not make these prosodic distinctions with F0. In both cases, high-information content words were contrasted with function words. While Wennerstrom’s studies are certainly relevant to the present study, the present study adds to this description of non-native English that L1 Chinese NNESs acoustically contrast between prosodic word stress and function words without the influence of sentence stress or prosodic focus. In addition, the rates of this contrast clearly distinguished proficiency groups.

To address the 3rd research question regarding which acoustic measurements L1 Chinese NNESs use to distinguish content and function words, we’ve seen how the three groups use duration, and we’ve seen that there is no statistical relationship between OEPT scores and p-values for F0 and intensity in the ordered logistic regression. It is nevertheless informative to
look in more detail at how each variable was used by the three OEPT levels. Once again, to establish a meaningful description of whether and how participants used F0 and intensity to create prosodic prominence between content and function words, z-scores were converted back to original values in hertz and decibels. Following that conversion, we need to know what constitutes noticeably different prosodic values from F0 and intensity.

Cullen & Long (1986) and Houtsma & Smurzynski (1990) investigated the “just noticeable difference” for computer-generated pure and complex tones, and both of these studies found that F0 changes of less than 1 Hz were enough for listeners to detect. But computer-generated pure tones and even complex tones are quite different from investigating the just noticeable difference of an F0 of 150 Hz that is contained within a co-occurring vowel that includes multiple formants and thousands of surrounding frequencies. Klatt (1973) conducted experiments on just noticeable differences of F0 values that were contained within computer-generated synthesized vowels, some of which were steady state and some of which changed formants as F0 also changed. In Klatt’s study, he also found that participants detected a F0 change of less than 1 Hz when a vowel was at a steady state, but he also found that when vowels shifted between formants, between a glide and a vowel, for example, then the lowest F0 changes that participants could detect ranged from 2 Hz to 4 Hz, depending on the steepness of the formant shift. This study will adopt 4 Hz as a threshold for causing acoustic prosodic difference.

Before discussing a threshold for intensity, it should first be restated that intensity alone has not been a reliable indicator of stress (Fry, 1955; Ladefoged, Draper, & Whitteridge, 1958; Lieberman, 1960), and Turk and Sawusch’s (1996) experiment found that the effects of intensity are additive, not independent of other cues duration and frequency, and not independently
reliable as an indicator of stress. That said, for the purposes of describing the prosodic production of participants, an intensity value to establish a prosodic difference must be adopted.

In read-aloud speech, Fry (1955) measured peak intensity and duration of five minimal pairs (e.g., diGEST vs. DIgest) to determine which cue was a stronger indicator of syllable stress. In Fry’s measurements of intensity as an indicator of stress, the highest mean value was 3 dB. Flanagan (1957) found just noticeable difference limens for intensity at about ±3dB within synthetically produced vowels. Nishinuma et al. (1983) studied CV syllables ([tatata…]) and found that a ±2.5dB change in intensity can affect stress judgments in durationally short syllables. Based on the results of these experiments of intensity and stress, I will adopt 3dB as a threshold for intensity to indicate a prosodic difference.

It would be dubious to call these thresholds of 4Hz and 3dB as “prosodically prominent,” so “acoustically different” will be used here in the absence of human judgments to confirm prosodic prominence.
Figure 19: Percentage of Acoustically Distinct Tokens that Meet or Exceed thresholds for F0 and Intensity

The percentages for F0 and intensity meeting or exceeding the 4Hz and 3dB thresholds are shown in Figure 19. F0 appears to be negatively associated with OEPT scores. The lowest OEPT group more frequently used higher F0 to distinguish function-content tokens than the other two OEPT groups. Recall that none of the data in this study were collected from sentence-final positions, because prosodically prominent pitch accents occur in sentence-final positions 66% of the time in extemporaneous English speech, according to a study by Calhoun (2010). This study also excluded data from sentence-initial positions because sentences often start with elevated F0, due to the physiological initial high air pressure from the lungs (Wennerstrom, 1998, 2001). These two most likely positions for high F0 values leave the sentence-medial position an unlikely place to encounter high F0 values, as the OEPT 35 group did 50% of the time. If, indeed, raters reacted negatively to a high number of nuclear pitch accents from the
OEPT 35 group, it may be because of the frequent use of the sentence-medial position to locate a nuclear pitch accent. The possibility also exists that participants used more than one pitch accent per sentence, which this study is unable to speak to. This finding also suggests that the methodological strategy to eliminate nuclear pitch accents by eliminating prosodic boundaries from data collection is only effective if the prosodic instincts of the participants conform to the prosodic expectations of native English speakers, which seems was not the case with the OEPT 35 group.

A negative association between F0 measurements and OEPT scores is interesting when we consider the prosodic treatment of Chinese function words. It is well known that Mandarin Chinese lexical words minimally carry one of four phonemic lexical tones: High level tone (T1), rising tone (T2), falling then rising tone (T3), and falling tone (T4). Not as well known is a neutral tone (T0, also called atonic or stressless tone) that is associated with Chinese function words and grammatical morphemes, among other constructions (B. Yang, 2015). When a lexical word that contains a lexical tone is compounded with a following neutral tone, tone sandhi, or contextual tonal shift, applies to the neutral tone (Yiya Chen & Xu, 2006). Chen and Xu found that the neutral tone has no specific contour as do the four lexical tones, but the neutral tone does have a mid-level F0 target range. They found that the shape of the neutral tone is highly mutable and subordinate to the adjacent lexical tone. For example, a high level lexical tone (T1) combined with a neutral tone (T0) (e.g., “dōng xi, 东西, ‘things’”) results in the neutral tone falling from the preceding high T1 to mid-level, whereas a falling lexical tone (T4) combined with a neutral tone (e.g., “dà yì, 大意, ‘careless’”) results in the neutral tone slightly rising from the preceding low off-glide of T4 to mid-level (M. Y. Chen, 2000, pp. 286–287). Complicated sandhi rules also apply to lexical compounds, for example, when a falling-then-rising tone (T3)
occurs adjacent to another T3, then the first tone changes to rising (T2) (B. Yang, 2015). M.Y. Chen (2000) has demonstrated in a variety of Chinese dialects that tone sandhi is the domain of Minimal Rhythmic Unit, which applies to both the neutral tone sandhi that cliticizes to the lexical tone, as well as sandhi rules for compounding lexical tones.

These Minimal Rhythmic Units may have some relationship to how examinees in the present study approach English function words. Lower scoring examinees more frequently used F0 to distinguish lexical content words from function words compared to the higher scoring examinees. The occasions when examinees used F0 to contrast English content and function words may be related to parsing them as a Minimal Rhythmic Unit and adapting the phonology of Chinese stress assignment to English text.

Intensity does not reveal any remarkable information, which is consistent with studies of stress reviewed above. Intensity differences on content words are low across all three groups, which may confirm that participants used intensity as additive to other prosodic cues of duration and F0.

**Cognitive Psychology and the Structure Building Framework**

In Chapter 3, I described prosody as a structural glue that holds discourse together for the listener/comprehender, and signals relationships among ideas in discourse. This claim was based on a review of studies from the perspective of linguistics.

Parallel to the linguistic investigation of language comprehension, cognitive psychology offers a complimentary explanation of how prosody facilitates the comprehension of spoken and written text. For over 30 years, Morton Ann Gernsbacher, along with many collaborators, has developed a general theory of language comprehension called the Structure Building Framework.
According to the Structure Building Framework, the goal of comprehension is to build coherent mental representations or structures. These structures represent clauses, sentences, paragraphs, passages, and other meaningful units. To build these structures, first, comprehenders lay foundations for their mental structures (Carreiras, Gernsbacher, & Villa, 1995; Gernsbacher & Hargreaves, 1992, 1988; Gernsbacher, Hargreaves, & Beeman, 1989). Then comprehenders develop their mental structures by mapping on information, when that incoming information coheres or relates to the previous information (Gernsbacher, 1996; Gernsbacher & Givón, 1995; Gernsbacher & Robertson, 1992; Haenggi, Gernsbacher, & Bolliger, 1993; Haenggi, Kintsch, & Gernsbacher, 1995). However, if the incoming information is less coherent, comprehenders employ a different process: They shift and initiate a new substructure (Foertsch & Gernsbacher, 1994; Gernsbacher, 1985). (Gernsbacher, 1997a, p. 85)

The framework above comes from the point of view of the comprehender, but when people are interacting verbally, both comprehender and speaker act simultaneously. The speaker formulates ideas and chooses language that expresses those ideas. Ideally, that language is clear, cohesive and coherent, but spontaneous language is sometimes incomplete, unclear or deficient in some respect. As a result, the speaker may choose to add information, amend information, or correct information.

While the speaker is engaged in these activities, the comprehender decodes the speech signal into meaningful units, i.e., words, phrases, and clauses, and builds a mental structure of ideas based on these meaningful units. New, amended, or corrected information may cause the comprehender to revise the mental structure. It is among a steady stream of information from a speaker that a comprehender has to make decisions to sort the more relevant information from
the less relevant information so that she can build mental structures. The comprehender makes use of overt markers clues, such as anaphoric and cataphoric signals from the speaker to link characters and actions together from clause to clause.

Cataphoric devices include such overt markers as vocally stressing a word in spoken discourse, or boldfacing a word in printed text. Presumably, speakers and writers mark certain concepts with cataphoric devices because those concepts will play a key role in the text or discourse. Thus, it would behoove listeners and readers if those key concepts had a privileged status in their mental structures. (Gernsbacher, 1997a, p. 92)

Just as speakers highlight relevant information with stress and focus, they also acoustically minimize grammatical function words by giving them no stress. Both speaker and comprehender use prosody that highlights more relevant information and acoustically minimizes less relevant information.

While [Gernsbacher] believe[s] that most people can appreciate that we need a mechanism that enhances relevant or related information, [she has] suggested that a mechanism that suppresses inappropriate or irrelevant information is perhaps even more crucial to the goal of comprehension, or in the words of the Structure Building Framework, the goal of building coherent mental representations. We need a mechanism of suppression because whenever we comprehend language, we experience various types of interference. (Gernsbacher, 1997a, p. 86)

To understand how the Structure Building Framework integrates with linguistic systems, I find it helpful to visualize the cataphoric suppression and enhancement, “vocally stressing” or “boldfacing” structure-building mechanisms. The suppression and enhancement model is shown in Figure 20, and will be integrated with linguistic systems below in Figure 22. The model
suggests that listeners give extra attention to the information that is prosodically focused by the speaker because it is likely to play a key role in the discourse, and by contrast, give different attention to the information that is suppressed because it is normally predictable (in the case of function words) or retrievable in discourse context (in the case of given information).

For example, given information is either retrievable from the prior discourse or predictable in the context of the discourse, which means that comprehenders only need be reminded of this information and it need not be prosodically flagged for special attention. Likewise, function words belong to closed classes, their numbers are relatively few, and they are normally predictable in English syntactic structure such that all function words may be completely deleted from a paragraph with little resulting difference in meaning\(^\text{11}\). The same

---

\(^{11}\) Example, given information either retrievable prior discourse predictable context discourse, means comprehenders only need reminded information need not prosodically flagged special attention. Likewise, function words belong closed classes, numbers relatively few, normally predictable English syntactic structure function words completely deleted paragraph little resulting difference meaning.
could not be said for the predictability of content words\textsuperscript{12}.

In Gernsbacher’s suppression model, alternate meanings of a lexical word (e.g., bug) can be suppressed by placing the word into a context that enhances one meaning (insect) while suppressing other meanings (to annoy, to listen covertly with an audio device). Suppression refers to multiple mechanisms, including the phonological focus on new and contrastive information and lack of prosodic recognition on grammatical function words.

Gernsbacher’s suppression and enhancement model mirrors the role of prosodic stress and focus. A schematic of three stress levels is presented in Figure 21, which is simplified from the phonological review in Chapter 3.

![Figure 21: Simplified schematic of three stress levels](image)

It is further helpful to visually align the complementary relationships among multiple complex systems of prosodic phonology, prosodic acoustics, and discourse analysis from the field of linguistics, and the Structure Building Framework from cognitive psychology. Prior to connecting these levels of stress to other systems in linguistics and cognitive psychology, recall that:

1. **Phonologically:** Prosodic words minimally receive lexical stress while function words remain prosodically unrecognized by English phonology (Selkirk, 2004). Focus refers

\textsuperscript{12} The could not be for the of.
to prosodic features that rise above lexical stress or sentence stress (Beckman et al., 2006; Bolinger, 1986; Cruttenden, 1997; Hayes, 1995; Ladd, 1996; Pierrehumbert, 1980)

2. Acoustically: Content words and stressed syllables align with higher measurements of acoustic cues, while function words align with lower measurements of acoustic cues. Content words with prosodic focus have higher measurements of acoustic cues than non-focused content words (Breen et al., 2010).

3. In discourse analysis: Prosodic focus aligns with new and contrastive information (Birch & Clifton, 1995; Bock & Mazzella, 1983; Nooteboom & Kruyt, 1987; Terken & Nooteboom, 1987). Given information aligns with intermediate measurements of acoustic cues that are greater than grammatical information (function words) but acoustically lesser than new or contrastive information (Cole, Mo, & Hasegawa-Johnson, 2010).

In Figure 22, in tier 1, the three simplified stress levels are illustrated. These stress levels align with word class in tier 2. Function words phonologically receive no stress (unless they are promoted to FOCUS), lexical content words receive a minimum of lexical stress, and when prosodic focus is employed under expected circumstances, it lands on the stressed syllables of the lexical content words of the topic to receive focus (Selkirk, 2004). In tier 3, word class aligns with measured acoustic values, which is the area of investigation in this study. Function words align with low acoustic values while content words align with acoustic values that are higher than measurements for function words. Acoustic values of words or syntactic constituents that have prosodic focus are higher than values for nearby lexical content words. These measurable acoustic cues of prosody align with the cataphoric message to the comprehender in tier 4. Low
acoustic values send a message to the comprehender that identifies function words (which provide structural information and are often predictable as these are closed classes), intermediate acoustic values suggest that the comprehender give some attention to lexical content words (for example, to retrieve given information from the discourse, or to maintain attention because the information may return to discourse), and high acoustic values suggest to the comprehender to pay extra attention to this text, which is likely to play a key role in the discourse (Gernsbacher, 1997a). Based on the prosodic acoustic values in tier 3, the expected information structure in tier 5 is that grammatical information aligns with word class in tier 2. Function words align with low acoustic values, given or contextually retrievable information aligns with content words that have intermediate acoustic values, and new or contrastive information aligns with content words that have higher acoustic values. Finally in tier 6, the suppression and enhancement model of the Structure Building Framework aligns with all of the above levels in the form of prosodic stress levels and information structure.\(^{13}\)

\(^{13}\) Although lexical selection, lexical repetition, and discourse strategy play roles in the suppression and enhancement model, only the prosodic role of suppression and enhancement will be discussed here.
### Description of the System

#### Elements of the System

<table>
<thead>
<tr>
<th>Description</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three levels of phonological stress</td>
<td>Focus</td>
</tr>
<tr>
<td>Stress levels align with</td>
<td>No stress</td>
</tr>
<tr>
<td>Word class</td>
<td>Lexical stress</td>
</tr>
<tr>
<td>Aligns with</td>
<td></td>
</tr>
<tr>
<td>Measurements of acoustic cues of prosody</td>
<td>Lower acoustic values</td>
</tr>
<tr>
<td>Prosody implicitly suggests</td>
<td>Intermediate acoustic values</td>
</tr>
<tr>
<td></td>
<td>Higher acoustic values</td>
</tr>
<tr>
<td>Cataphoric prosodic message to comprehender</td>
<td>This information is structural and predictable</td>
</tr>
<tr>
<td></td>
<td>Give some attention to this information</td>
</tr>
<tr>
<td></td>
<td>Give extra attention to this information</td>
</tr>
<tr>
<td>Expected information structure suggested by prosodic structure</td>
<td>Grammatical information</td>
</tr>
<tr>
<td></td>
<td>Given information</td>
</tr>
<tr>
<td></td>
<td>New or contrastive information</td>
</tr>
<tr>
<td>Mechanism of suppression and enhancement</td>
<td>Suppression</td>
</tr>
<tr>
<td></td>
<td>Enhancement</td>
</tr>
</tbody>
</table>

Figure 22: Schematic of relationship of multiple systems in linguistics and cognitive psychology
The multi-system model in Figure 22 depicts an idealized expectation that a speaker will signal among these linguistic and cognitive systems. In light of the results of the present study, we can modify the model to illustrate what may be occurring with the relationships among the systems of language and cognition depicted in Figure 22.

The groups of participants in this study did not reliably acoustically distinguish function and content words, but participants with higher OEPT scores did so more frequently than participants with lower OEPT scores. Figure 23 illustrates a suggestion of what may be occurring in the multi-system model when physical acoustic cues do not distinguish word class and activate suppression and enhancement mechanisms. In this model, we have to start from the area of investigation in this study, which is tier 3. In those instances when participants did not acoustically distinguish (in tier 3) content and function words (which for OEPT 35 participants, was 61% of the tokens at a 1:1.5 durational ratio), for those two-word tokens, there was no clear cataphoric prosodic message to the comprehender (in tier 4). The cataphoric message suggests how much attention to devote to the text in the token, and there was therefore no clear distinction (in tier 5) between grammatical information expressed by function words and meaningful information expressed by content words (in tier 2). As a result, information structure was not reflected in those two-word tokens. It appears, then that in the cases where two-word tokens are undifferentiated by the speaker by using acoustic cues, the first two columns under tier 1 of “no stress” and “lexical stress” may be acoustically and perceptually collapsed into a single column. Finally, because of this lack of prosodic distinction, the crucial element of suppression within the Structure Building Framework in tier 6 is likely missing from these tokens.
### Description of the system

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levels of phonological stress</td>
</tr>
<tr>
<td>Lexical stress</td>
</tr>
<tr>
<td>Focus</td>
</tr>
</tbody>
</table>

1. Stress levels align with

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function words and lexical content words</td>
</tr>
<tr>
<td>Lexical content words</td>
</tr>
</tbody>
</table>

2. Aligns with

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undifferentiated acoustic values</td>
</tr>
<tr>
<td>Higher acoustic values</td>
</tr>
</tbody>
</table>

3. Prosody implicitly suggests

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undifferentiated, give equal attention to all information</td>
</tr>
<tr>
<td>Give extra attention to this information</td>
</tr>
</tbody>
</table>

4. Aligns with

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical information undifferentiated from given information</td>
</tr>
<tr>
<td>New or contrastive information</td>
</tr>
</tbody>
</table>

5. Information structure suggested by prosodic structure

<table>
<thead>
<tr>
<th>Elements of the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undifferentiated, unsuppressed</td>
</tr>
<tr>
<td>Enhancement</td>
</tr>
</tbody>
</table>

6. Mechanism of suppression and enhancement

---

Figure 23: Schematic of collapsed columns based on lack of acoustic distinction between function and content words
It is reasonable to infer that within the limited domain of that two-word token, lack of acoustic distinction between function and content words likely diminishes a comprehender’s ability to build coherent mental representations in the Structure Building Framework. Terken and Nooteboom’s (1987) study that was discussed in Chapter 3 is relevant here when we consider their findings in light of Gernsbacher’s Structure Building Framework. Terken and Nooteboom observed that in the context of an existing discourse, comprehenders reacted faster to new information that was prosodically focused (i.e., expected prosody), and comprehenders reacted slower to new information that was prosodically unfocused (i.e., unexpected prosody). Terken and Nooteboom hypothesized that comprehenders use prosodic cues before textual cues to build a mental map of discourse. Prosodically focused nouns are expected to align with new information that coheres to the existing discourse. Likewise, prosodically unfocused nouns are expected to align with given information that was previously introduced into the discourse.

However, because prosodically unfocused new information was processed by the comprehenders more slowly than prosodically focused new information, Terken and Nooteboom hypothesized that this delay was caused because comprehenders created a mental map of the discourse based on the unfocused prosodic cue that signals the comprehender to map this information onto previously mentioned discourse. Then, finding no given information to match this information to, the comprehender revises the mental map to restructure the new information. As Gernsbacher describes the process, “…if the incoming information is less coherent, comprehenders employ a different process: They shift and initiate a new substructure” (Gernsbacher, 1997a, p. 85). In the context of this study of non-native speech in which various forms of “interference” is expected, the processes of initiating a foundation for the mental structure or shifting to a new substructure are very important. “[A] large body of converging data demonstrated that comprehenders slow
down when they are presumably laying mental foundations for their mental structures” (Gernsbacher, 1997b, p. 267). This tendency of comprehenders to slow down when laying foundations for mental structures applies to multiple modes of comprehension: reading text, viewing picture stories, as well as listening to speech (Gernsbacher, 1997b). The suppression and enhancement model in the Structure Building Framework matches the anecdotal classroom experiences that initially led me down this investigatory path. In some conversations with my students, and to different degrees that varied from one individual to another, I personally found that my cognitive task was dependent on how much prosodic guidance I was getting from the speaker to enable me to separate more relevant information from less relevant information. I recognize now that when grammatical or given information is not prosodically suppressed, the task of separating grammatical information from new or contrastive information has to be performed on a more conscious, textual basis rather than an unconscious, prosodic basis.

**Future research**

This study intentionally targeted the measurement of content and function words in an unfocused context. The methodology excluded content that was prosodically focused, such as new or contrasted information, which opens up avenues for future research. Wennerstrom’s (1998) study, described in Chapter 1, like the present study, also measured prosody of function and content words of native Chinese speaking teaching assistants from a range of English proficiency levels. However, Wennerstrom’s methodology measured only new information and intentionally excluded given information. Future research that would offer a more detailed description of prosodic production of native Chinese speakers of ESL would combine methodologies of Wennerstrom’s 1998 study and the present study. Like the present study, pairs
of single syllable function and content words may be identified, and from those tokens, a
discourse analysis performed to identify new and given information, and prosodic measurements
taken to compare prosodic treatment of function and content pairs in both new and given
discourse positions. Once again, as we saw in Figure 22, these prosodic measurements may be
used to indicate to what degree prosody aids predictability of new and given, and content and
functional elements of discourse, which allows inferences to be made about participants’ use of
suppression and enhancement mechanisms to build mental structures.

A second area of future research may concentrate on how much a comprehender’s ability
to build coherent mental representations is diminished by a lack of suppression and enhancement
mechanisms. That is, does comprehension decrease with the decrease in prosodic distinctions
between function and content words? Hahn’s (2004) study described in chapter 2 studied the
comprehension effects caused by differences in sentence stresses. She varied three versions of an
academic lecture given by a non-native English speaking ITA that included appropriate sentence
stress on new information in condition 1, inappropriate sentence stresses on given information in
condition 2, and did not include sentence stress on any information in condition 3.
Comprehension test results of 90 undergraduate participants showed that listener comprehension
was best in condition 1, and similarly weak in both conditions 2 and 3. A second area of future
research that I would pursue would be based on Hahn’s 2004 study and the present study, which
would create three similar conditions of an academic lecture and test participants’
comprehension. Similar to Hahn’s study, the three conditions can be achieved via digital
manipulation that makes content words more prosodically prominent than function words
(condition 1), content words less prosodically prominent than function words (condition 2), and
no prosodic difference between content and function words. The same study may also repeat
Hahn’s methodology to first confirm the results of her 2004 study, and second compare the effects of inappropriate focus marking on given information to the effects of inappropriate prosodic prominence on function words. Depending on the results of this study, inferences again may be made about participants’ use of suppression and enhancement mechanisms to build mental structures.
REFERENCES


https://doi.org/10.1080/01690965.2010.504378

https://doi.org/10.1080/01690965.2010.491682

https://doi.org/10.3758/BF03214418


Ginther, A. (2016, December 5). Personal communication.


<table>
<thead>
<tr>
<th>Level</th>
<th>General Proficiency Level is… Performance is…</th>
<th>Possible Response Characteristics at this Level: the symbol / means “and/or”</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td><strong>Excellent and Consistent across items. Majority of items are 60s. Minimal listener effort required to adjust to accent. Displays of lexico-syntactic sophistication.</strong> Speaker is at ease and confident fulfilling task, elaborating clear message, using accurate English.</td>
<td>Fluent and coherent on all tasks; normal discourse-level pausing and hesitation expected; pace may vary. Good to excellent prosody. Minor, rare pronunciation difficulties. Effective use of wide range of vocab with idiomatic expression. Syntax and tense complex and varied; appropriate use of modality. Grammar and vocab errors minor, rare, and do not affect meaning.</td>
</tr>
<tr>
<td>55</td>
<td><strong>More than Adequate. Little listener effort required to adjust to accent/prosody/intonation. Consistently comprehensible, coherent.</strong> Strong skills across items. Wide range of vocab and syntactic structures. Speaker exerts little noticeable effort in elaborating a clear message to fulfill task.</td>
<td>May have strong accent, sound substitutions, some syllable stress errors, but listener adjusts quickly. Good prosody. May have pausing/hesitation, variable speech rate, but generally fluent. May have some distracting sounding or flat intonation. Variety and range of vocab with idiomatic expressions; minor usage errors. Complex and varied syntax; syntax errors minor; some minor bound morphology errors may be systematic but do not affect meaning.</td>
</tr>
<tr>
<td>50</td>
<td><strong>Adequate for classroom without support.</strong> Mostly 50s with some 55s or very few 45s. Acceptable amount of listener effort required to adjust to accent/prosody/intonation. Consistently intelligible and comprehensible.** Speaker may exert a little noticeable effort but, despite minor errors of grammar/vocab/stress/fluency, message is adequately coherent, with correct information, some sophistication.</td>
<td>May have sound substitutions/stress errors/L1 characteristics in pronunciation and intonation, but listener adjusts. Speaker articulates sounds and words with minimal effort. Prosody is acceptable. May have hesitations/pauses/repetitions between spurts of fluency; possibly slow delivery speed; no consistent disfluencies such as searches, restarts. Adequate vocab with some idiomatic expressions; may have vocab usage, article and preposition errors that do not derail message. May have bound morphology, minor grammar errors that don’t derail. Range and complexity of syntax sufficient to fulfill task successfully. May lose path but recover successfully, self correct.</td>
</tr>
<tr>
<td>Score</td>
<td>Category</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>45</td>
<td>Borderline - Inconsistent – Minimally adequate for classroom with support. Mix of 45s and 50s. Very few, if any, 40s. Tolerable listener effort required to adjust. Generally intelligible, comprehensible. Strengths &amp; weaknesses across characteristics or items. Message is generally coherent (but occasionally not), but may require more than a little effort for speaker to compose. Or message may be clear and expressed fluently, but language use is somewhat simplistic.</td>
<td>May have pronunciation/intonation/stress issues that require listener adjustments, but still intelligible. Articulation may require noticeably small effort on speaker’s part. Prosody may be borderline. May have a few disfluencies: restarts/repeats/searches/consistent hesitation or slow speed, but still comprehensible Responses fulfill task adequately, but may be limited, simplistic; may digress Vocab range may be somewhat limited/more than a few vocab errors but most do not affect meaning May have systematic grammar errors, but most do not affect meaning Good listening comprehension: accurate information Profiles vary: May sound confident/fluent, but deficient in accuracy or complexity, or speak too fast to follow easily; May have good pronunciation/ prosody, but deficient in fluency and accuracy</td>
</tr>
<tr>
<td>40</td>
<td>Limited - Not ready for the classroom. Mostly 40s with some 45s or a few 35s, if any. Generally able to address prompts and complete responses. Consistent listener effort may be necessary. Message may be simplistic/unfocussed/incomplete/incorrect. May struggle somewhat to build sentences/argument or to articulate sounds.</td>
<td>May have pronunciation/articulation issues that affect intelligibility; may be occasionally unintelligible. May have disfluencies such as restarts/repeats/searching/choppy delivery; may compose slowly, with effort; prosody/stress may cause problems Argument may be dependent on repetition; may digress or run on in an unorganized manner; information may be occasionally incorrect Limited vocab or grammar may affect ability to respond in a sophisticated manner; vocab usage errors may be common; may lack idiomaticity May have a variety of noticeable, systematic grammar/word order errors</td>
</tr>
<tr>
<td>35</td>
<td>Restricted - May need more than 1 semester of support. Mix of 35s and 40s. Listening requires considerable effort or patience. May be unintelligible or incoherent more than occasionally OR have marked deficiencies in at least 3 areas: fluency, articulation/ pronunciation, prosody, comprehensibility/ coherence, listening comprehension, accuracy/range of vocab or grammar.</td>
<td>May be difficult to understand/follow/listen to more than occasionally, due to poor articulation/ pronunciation or intonation/stress/rhythm issues May have marked disfluencies: false starts/unproductive repetition/long pauses/struggling/topics abandoned, or syllabic/choppy/halting delivery Sentences/argument may be difficult to follow/often incomprehensible or incoherent. May have: restricted range of vocab/syntax,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>many vocab errors, a variety of grammar errors that may be difficult to classify, major syntax errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information on listening items may be incorrect or incomplete</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>