

© 2010 Yoonsook Mo

PROSODY PRODUCTION AND PERCEPTION
WITH CONVERSATIONAL SPEECH

BY
YOONSOOK MO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Jennifer S. Cole, Chair
Associate Professor Mark Hasegawa-Johnson
Associate Professor Chilin Shih
Assistant Professor Ryan K. Shosted
Assistant Professor Duane Watson

Abstract

Speech utterances are more than the linear concatenation of individual phonemes or words. They are organized by prosodic structures comprising phonological units of different sizes (e.g., syllable, foot, word, and phrase) and the prominence relations among them. As the linguistic structure of spoken languages, prosody serves an important function in speech communication: prosodic phrasing groups words into pragmatically and semantically coherent small chunks and prosodic prominence encodes discourse-level status and rhythmic structure of a word within a phrase. In speech communication, speakers shape spoken language through the modulation of multiple acoustic parameters related to tempo, pitch, loudness, vocal effort, and strength of articulation in order to signal prosodic structures. Prosody is therefore a major source of phonetic variation in speech and in particular, elements at the edges of prosodic units and those assigned prominence are phonetically distinct from similar elements in different prosodic contexts. From a listener's standpoint, one must attend to this phonetic variation, and, more specifically, to acoustic variation in order to reconstruct the prosodic context and to understand the meaning of an utterance as intended by the speaker.

This thesis concerns the communication of prosody in everyday speech, with a primary focus on acoustic variation arising from prosodic context and its interaction with other factors including syntactic, semantic, pragmatic structure, and word predictability. More specifically, the goal of the thesis is to understand prosody in terms of the mechanisms of speech production, to identify the cues that guide listeners' in-

terpretation of prosodic structure, and to establish statistical models of the acoustic encoding of prosody, in everyday conversation.

This thesis introduces a new method of prosody annotation, called *Rapid Prosody Transcription (RPT)*, which provides reliable and consistent prosody annotations, is comparable to highly trained, expert listeners', and better approximates prosody perception in every speech communication. In RPT, prosody annotation is obtained through the real-time tasks of prosody transcription by a large group of "ordinary" (untrained, non-expert, and thus naïve in terms of the phonetics and phonology of prosody annotation) listeners, on the basis of auditory impression only.

On the basis of sets of prosodically-annotated speech excerpts extracted from the Buckeye Corpus of spontaneous conversational speech of American English through RPT, the rest of this thesis reports findings regarding prosody production and perception in everyday speech communication. With various statistical methods including non-parametric Spearman's correlation and multiple linear regression analysis, this thesis demonstrates that given the invariance in a set of acoustic parameters, prosodic prominence is signaled through a combination of multiple acoustic parameters from which each speaker may choose any subset as their selection, and prosodic boundary is cued by a single acoustic parameter relating to speech tempo, suggesting that the production mechanisms of prosodic prominence are underlyingly different from those of boundary production. Such difference in the acoustic encoding of prosodic features is further evidenced in the temporal structure of subsyllabic components of monosyllabic CVC words.

Evaluating the role of speakers and listeners in the communication of prosody, this thesis reveals speaker-dependent variability in the acoustic encoding of prosody and listener-dependent variability in the decoding of prosody. These findings suggest that given the multiplicity of acoustic parameters, speakers choose any subset as their selection in order to signal prosodic structures and depending on the nature of the

acoustic parameters, listeners attend to acoustic variation in particular forms (raw vs. normalized), within particular comparison domains (syntagmatic vs. paradigmatic) in order to correctly interpret prosodic structures. Confirming that acoustic variation in the speech signal guides a listener to perceive prosodic structure as produced by a speaker, this work further show that other factors (syntactic and semantic expectation and word predictability in discourse and in the language) interplay with acoustic variation in prosody perception.

This research contributes both to large scale prosody research by introducing a new and innovative method for prosody annotation and to our understanding of the communication of prosody in everyday speech, by highlighting variation in the acoustic encoding of prosody depending on prosodic features as well as on speaker identity and the nature of prosody as an interface phenomenon relating various factors including phonology, syntax, discourse structure, and lexical entropy. Taking into account speaker-dependent variability in the implementation of prosody and a large role of listeners in the normalization of such variability, this thesis proposes the best models of the acoustic encoding of prosody in everyday speech communication.

To my parents, Seungwon, and Yoonsoo.

Acknowledgments

Through this opportunity, I wish that I could thank everybody who made my long journey possible and successful with consistent and patient care and encouragement although it can never be possible to express my gratitude in any satisfactory manner.

First and foremost, I would like to thank my advisor Jennifer Cole for her constant support and guidance throughout my doctoral research at UIUC. I have benefited from discussions and countless meetings with her, and her insightful comments and constructive criticisms. Her insightful comments and constructive criticisms have trained me as a linguist with excellent preparation and enthusiasm. It was my great privilege to have her as my advisor and to learn from her.

Another very important person in my research is Mark Hasegawa-Johnson. He has also guided me with his comments from a different point of view and constant support and encouragement. It was my great pleasure to interact with him through numerous individual and group meetings. I felt greatly fortunate to have opportunities to work with him and to have him in my committee.

I am also greatly grateful to the members of my committee, Chilin Shih, Ryan Shosted, and Duane Watson for their comments and support. They have contributed to my research in various aspects. Working with Chilin Shih and Ryan Shosted, I have been trained to work with physiological data including articulatory as well as aerodynamic data. Duane Watson has provided me with opportunities to understand human language processing with the eyetracking system.

I have much benefited from scholars at UIUC outside my thesis committee. In

particular, I would like to thank Chin-Woo Kim and Wayne Dickerson in Linguistics. Wayne Dickerson motivated me to pursue my research on the spoken languages while I was in the master program in Teaching English as a Second Language. In addition to his great service as the departmental head, Chin-Woo Kim has always been like a parent of all students in Linguistics, remembering students birthdays, giving them warm and kind words, and inviting all the students to his war house every year. James Yoon is another scholar who I would thank. He has always supported me in many aspects. Through meetings and discussions with him, I have been able to complete my research in partial fulfillment of the requirements for the Master degree in Linguistics.

Without my friends and colleagues in various disciplines, I could not complete my doctoral study. I would like to thank the members of the interdisciplinary Prosody-ASR group: Margaret Fleck, Sarah Borys, Jui-Ting Huang, Xiaodan Zhuang, Arthur Kantor, Tim Marht, and Tae-Jin Yoon. My friends have been a great support for me in my life at UIUC: Heeyoun Cho, Heejin Kim, Eunah Kim, Su-Youn Yoon, Kyung-Young Kim, Chaeyoon Park, Soorim Noh, Junghyun Min, Meryl Garrison, and many other friends.

Finally but most importantly, my family made it possible what I have completed. My parents have been my inspiration for my life and their devotion and endless love has stood me up at every moment. In particular, my mom who is now in bed in the hospital and is missing her daughter has been the source of my energy. My parents-in-law have also been the greatest support for me. I also wish to thank my brother and his family and brothers-in-law and their families for their support as well as for filling my empty space for parents and caring for them. I have two men who I have been debted the most: my loving husband, Yoonsoo Pang and my beloved Seungwon Pang. They have not only supported me but also been greatly patient in every single aspect at every single moment. Especially, I owe the most to my son, Seungwon.

Since he was 21 month olds, he has spent more time in libraries than at home. He slept on the office desk or on the chairs. For three years, he lived without his dad who was at Berkeley, CA. He always missed his grandparents, uncles, nephews and nieces. Without his understanding and patience, his mom could not imagine her success in her doctoral program.

My doctoral research was funded in part by the School of Literature, Culture, and Linguistics Dissertation Completion Fellowship (2009-2010) and by the National Science Foundation grants funded to Mark Hasegawa-Johnson and Jennifer Cole, and Mark Hasegawa-Johnson, Jennifer Cole, and Chilin Shih.

Table of Contents

	Page
List of Tables	xiii
List of Figures	xv
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Defining prosodic prominence and boundary	3
1.3 Challenges for obtaining prosody annotation in spontaneous conversational speech	6
Chapter 2 Rapid Prosody Transcription	10
2.1 Buckeye corpus	10
2.2 Materials and Subjects	12
2.3 Procedure of Rapid Prosody Transcription (RPT)	14
2.3.1 Experiment 1	14
2.3.2 Experiment 2	18
2.3.3 Results	18
2.3.4 Testing the reliability of RPT	22
2.4 Summary	28
Chapter 3 The Distribution of Prosodic Scores by Phone Identity	30
3.1 Introduction	30
3.2 Distribution of prosody scores by phone	31
3.2.1 Distribution of prominence scores by phone	31
3.2.2 The distribution of boundary scores	34
3.3 Conclusion and Discussion	34
Chapter 4 An Acoustic Investigation of Prosodic Prominence	37
4.1 Introduction	37
4.1.1 Fundamental frequency (F_0)	38
4.1.2 Other acoustic correlates of prominence	41
4.2 Acoustic analyses	46
4.2.1 Measurements	46
4.2.2 Normalization of the acoustic measures	48
4.3 Results	49

4.3.1	How closely is each acoustic measure related with the perception of prosodic prominence by ordinary listeners?	49
4.3.2	How much do different acoustic measures contribute to the ordinary listeners' perception of prosodic prominence?	58
4.4	Summary and Discussion	63
4.5	Conclusion	69
Chapter 5 An Acoustic Investigation of Prosodic Boundary		71
5.1	Introduction	71
5.1.1	Fundamental frequency (F_0)	72
5.2	Acoustic analyses	82
5.2.1	Measurements	83
5.3	Results	83
5.3.1	How closely is each acoustic measure related with the perception of prosodic boundary?	84
5.3.2	To what extent do different acoustic measures contribute to listeners' perception of prosodic boundary?	92
5.4	Summary and Discussion	95
5.5	Conclusion	100
Chapter 6 Prosodic Effects on the Temporal Structure of Monosyllabic CVC Words		102
6.1	Introduction	102
6.2	Acoustic analyses	105
6.2.1	Acoustic measurements	105
6.3	Results	106
6.3.1	How does prosodic prominence influence the temporal characteristics of monosyllabic CVC words?	107
6.3.2	How does prosodic boundary influence the temporal characteristics of the monosyllabic CVC words?	110
6.3.3	How do prosodic features influence the internal temporal structure of the monosyllabic CVC words in spontaneous conversational speech of American English?	115
6.4	Summary and Discussion	115
6.5	Conclusion	122
Chapter 7 How do Ordinary Listeners Perceive Prosodic Features? Syntagmatic VS. Paradigmatic Comparison		124
7.1	Introduction	124
7.2	Normalization	128
7.2.1	Paradigmatic normalization	128
7.2.2	Syntagmatic normalization	130
7.3	Results	131
7.3.1	Prosodic prominence	131
7.3.2	Prosodic boundary	135

7.4	Summary and Discussion	139
7.5	Conclusion	142
Chapter 8 How do Ordinary Speakers Signal Prosodic Features?		
Speaker-Dependent VS. Speaker-Independent Models		143
8.1	Introduction	143
8.2	Acoustic measurements	145
8.3	Results	145
8.3.1	Prosodic prominence	146
8.3.2	Prosodic boundary	150
8.4	What are the best acoustic models of prosody?	154
8.4.1	Best acoustic models of prosodic prominence and boundary . .	154
8.5	Summary and Discussion	162
8.6	Conclusion	165
Chapter 9 What Other Factors Affect Ordinary Listeners' Perception of Prosody?		
		166
9.1	Introduction	166
9.2	Does syntactic information fully predict ordinary listeners' perception of prosody?	170
9.2.1	Data collection	170
9.2.2	Results	170
9.2.3	Discussion	176
9.3	How does word predictability relate to the perception of prosodic prominence?	179
9.3.1	Analysis	179
9.3.2	Discussion	184
9.4	Conclusion	187
Chapter 10 Conclusions		
		188
10.1	Summary of findings	188
10.2	Conclusion	194
Appendix A Buckeye Perception Projects		
		197
A.1	Subject PowerPoint® and Transcript examples	197
A.1.1	A sample Microsoft PowerPoint® page (chunk boundaries) constructed for the sound file presentation for each transcriber . .	197
A.1.2	A sample Microsoft PowerPoint® page (prominence) constructed for the sound file presentation for each transcriber	198
A.1.3	A sample corresponding orthographic word transcript	199
A.2	Consent and instruction forms	205
A.2.1	5-minute instruction for subjects	205
A.2.2	Subject consent form	206
A.2.3	Subject language survey form	207

Appendix B Programming Scripts	208
B.1 Sample scripts for Praat TM version 5.2.03	208
B.2 Sample scripts for Python TM version 2.6	211
References	223

List of Tables

Table	Page
1.1 Comparisons of the agreement rate of prosody annotation following the ToBI transcription convention reported in different studies	8
2.1 Comparisons of the Fleiss' kappa scores and the corresponding z -normalized scores in Experiment 1 and 2	24
3.1 The distribution of lexically stressed vowels and lexically stressed word final vowels	32
4.1 Summary of the Spearman's non-parametric correlation analyses between P-scores and acoustic measures of each vowel (I). Spearman's ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.	51
4.2 Summary of the Spearman's non-parametric correlation analyses between P-scores and acoustic measures of each vowel (II). Spearman's ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.	52
5.1 Summary of the Spearman's non-parametric correlation analyses between B-scores and acoustic measures of each vowel (I). Spearman's ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.	85
5.2 Summary of the Spearman's non-parametric correlation analyses between B-scores and acoustic measures of each vowel (II). Spearman's ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.	86

9.1	Confusion matrix of prosodic feature assignment based upon listeners' expectation and prosodic features assigned by speakers in the speech signal: prosodic prominence and prosodic boundary	175
9.2	Spearman's non-parametric correlation and linear regression analyses of P-scores and log-frequency, and P-scores and word repetition (1, 2, 3, and 4 more), from words in three data sets: all long excerpts, long excerpts minus frequently reduced words, and long excerpts minus function words.	181

List of Figures

Figure	Page
1.1 A schematic representation of the prosodic hierarchy	4
2.1 Transcription scheme	16
2.2 The distribution of probabilistic prominence (P, solid line) scores and boundary (B, dotted line) scores for each word in a sample utterance from Speaker 26	20
2.3 The distribution of the intervals between prosodic prominences (dark grey) and prosodic boundaries (light grey) in Experiment 1 on the left and in Experiment 2 on the right	21
2.4 The boxplots of pairwise Cohen’s kappa scores between the agreed prosody transcription and the prosody transcription (prominence on the left and boundary on the right)	26
3.1 The distribution of the mean probabilistic P(rominence)–scores (solid) and B(oundary)–scores (oblique) of each phone	33
4.1 The distribution of the total variation (R^2) of the ordinary listeners’ perception of prosodic prominence (oblique bars) and prosodic boundary (dotted bars)	61
4.2 The distribution of the variation (R^2) in the ordinary listeners’ response to prosodic prominence predicted by stepwise multiple linear regressions of the acoustic measures and P–scores	62
5.1 The distribution of the variation (r^2) in the ordinary listeners’ response to prosodic boundary predicted by stepwise multiple linear regressions of the acoustic measures and B–scores	94
6.1 Scatterplot with regression lines between Prosody scores (P– and B–scores) and the raw durations (in seconds) of the monosyllabic CVC words	108
6.2 Scatterplot with regression lines between P–scores and the raw durations of the onset (—*—), nucleus (- - - Δ - - -), and coda (- \cdot - \diamond - \cdot -) of the monosyllabic CVC words	109

6.3	The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic prominence (oblique bars) and prosodic boundary (dotted bars), explained from the various durational measures of the monosyllabic CVC words	111
6.4	Scatterplot with regression lines between B-scores and the raw durations of the onset ($-*-$), nucleus ($- - -\Delta- - -$), and coda ($- \cdot - \diamond - \cdot -$) of the monosyllabic CVC words	113
6.5	Scatterplot with regression lines between P-scores and the ratio of the onset ($-*-$), nucleus ($- - -\Delta- - -$), and coda ($- \cdot - \diamond - \cdot -$) duration to syllable duration in the monosyllabic CVC words	116
6.6	Scatterplot with regression lines between B-scores and the ratio of the onset ($-*-$), nucleus ($- - -\Delta- - -$), and coda ($- \cdot - \diamond - \cdot -$) duration to syllable duration in the monosyllabic CVC words	117
6.7	Schematic representation of temporal structure of the monosyllabic CVC word: non-prominent word (P-score = 0, top) vs. prominent word (P-score = 1, bottom)	121
6.8	Schematic representation of temporal structure of the monosyllabic CVC word: phrase-medial word (B-score = 0, top) vs. phrase-final word (B-score = 1, bottom)	122
7.1	Distribution of the total variations (r^2) of the ordinary listeners' perception of prosodic prominence, explained from the acoustic measures: raw, paradigmatically ($z-$, $z(\log)-$, and $\gamma-$), and syntagmatically (window size: 3 adjacent vowels, 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words) normalized acoustic measures from left to right.	132
7.2	Contribution of each acoustic measure (duration, local $F_{0,max}$, overall intensity, subband intensities in 0-500, 500-1000, 1000-2000, and 2000-4000 Hz) to predicting ordinary listeners' response to prosodic prominence: with raw acoustic measures, z -normalized acoustic measures by phone, z -normalized log-transformed acoustic measures by phone, γ -normalized acoustic measures by phone, syntagmatically normalized acoustic measures over a dynamic window of 3 adjacent vowels, 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words, in order from left to right.	134
7.3	Distribution of the total variations (r^2) of the ordinary listeners' perception of prosodic boundary, explained from the acoustic measures: raw, paradigmatically ($z-$, $z(\log)-$, and $\gamma-$), and syntagmatically (window size: 5 adjacent vowels and 3 adjacent stressed vowels) normalized acoustic measures from left to right.	136

7.4	Contribution of each acoustic measure (duration, local $F_{0,max}$, overall intensity, subband intensities in 0–500, 500–1000, 1000–2000, and 2000–4000 Hz) to predicting ordinary listeners’ response to prosodic boundary: with raw acoustic measures, z –normalized acoustic measures by phone, z –normalized log–transformed acoustic measures by phone, gamma–normalized acoustic measures by phone, syntagmatically normalized acoustic measures over a dynamic window of 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words, in order from left to right.	138
8.1	The distribution of the total variation (r^2) of the ordinary listeners’ perception of prosodic prominence by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of P–scores with the acoustic measures	147
8.2	The distribution of the total variation (r^2) of the ordinary listeners’ perception of prosodic prominence by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of P–scores with the acoustic measures	149
8.3	The distribution of the total variation (r^2) of the ordinary listeners’ perception of prosodic boundary by speaker (blue) and by phone (red), obtained from multiple linear regression analyses of B–scores with the acoustic measures	152
8.4	The distribution of the total variation (r^2) of the ordinary listeners’ perception of prosodic boundary by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of B–scores with the acoustic measures	153
8.5	The distribution of the total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners	156
8.6	The distribution of the total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners	157
8.7	The distribution of the adjusted total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners	158
8.8	The distribution of the adjusted total variation (R^2) in the speaker-dependent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners	159
9.1	The distribution of probabilistic prominence scores (P–scores) of a word in a sample utterance from Speaker 26, with (solid line) and without (dotted line) hearing sound files	172
9.2	The distribution of probabilistic boundary scores (B–scores) of a word in a sample utterance from Speaker 26, with (solid line) and without (dotted line) hearing sound files	173

9.3	Boxplots of P-scores for words in “long” excerpts, grouped by repetition index. These plots include only words that occur with at least two instances within a discourse segment: (1) first mention, (2) second mention, (3) third mention, and (4) fourth or more mention.	183
9.4	The total variation (r^2) in ordinary listeners’ perception of prosodic prominence from multiple linear regression analyses on the left and each major correlate’s contribution to modeling listeners’ perception of prosodic prominence on the right	185

Chapter 1

Introduction

1.1 Introduction

Speech utterances are more than the linear concatenation of individual phonemes or words. They are organized by prosodic structures comprising phonological units of different sizes (e.g., syllable, foot, word, and phrase) and the prominence relations among them. Prosodic units at the phrase level group together sequences of adjacent words that cohere semantically. Within a prosodic phrase, one or more words may be assigned prominence as a phonological means of “highlighting” a word or a phrase that carries information marking the message as discourse-new or focused.

The transfer of information between speaker and listener is the main goal of speech communication. Beyond lexical meaning and the meaning conveyed through syntactic structure, speakers communicate pragmatic and discourse meaning in everyday speech through prosody. Prosodic structures are encoded in phonetic form through the modulation of pitch, loudness, tempo, vocal effort and strength of articulation. Prosody is therefore a major source of phonetic variation in speech, and elements at the edges of prosodic units and those assigned prominence are phonetically distinct from similar elements in different prosodic contexts. From a listener’s standpoint, one must attend to this phonetic variation, and, more specifically, to acoustic variation in order to reconstruct the prosodic context and to understand the meaning of an utterance as intended by the speaker.

The main objective of this thesis is to investigate the production and perception

of prosody in spontaneous, conversational speech of American English, as produced for natural communicative purposes, with a focus on acoustic variability arising from prosodic contexts. An innovation of this study is that it looks at prosody simultaneously from the perspectives of the speaker and listener, by examining the production correlates of prosody in acoustic form and at higher levels of linguistic organization, on the basis of the prosodic elements that ordinary listeners perceive in conversational speech. To achieve these goals, the following research questions will be addressed: (1) how does prosody influence the acoustic patterns of the speech signal?; (2) is there any speaker-dependent variability in the acoustic encoding of prosody?; (3) do listeners reliably identify prosodic prominence and boundary?; (4) do they attend to acoustic variation in the speech signal in prosody perception?; (5) what kinds and forms of acoustic parameters, individually or in combination, contribute to the listener's perception of prosody?; (6) what is the relationship between prosody and other linguistic factors including syntactic structure and word predictability (e.g., word token frequency and word repetition in a discourse)?

The research questions posed here have been examined based on prosody annotations obtained by rapid prosody transcription, RPT, by which prosodic features in spontaneous speech are identified by multiple "ordinary" (non-expert, untrained and naïve in terms of phonetics and phonology of prosody transcription) listeners during real-time tasks of auditory prosody transcription. In other words, this research investigates the phonetic nature of prosody in American English, as produced by ordinary speakers and as identified by ordinary listeners. By obtaining ordinary listeners' annotations of prosodic features in real time, the current study better approximates prosody perception in everyday communication than most prior studies in which only a few trained transcribers made linguistic judgments on the locations of prosodic features, following the ToBI (Tones and Break Indices) transcription convention. The present study is also motivated by several facts including the following: First, al-

though many prior studies have shown significant agreement rates of prosody annotation among trained labelers who had a substantial amount of training for prosody transcription (Dilley et al., 2006; Ostendorf et al., 1995; Pitrelli et al., 1994; Syrdal and McGory, 2000; Yoon et al., 2004), there still remains considerable disagreement among prosody transcriptions made by highly trained transcribers. Secondly, there is deviation in transcription methodology from prosody perception in everyday communication: listeners do not have multiple chances to hear speech as many times as they may need and do not have time for a careful, visual inspection of speech. Therefore, this research strives to be a more accurate approximation of prosody production and perception as it occurs in authentic, everyday conversations; it not only continues the investigation of the phonetic nature of prosody in American English in tradition with prior studies, but additionally looks at the phonetic variation arising from prosodic context as produced by ordinary speakers in spontaneous conversational speech and as perceived by ordinary listeners in the real time prosody perception tasks. Before examining the phonetic nature of prosody in everyday spontaneous speech, I provide a brief overview of the form and function of prosodic structures from prior work.

1.2 Defining prosodic prominence and boundary

Prosody is an aspect of phonological structure above the level of the individual phone (consonant or vowel). Prosodic structure comprises hierarchically organized domains, from the syllable up to the utterance level, as illustrated in Figure 1.1. Within each level one or more elements (e.g., syllables, feet, and words) may be assigned prominence. There are a number of competing proposals concerning the number of distinct prosodic domains, and the factors that determine the assignment of prosodic structure (e.g., Abercrombie, 1964; Jun, 1993; Ladd, 2008; Liberman and Pierrhumbert, 1984; Liberman and Prince, 1977; Nespor and Vogel, 1983, 1986; Pierrehumbert,

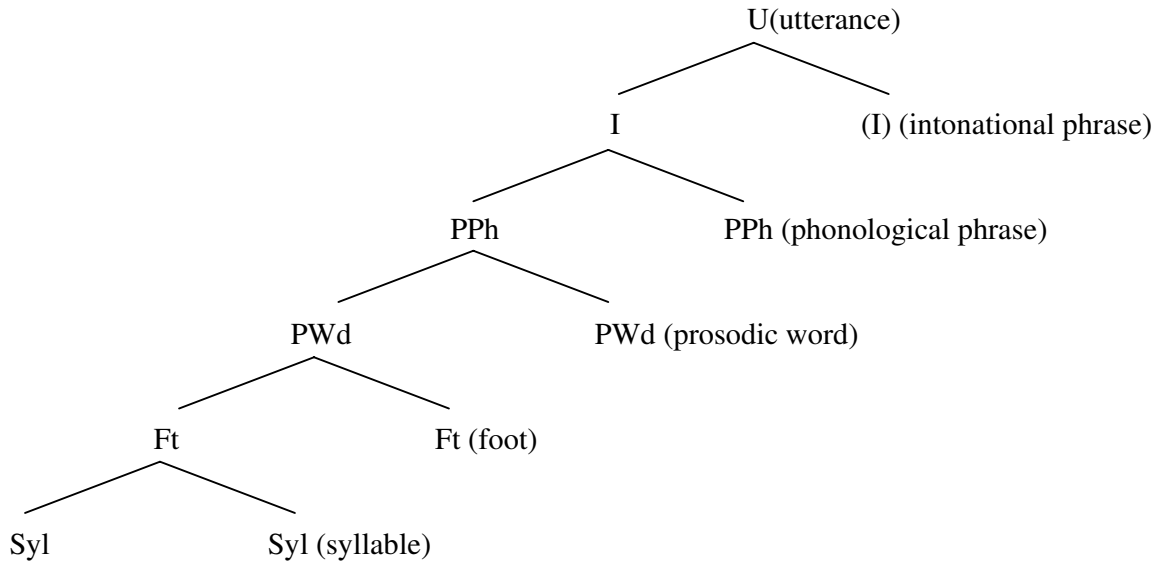


Figure 1.1: A schematic representation of the prosodic hierarchy

1980; Selkirk, 1986; Truckenbrodt, 1995). Common to all accounts of prosody is the recognition that prosodic structures at the phrase level are not a perfect reflection of syntactic structures.

Thus, Fox (2000) states that prosodic phrases are not constructed solely on the basis of morphosyntactic rules, but also by “independently motivated rules”. Although these independently motivated structures are closely related to morphosyntactic structures, prosodic phrases are not always isomorphic to morphosyntactic structures as attested by studies such as Gee and Grosjean (1983); Watson and Gibson (2004), among others. For example, Kang and Speer (2004) discuss the importance of prosodic structure in spoken language processing with a sentence containing a participle phrase which is possibly grouped together with the previous noun phrase as in (example 1.1a), or comprises its own prosodic phrase, separated from the previous NP as in (example 1.1b). With prosodic structure in (example 1.1a), the sentence is interpreted as *Aaron followed a poor guy who was drinking his soda*. With prosodic structure in (example 1.1b), on the other hand, the same sentence is interpreted as

Drinking his soda, Aaron followed a poor guy.

Aaron followed a poor guy drinking his soda.

- a. Aaron followed // a poor guy drinking his soda. (1.1)
- b. Aaron followed a poor guy // drinking his soda.

Prosodic features are also understood to convey information about the intentional and attentional structure of discourse (Hirschberg and Pierrehumbert, 1986; Pierrehumbert, 1980; Pierrehumbert and Hirschberg, 1990), and thus play a role in sentence or discourse processing (e.g., Arnold, 2008; Cutler et al., 1997; Dahan et al., 2002; Kjelgaard and Speer, 1999; Kraljic and Brennan, 2005; Nakatani, 1997; Terken and Nootboom, 1987). Cutler et al. (1997) reviews an extensive body of studies that show how the prosodic structures produced by speakers act as cues to guide the listener’s interpretation of syntactic and discourse structure. For instance, several studies (Carlson et al., 2001; Kang and Speer, 2002; Kjelgaard and Speer, 1999; Kraljic and Brennan, 2005, among others) provide evidence that prosodic structures aid a listener in parsing the syntactic structures of speech disambiguating syntactically ambiguous sentences. More recently, Dahan et al. (2002) and Arnold (2008) showed that accenting (as a mark of prosodic prominence) biases a listener’s comprehension of nouns-unaccented nouns are interpreted as anaphoric references.

This thesis follows prior work in considering prosodic structure as an aspect of phonological representation, comprising prosodic domains and prominence relations among the elements in a given domain. The following definitions are adopted here. *Prosodic boundaries* serve to demarcate chunks of speech that group together adjacent words that cohere semantically. *Prosodic prominence* serves to highlight a word or a phrase that carries important information relative to the communicative goals of the discourse. Prominence conveys the status of these words or phrases as focused or discourse-new.

Such prosodic structure is signaled through the phonetic implementation of various physiological as well as acoustic parameters as indicated by a large body of research. Bolinger (1958) first introduced the notion of pitch accent as a physical realization of stress and suggested that if a word is prominent in a sentence, then this prominence is realized as pitch accent. Lieberman (1967) explained that prominence results from positive subglottal pressure which in turn makes a prominent element louder, and claimed that prominence has two sources, namely stress which can be predicted by the stress cycle, and emphasis which cannot. Pierrehumbert (1980) and Liberman and Pierrhumbert (1984) define prominence in terms of the Strong-Weak patterning of phonological metrical structure, and consider the F_0 contours that mark pitch accent as one aspect of the phonetic expression of prominence. The current study examines prosody production and perception, focusing on acoustic variation arising from such prosodic structure.

1.3 Challenges for obtaining prosody annotation in spontaneous conversational speech

In order to study the form and function of prosody in spoken language, researchers need appropriate tools for marking those landmarks that encode the prosodic structures of speech. Most of the existing research relies on manual transcription of these prosodic landmarks based on auditory impression, sometimes aided by a visual analysis of the speech display. The lack of a unified, or even a widely accepted, transcription system has been a central problem in the study of prosody, impeding the development of prosody research. In 1991, in response to this dilemma, a group of researchers gathered and developed a unified system called the Tones and Break Indices (ToBI) transcription convention within the autosegmental-metrical framework of phonological theory. Central to this system are two tonal events: one associated with prosodic

boundaries (boundary tones), and the other associated with accented syllables (pitch accents). Two tonal targets (H and L) are used to express different types of pitch accents and phrasal tones either by themselves or in combination with each other. In addition, the depth and strength of a juncture between words is expressed by a break index (0 to 4). This system provides a transcription method that can be interpreted by various researchers in a consistent way, as well as enabling researchers to compare their results with those of other studies.

On the other hand, there are several drawbacks to employing this widely accepted transcription convention. First, prosody annotation following the ToBI transcription convention does not reflect a listener's perception of prosody in everyday communication. An ordinary listener does not have time or the linguistic knowledge to explicitly identify the types and location of prosodic boundaries and prominence in everyday speech communication. In addition to multiple opportunities for auditory playback in ToBI system, transcribers are highly trained, and are allowed to inspect the visual displays of speech before they make a final decision about the prosodic features contained therein. In most instances of everyday communication, however, a listener must not only recover prosodic structures based solely upon auditory impression, but she or he is not aided by the visual speech. Secondly, the monetary as well as time commitment of prosody annotation using the ToBI system is prohibitively high, as prosody annotation requires time for extensive transcriber training and for the transcription task itself. For example, in Pitrelli et al. (1994), participants were trained as follows:

Each transcriber was provided with a document describing the ToBI standard, and the ToBI training materials. The training materials contain a short tutorial explaining each of the labels in ToBI, along with recorded examples of transcribed utterances for listening at key points in the tutorial narrative. Interspersed in the tutorial are lists of untranscribed utterances

Agreement Rate	Pitrelli et al. (1994)	Grice et al. (1996)	Syrdal and McGory (2000)	Yoon et al. (2004)	Dilley et al. (2006)
Prominence labeling	81%	87%	91%	87%	87%
Boundary labeling	93%	N.A.	93%	90%	88%

Table 1.1: Comparisons of the agreement rate of prosody annotation following the ToBI transcription convention reported in different studies

similar to the examples, which the transcribers could use to practice the labels described up to that point in the text. Transcribers were encouraged to discuss these examples with others; however, the training materials are designed to be self-paced, so that the user need not have an expert on site.

Similarly, in Dilley et al. (2006), five undergraduate students were trained with the same manual as well as by computerized exercises. In addition, they received feedback from an expert labeler and had bi-weekly meetings with expert labelers where they transcribed two one-minute speech files and received feedback from expert labelers. Before participating in the prosody annotation study, they transcribed 90 second long speech files, and their transcriptions were evaluated by expert labelers.

Despite the extensive training of transcribers, there still remains a considerable amount of disagreement among transcribers as summarized in Table 1.1 (Dilley et al., 2006; Grice et al., 1996; Pitrelli et al., 1994; Syrdal and McGory, 2000; Yoon et al., 2004). Table 1.1 compares the agreement rates among transcribers regarding the presence or absence of pitch accent as well as of boundary tone in five different studies. Although the agreement rates are over 80% in the annotation of the location of pitch accent, and over 90% in boundary location, the disagreement rates are still quite high: around 20% for pitch accent and around 10% for boundary tones. Considering also the types of pitch accent and boundary tones in question, the agreement rates are even lower in all studies. In the present study, therefore, a new transcription method called

Rapid Prosody Transcription, has been designed for prosody annotation by a group of ordinary listeners, reflecting their perception of prosody in everyday communication. Rapid Prosody Transcription is further described in Chapter 2.

Chapter 2

Rapid Prosody Transcription

This section introduces the rapid prosody transcription (RPT), as a new, successful prosody annotation system.¹ RPT is distinct from other prosody transcription methods such as ToBI in that prosody annotation is done by groups of untrained, non-expert (“ordinary”) listeners in real-time online perception tasks and is solely based on the auditory impression. The main goals of this chapter are not only to introduce, but also to evaluate the reliability of RPT as a new prosody annotation method. In this chapter, I also evaluate whether groups of ordinary listeners’ transcription of prosody in spontaneous conversational speech from American English is consistent and reliable across listeners as well as compared to trained, “expert” listeners’ prosody transcription. In this chapter, we have used the RPT method for transcription of both prosodic prominence and boundary in spontaneous conversational speech, and the findings from reliability tests of RPT are reported.

2.1 Buckeye corpus

The Buckeye Corpus of Spontaneous Conversational Speech contains interviews of between 30 to 60 minutes duration between an interviewer and a single interviewee for a total of about 40 hours of speech (Pitt et al., 2007). Forty middle-class Caucasian interviewees (20 males and 20 females, all native of Central Ohio) were recruited

¹RPT was developed as a part of the NSF-funded interdisciplinary project, to create a training database for the development of an automatic prosody detection algorithm through collaborative effort under the direction of Jennifer Cole.

from the Columbus, Ohio community in 2000. The participants were split into two age groups (under 30 and over 40) which were balanced in terms of gender. The interviews were conducted in a small seminar room by one of two interviewers, either a 32-year-old male or a 25-year-old female. During the interviews, the interviewer asked questions about the interviewee as well as questions about which the interviewee could express their 'everyday' opinions on topics such as education, religion, school life, and politics. The interviewer did not often interrupt the interviewee's turn but sometimes they did as in everyday conversations. The interviewees were told that the purpose of the research was to determine how people express their opinions in conversation. The conversations between the interviewer and the interviewee were spontaneous and natural, and only the interviewee's turns were recorded, while the interviewer's turns were able to be heard as back channels.

The recorded interviews were subsequently orthographically and phonemically transcribed. After completing the orthographic transcriptions, phonemic representations were constructed for each word, and aligned with the audio speech recordings in two phases: (1) automatic phone alignment using Entropic Aligner software, and (2) subsequent manual correction. The transcriptions contained both speech and non-speech events such as silences, non-vocal and vocal noises, disfluencies, cut-offs, errors, fillers, as well as lexical items and phones. The DARPA-based phone set was employed for phone transcription, and four more symbols were added for syllabic nasals, the rounded reduced vowel, and the glottal stop. The labeling consistency across the four phone transcribers was evaluated by measuring six pairs of inter-transcriber agreement rates in lexical and in phone labeling. As reported in Raymond et al. (2002) and Pitt et al. (2005), the average pairwise agreement rate was 80.3%, and the agreement rates for consonants were generally higher than those for vowels. More specifically, the transcribers agreed most when labeling stops and fricatives for consonants, and when labeling diphthongs as well as the point vowels among monophthongs for vow-

els. Reduced vowels, however had considerably low transcription agreement rates. In addition to phone labeling, they reported that the mean deviation in boundary placement across transcribers was 16ms. In sum, this corpus contains reliably transcribed words, dictionary pronunciation and phone transcriptions all with time aligned sound files.

2.2 Materials and Subjects

A total of 54 speech excerpts from 25 randomly chosen speakers were extracted from the Buckeye Corpus of Spontaneous Speech of American English for the transcription experiments reported in this thesis. Two excerpts from another speaker were selected for demonstration purposes but were not included in any of the analyses reported here. The 56 total speech excerpts were used in the transcription session (2 for demonstration and practice and 54 for transcription). The excerpts were further divided into two groups. Two *short* excerpts of between 11–22 seconds duration were extracted from the interviews of each of 18 speakers. A second set of long excerpts of 31–58 seconds duration were selected from the same 18 speakers. Excerpts were selected according to the following criteria. First, I selected speech excerpts from segments of the interviews in which there are no technical recording problems. Due to this criterion, all the recordings from one particular interviewee (S23) were removed. Second, each speech excerpt was carefully selected in order to contain roughly the same proportion of the interview because there might be differences in linguistic, paralinguistic, physiological, and emotional factors, along with the timeline of the interview. That is, among 36 short excerpts, 12 speech excerpts were extracted from the beginning of the interview, another 12 from the middle portion of the interview, and the other 12 from the interview last part. Therefore, each set of the speech excerpts contain about 30% of the beginning, middle, and end of the interview. Third,

speech excerpts were selected to minimize the occurrence of disfluencies, though it was not possible to avoid all the disfluent regions for the selection of speech excerpts, especially in the long excerpts. Fourth, speech excerpts in the short and long sets were determined not to overlap.

There were some differences between short and long speech excerpts beyond their average length. First, the long speech excerpts were prepared with the purpose of investigating the influence of word token frequency and the repetition of individual words on prosodic form and acoustic implementation, and therefore the long excerpts contained at least one repetition of one or more content words. Although word repetition was not a factor in selecting short excerpts, they also contained some instances of repeated content words. Secondly, for the short excerpts, two speech intervals from each speaker were extracted, while one or two speech excerpts from each speaker were extracted to create the long excerpts. Due to the requirement such that each long excerpt must contain as many as possible content words that are repeated more than one time, it was not possible to extract one long speech excerpt from each speaker: some speakers interviews contain a great deal of repetitions of content words while others do not.

After extracting all the speech excerpts, loudness was normalized by dividing the mean RMS intensity of each sound file by the maximum mean RMS intensity, and then by scaling the maximum peak value to 1. In the short excerpts, a single speech excerpt was used in the transcription task for prosodic prominence transcription, and the other excerpt from the same speaker was used for prosodic boundary perception. In the long excerpts, the same speech excerpt was used for both prosodic prominence and boundary perception.

The excerpts were presented to transcribers in blocks according to the transcription focus (prominence annotation or boundary annotation). The short excerpts and long excerpts were transcribed in separate tasks performed in different sessions with

different groups of transcribers. In a transcription session, the excerpts were randomized within each block (prominence or boundary annotation), and the corresponding orthographic word transcripts were also prepared in the same randomized order on a printed page. In the transcripts, words were separated by a space and no punctuation or capitalization was used. Speech errors and disfluencies, if any, were included in the transcripts. For example, if, in a speech excerpt, a speaker intended to say “choose ” but the sound was cut-off at the onset of the word, the transcript included the cutoff word: ch- choose. Microsoft PowerPoint® files were constructed for the presentation of sound files to each transcriber, and the corresponding orthographic word transcripts were also provided (see Appendix A.1 for a sample transcript).

2.3 Procedure of Rapid Prosody Transcription (RPT)

In this thesis, prosody annotations of spontaneous conversational speech of American English were obtained from two separate experiments, which followed the same protocol, described here.

2.3.1 Experiment 1

SUBJECTS 74 participants, who are naïve in terms of phonetics and phonology, and without prior training of prosody transcription, were recruited from undergraduate linguistics courses at the University of Illinois at Urbana-Champaign. Although participants were not screened out before the experiments, the language background of each participant and their prior experience in prosody transcription were surveyed and later in the data analysis, prosody transcription from participants whose first or dominant language is not English or who had prior experience in prosody transcription was excluded. Listeners included in the analyses, therefore, were monolingual

native speakers of American English and had no prior training of prosody transcription. They participated in one of two sessions of rapid prosody transcription.

SETTING Prosody transcription took place in a computer laboratory, where 40 personal computers were installed in two sides separated by an aisle (about 20 computers in each side). Each computer was equipped with headphones, where a Microsoft PowerPoint® presentation file containing a set of speech excerpts for each participant was uploaded. The master computer is connected to a large screen on the front wall, where the instruction for prosody transcription was projected.

MATERIALS The speech excerpts extracted from the Buckeye corpus of spontaneous conversational speech of American English were first divided into two groups so that each group included one excerpt from each speaker. The 18 excerpts in each group were then divided into two separate blocks, one intended for prominence transcription and the other for boundary transcription. Within each block, the sound files were randomized for each participant. The other group contained the remaining 18 excerpts, which were prepared following the same procedure as described above. A printed transcript of the content in each speech excerpt (de-punctuated and no capitalization) was prepared to each participant, with the excerpts ordered to match the ordering of the sound files they would hear.

PRE-TASK Before starting the transcription task, the participants were provided a 5-minute introduction in which they were told the goal of the study (introduction, Appendix A.2.1) and were administered the informed consent form attached in Appendix A.2.2. The participants also completed a language survey form attached in Appendix A.2.3, where they listed their first language or any primary language that they used for daily life, if different from their first language, as well as any languages they had learned for more than one year, including the length of their education in each language. The participants were then provided simple definitions of prosodic prominence and boundary (Appendix A.2.1). A prominent word was defined as a

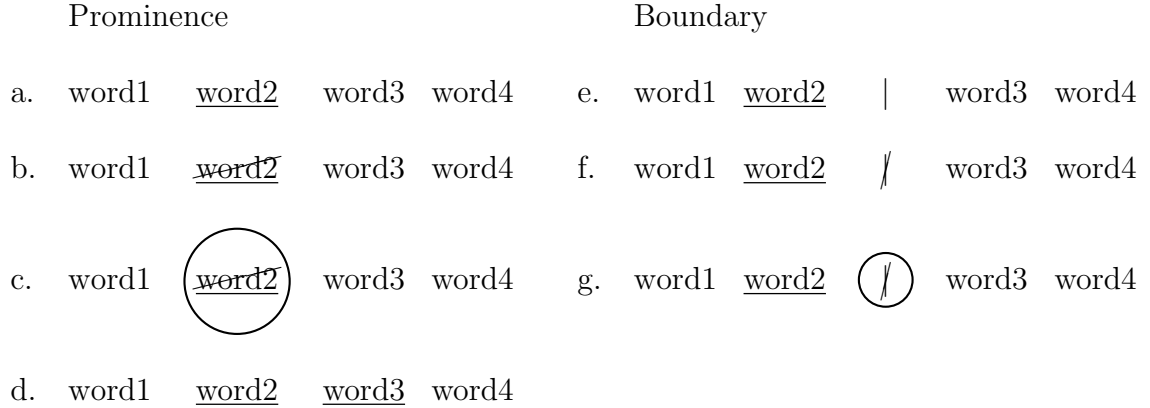


Figure 2.1: Transcription scheme

word that is “highlighted for the listener, and stands out from other non-prominent words”, while a chunk was defined as a grouping of words “that helps the listener interpret the utterance”, and that chunking is “especially important when the speaker produces long stretches of continuous speech”.

The whole transcription procedure was projected on the screen as schematized in Figure 2.1a–g. The participants were instructed to mark up their transcripts by underlining words they hear as “prominent” and by marking a vertical bar between words that belong to different “chunks” of the utterance, while listening to speech excerpts played in real time (Figure 2.1a and 2.1e). Changes to the transcription were able to be made during any play of the sound file. If listeners wanted to withdraw their decision on the locations of prominence or of boundary, they could cross out the markings as shown in (Figure 2.1b and 2.1f). The cancelled-out markings were able to be recalled with a circle (Figure 2.1c and 2.1g). In the end, a group of 15–23 subjects transcribed prosodic features, namely prosodic prominence and boundary, for each speech excerpt.

TASK 74 participants were divided into two groups: one for the first 18 speech excerpts, and the other for the remaining 18 speech excerpts. In this way, it was guaranteed that each subject would hear only one speech excerpt from each speaker.

During the transcription task, in one group sitting in the same side of the computer lab participants completed prominence transcription in the first block, and boundary transcription in the second. The other group sitting in the other side had the reverse order. By doing so, it was ensured that the order of prosody perception task for prominence and for boundary was balanced. In sum, each excerpt from each speaker received both prominence and boundary markings by different transcriber groups with a balanced order.

Then the participants were asked to check the volume of their headset and to follow the directions contained within the presentation files. Each presentation began with either prominence block or with boundary block, thus splitting the participants into two groups. They were provided with one practice sound file in the beginning of each block: one for prominence and the other for boundary. Once beginning the transcription task, each participant was able to listen to each sound file twice in the predetermined order, and at their own pace. In other words, the intervals between the play of each speech excerpt were regulated by the participant.

There were some rules that listeners needed to follow. Participants were instructed not to underline a whole phrase containing multiple words, but rather to underline each word separately, as prominent (Figure 2.1d). That is, their judgment must be made on a word-by-word basis. Participants played each sound file twice and at their own pace, making their transcripts as they listened, but they were not able to stop or resume the sound files in the middle of play, or after two times of play. Since the transcription task was done in real time, and listeners did not have much time, they were not allowed to correct their markings with an eraser, which would slow down their transcriptions. It is important to note that participants did not view any graphical display of the speech signal, and thus that the transcriptions were made solely on the basis of auditory impressions in concert with the printed orthographic transcript.

2.3.2 Experiment 2

23 subjects were recruited from the same pool of undergraduate students at the University of Illinois at Urbana-Champaign. Like in Experiment 1, the subjects were naïve in terms of the phonetics and phonology of prosody transcription. No participants who had previously participated in Experiment 1 participated in Experiment 2. They transcribed the locations of prominence and boundary for 18 long excerpts selected from 14 speakers’ interviews drawn from the Buckeye Corpus: one or two excerpts from each speaker’s interview. The procedure of prosody transcription in Experiment 2 was the same as in Experiment 1. Yet, while the participants listened to sound files from different speakers for both prominence perception and for boundary perception in Experiment 1, the participants in Experiment 2 marked prosodic prominence and boundary for the same speech excerpts from the same speakers. As a result, a group of 10–13 subjects marked the locations of prosodic prominence and boundary for each speech excerpt.

2.3.3 Results

After having collected prosody annotations from each experiment, data from 3 transcribers in Experiment 1 and from 3 transcribers in Experiment 2 were excluded due to either their failure to follow the transcription guidelines or their language backgrounds (they were not monolingual English speakers: they indicated either that their first language is not English or that they primarily speak another language in their life). The number of transcribers for each excerpt ranged from 10 to 22. For each excerpt, the transcriptions from all transcribers were pooled together, and each word was assigned a probabilistic prominence score (P-score) and a probabilistic boundary score (B-score) that codes the proportion of transcribers out of the whole group who marked that word as prominent or as followed by a prosodic boundary (e.g., final in a “chunk”). Figure 2.2 illustrates the probabilistic P- and B-scores of a part of

one excerpt from Speaker 26. For example, in this figure, about 33% of the subjects marked the first word, *I*, as prominent and nobody heard it as followed by a prosodic boundary and therefore, the first word, *I*, has 0.33 as a P-score and 0 as a B-score.

The pattern of prominence and boundary perception can be assessed in terms of the interval between prominence or boundary marks on each transcription sheet, based on the number of words between each prominence or boundary annotation. The means of the prominence intervals and boundary intervals are then calculated for each speaker, over all transcribers, as in (2). The distribution of these mean intervals by speaker is illustrated in Figure 2.3. The mean intervals were calculated by the following equation 2.2. The mean interval between prosodic prominences in Experiment 1 ranges from 5.46 to 8.13 words and from 5.45 to 9.67 words in Experiment 2, and the mean intervals between prosodic boundaries range from 4.43 to 11.5 words in Experiment 1 and from 6.81 to 11.5 words in Experiment 2.

$$\begin{aligned} \text{Mean intervals} &= \frac{(\text{Total no. of words})}{(\text{Total no. of prosodic markings})} & (2.1) \\ &= \frac{(\text{No. of words per excerpt}) \times (\text{No. of transcribers})}{(\text{Sum of prosodic markings by each transcriber})} \end{aligned}$$

These prosody scores have the following important characteristics: (1) they are probabilistic and pseudo-continuous rather than dichotomous; (2) perfect agreement of the presence of a prosodic feature across transcribers is reflected in the maximum prosody score, 1; and (3) perfect agreement of the absence of a prosodic feature is reflected as the minimum prosody score, 0. An important feature of this method of prosody annotation is that it directly encodes the variability in prosody perception across listeners, while the more commonly used method attempts to resolve or minimize inter-transcriber differences, e.g., through extensive training prior to transcription, through conferencing among transcribers to achieve consensus, or through majority rule. But in all cases the resulting transcription represents a dichotomous

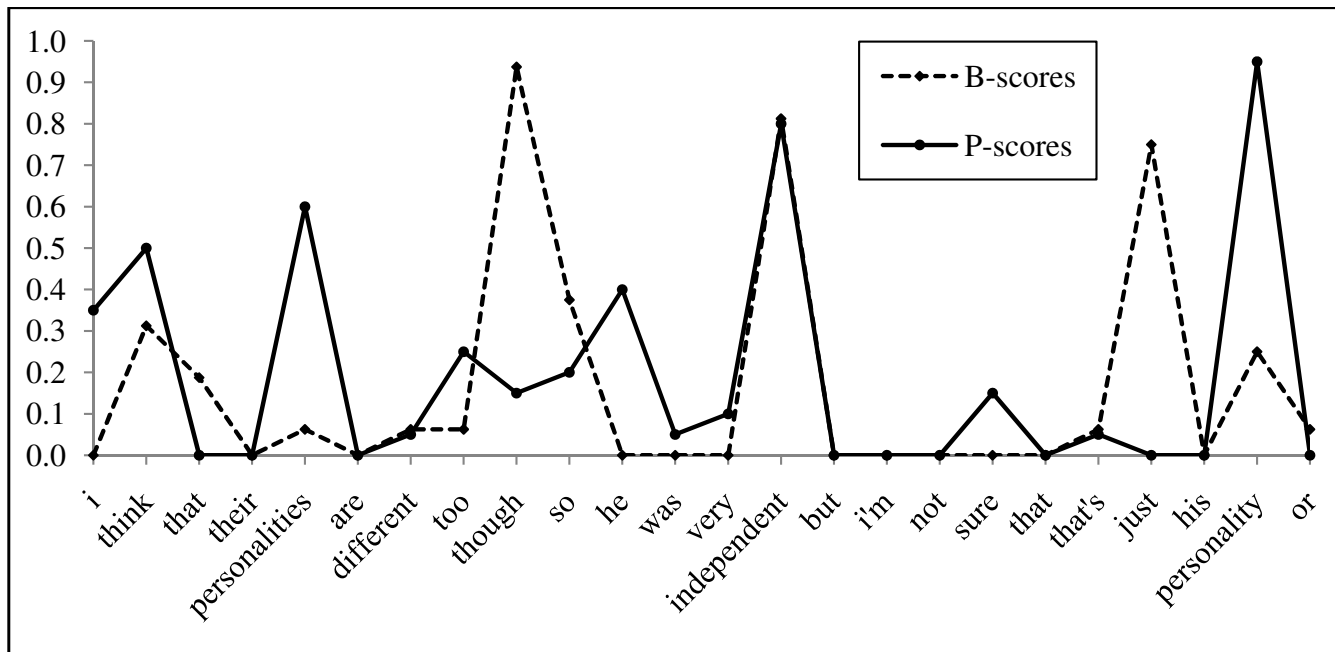


Figure 2.2: The distribution of probabilistic prominence (P, solid line) scores and boundary (B, dotted line) scores for each word in a sample utterance from Speaker 26

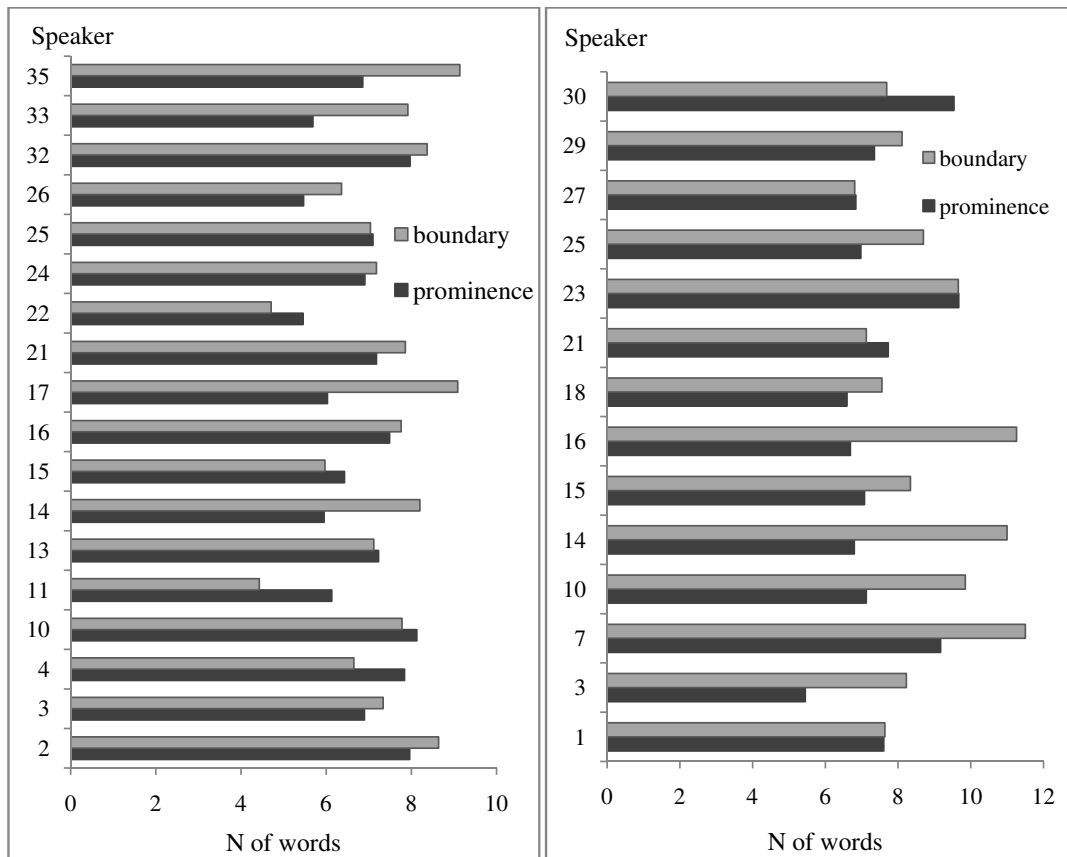


Figure 2.3: The distribution of the intervals between prosodic prominences (dark grey) and prosodic boundaries (light grey) in Experiment 1 on the left and in Experiment 2 on the right

distinction between a boundary or prominence label and its absence, with no encoding of the level of inter-transcriber agreement on any individual word or prosodic label. The prosody scores derived from the RPT method allow us to look directly at variation in prosody production and perception, examining prosody as an aspect of variable linguistic behavior.

At the same time, the prosody scores obtained through RPT require employing new statistical methods for data analysis. Previously, researchers made dichotomous or categorical comparisons of conditions that obtain when a certain prosodic landmark is present with conditions observed when it is not: prominent vs. non-prominent and phrase-initial or -final vs. phrase-medial. With the quasi-continuous-valued prosody scores obtained from RPT, correlations between prosodic features and other measures of the linguistic context can be calculated and regression analysis used to establish new models of prosodic prominence and boundary. Before looking into the relationship between prosody scores and other aspects of the linguistic context, however, it is necessary to establish that these prosody scores, reflecting ordinary listeners' perception of prosody, are consistent and reliable. In the following section, the reliability of annotations derived from RPT is evaluated.

2.3.4 Testing the reliability of RPT

As can be seen in Figure 2.2, many of the words in a given excerpt have low P- and B-scores, suggesting that transcribers reach high agreement rates for the absence of a prosodic feature on those words. Some words receive high scores, on the other hand, suggesting that transcribers also reached high agreement rates for the presence of a prosodic feature on those words. Regarding prominence, the transcribers agree on the non-prominence of many words as well as the prominence of some others. As for phrasing, they agree that on many words are not followed by a phrase boundary (i.e., are phrase-internal), and that some words are phrase-final. The reliability of the

prominence and boundary transcriptions by untrained non-expert ordinary listeners was evaluated in the following ways.

The first method used to evaluate if and how ordinary listeners' perception of prosody is reliable was to look at the Fleiss' kappa statistic for multi-transcriber agreement across all transcribers (Fleiss, 1971). To my knowledge, this work is the first to use Fleiss' kappa statistic for multi-transcribers agreement to assess the reliability of prosody annotation—a method which has subsequently been adopted to evaluate the reliability of transcribers' agreement scores in other studies (Cole et al., 2008; Prieto et al.; Yoon, 2010). Use of Fleiss' kappa statistic for multi-transcribers agreement has several benefits when compared to other ways of assessing agreement rates. First, similar to Cohen's kappa statistic for pairwise inter-transcriber agreement, Fleiss' kappa statistic takes into account pairwise agreements by chance. Other methods of assessing agreement, such as the percentage of agreements over the total number of agreements and disagreements, do not consider the agreement ratio by chance and therefore always overestimate agreement. Second, using Fleiss' kappa statistic provides a single coefficient as a measure of agreement among all pairs of transcribers, while Cohen's kappa calculates pairwise agreements, and multi-transcriber agreement is approximated using the mean kappa score. Third and most importantly, Fleiss kappa statistic for multi-transcriber agreement allows us to evaluate of the reliability of prosodic transcription using statistical methods, through the use of the z -statistic. The variance of the pooled prosody annotations relies on the number of transcribers as well as the number of agreements and disagreements.

The reliability of RPT transcriptions pooled over all 71 transcribers in Experiment 1 and all 20 transcribers in Experiment 2 were evaluated using Fleiss' kappa coefficient, and Fleiss' kappa scores were then z -normalized, as shown in Table 2.1. Fleiss' kappa statistic for prominence annotation ranged from 0.345 to 0.448 and the corresponding z -scores, from 21.5 to 33.4. Fleiss' kappa statistic for boundary an-

$z = 2.33, \alpha = 0.01$		Experiment 1				Experiment 2
		Run 1		Run 2		
		Group1	Group 2	Group 1	Group 2	
Prominence	Kappa	0.377	0.399	0.346	0.448	0.377
	z	25.2	22.7	21.5	33.4	32.8
Boundary	Kappa	0.601	0.587	0.532	0.640	0.580
	z	33.0	31.4	26.8	37.3	44.3

Table 2.1: Comparisons of the Fleiss’ kappa scores and the corresponding z -normalized scores in Experiment 1 and 2

notation ranged from 0.532 to 0.640, and the corresponding z -scores, from 26.8 to 44.3. The z -scores were significant with a 99% confidence level ($z = 2.33$), indicating that agreement among ordinary listeners regarding the perception of prominence and boundary is greatly above chance, and that prosody perception by ordinary listeners is consistent and reliable. Additionally, Table 2.1 indicates that ordinary listeners agree upon the presence or absence of a prosodic boundary more reliably than on the presence or absence of prosodic prominence.

The second method used to evaluate the reliability of RPT was to compare prosody annotations made by the ordinary listeners with expert transcribers’ prosody annotations. Although high Fleiss’ kappa multi-transcriber agreement scores showed that ordinary listeners agreed on prosody annotations among themselves, it should be determined whether the ordinary listeners’ perception of prosodic features is the same as expert listeners who were intensively trained for such annotation. I selected eight speech excerpts from those used in Experiment 1, four of which reached the highest agreement rates among the ordinary listeners, and the other four of which received the lowest agreement rates among them. These eight speech excerpts were prosodically annotated by three trained expert transcribers, two of whom had more than two years of training in the ToBI transcription method including the author. The expert labelers marked the locations of prosodic prominence and boundary with the aid of

visual displays derived from the speech signal, including sound waves, spectrograms, pitch, intensity, and formant tracking contours. They were also allowed to listen to the sound files as many times as needed, and to zoom in and out on some parts of speech if necessary. When identifying prosodic prominence and boundary, the “expert” labelers referred to any available information in the speech stream including the height, movement, and the slope of F_0 , as well as the lengthening or shortening and the hypo or hyper-articulation of the part of speech, following the phonology and the phonetics of prosody annotation.

After collecting prosody transcription by three expert transcribers, the agreement rates among three expert transcribers were first calculated using Fleiss’ kappa statistic. Although the actual agreement rate for prominence transcription ($p(A) = 0.866$) is slightly lower than the boundary transcription agreement ($p(A) = 0.890$), the Fleiss’ kappa scores for prominence transcription ($\kappa = 0.719$) is higher than for boundary transcription ($\kappa = 0.619$) because in boundary transcription there is a greater chance of marking no boundaries after each word than in prominence transcription. Differences among the three expert transcribers were resolved through consensus during discussions that followed each transcriber’s individual transcription work. Differences that could not be resolved through consensus were resolved based on the majority transcription.

Pairwise Cohen’s kappa scores between the agreed transcription by the expert labelers and each of the ordinary labelers were also calculated. The boxplot in Figure 2.4 displays the distribution of the total 56 pairs of pairwise Cohen’s kappa agreement scores for prominence as well as the 51 pairs for boundary perception. The kappa scores of prominence transcription range from 0.115 to 0.809 with a mean agreement score of 0.366 and those of boundary transcription from 0.361 to 0.760, with a mean agreement score of 0.580. As expected, great variation was found in pairwise Cohen’s agreement scores between the agreed prosody transcription by the expert transcribers

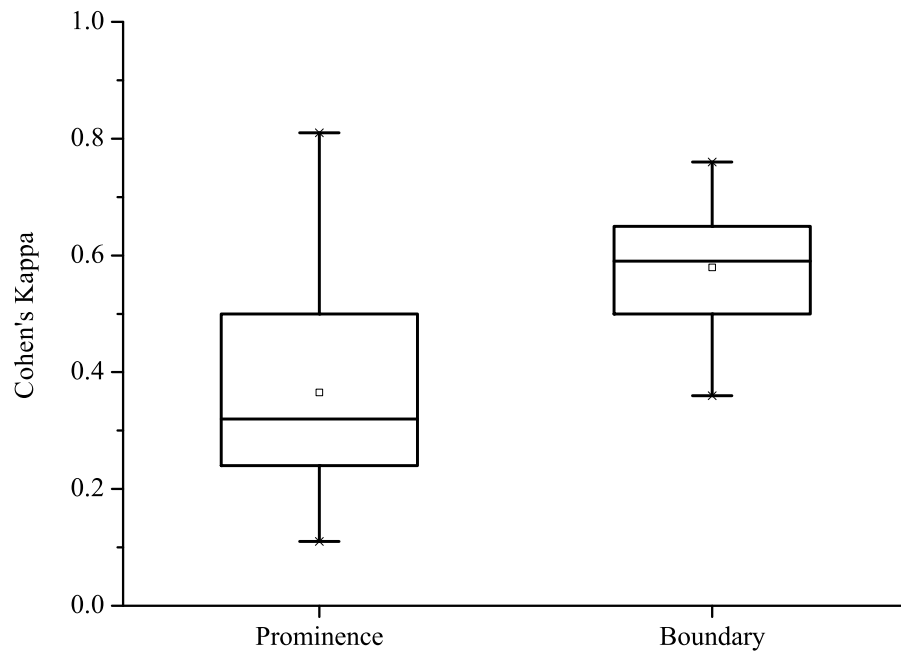


Figure 2.4: The boxplots of pairwise Cohen's kappa scores between the agreed prosody transcription and the prosody transcription (prominence on the left and boundary on the right)

and the prosody transcription by each of the ordinary listeners. As for prominence annotation, variation arose from the fact that the ordinary listeners identified fewer words as prominent than did the expert labelers. That is, in most cases, the ordinary transcribers identified a subset of prominence markings from those marked by the expert transcribers. In relation to boundary transcription, on the other hand, variation originated mostly from one type of disfluency - filled pause. Overall, there was greater variation in pairwise agreement scores for prominence transcription than for boundary transcription.

It is evident that the ordinary listeners identified a subset of the prosody labels marked by expert transcribers, which could be due to many factors, such as the time limitation, or the difficulty of interpreting weak cues. Therefore, when looking at

the confusion matrices of agreements and disagreements between the expert and the ordinary listeners, it is interesting to look at which words at least one ordinary listener and at least one expert labelers marked as prominent or as preceding a prosodic boundary. In other words, I treated any prosodic feature identified by a single labeler as valid. This analysis assigns a prosody value of “1” to all words with a prosody score (prominence or boundary) greater than zero, with zero-valued prosody scores assigned the “zero” label. This transforms the quasi-continuous valued prosody scores into dichotomous prosody features for prominence and boundary. The agreement rates between the ordinary listeners’ prosody transcription and prosody transcription by the expert listeners were 0.563 for prominence transcription and 0.597 for boundary transcription, suggesting that words where prosodic features were identified by one or a small number of the ordinary listeners coincide with words where the expert labelers marked prosodic features. That is, it was shown that the ordinary listeners’ perception of prosodic features is not much different from the expert transcribers’ perception of prosodic features, although their prosody transcriptions still disagree upon prosody transcription to some extent. Yet, the results also show that the agreement rates between the expert and the ordinary labelers for prominence transcription greatly increase, but there is no improvement in the agreement rates for boundary perception, when comparing the mean pairwise Cohen’s kappa scores ($\kappa = 0.366$ for prominence and 0.580 for boundary) with the pairwise Cohen’s kappa scores ($\kappa = 0.563$ for prominence and 0.597 for boundary) listed above. This indicates that in boundary perception, there are few words after which a few ordinary listeners marked a prosodic boundary but a majority of them did not, while in prominence perception there are many words which only few ordinary listeners marked prosodic prominence, revealing great variation in prominence markings across ordinary listeners but not such variation in boundary markings. In other words, an ordinary listener tends to mark a subset of prominent words that the expert labelers indicate as promi-

ment. On the other hand, in boundary perception, an ordinary listener is likely to mark more or less the same number of boundaries at almost the same locations as the expert listeners. There are several possible reasons for these results. First, it may be attributed to the fact that the interval between prominences in words is usually much shorter than the interval between boundaries, and therefore prominence perception requires more cognitive attention or more time to mark all prominent words than boundary perception. Or, it may also result from the fact that there are many ways of acoustic encoding of prosodic prominence and therefore acoustically weak prominence is perceived by fewer listeners.

2.4 Summary

In this study, the Rapid Prosody Transcription (RPT) method has been introduced as a new method of prosody annotation, and its benefits, relevance, and the reliability of this method have been discussed. Various evaluations of the reliability of RPT demonstrate that (1) prosody annotation obtained by a group of ordinary listeners is consistent and reliable within the group; (2) prosody annotation by ordinary listeners is comparable to and not greatly different from the expert listeners' prosody annotation, especially when pooling prosody annotations across all ordinary listeners; (3) although ordinary listeners' prosody transcription is comparable to the expert ones', there still remains great variation in prosody transcription by ordinary listeners; and finally (4) the variation in prosody transcription by any individual ordinary listener, possibly induced by performance errors, can be reduced by obtaining prosody transcriptions by multiple ordinary listeners, approximating professional labelers' transcription. Ordinary listeners' performance errors may result from time limitation as well as possibly limited sources of acoustic information. As previously described, RPT is real-time prosody transcription, and the labelers can only listen to

speech excerpts twice in real time. Therefore, it is possible that some weak prosodic features, in particular prosodic prominence which occurs more often than prosodic boundary in nature, may be perceived by only a few ordinary listeners. The findings from this study altogether suggest Rapid Prosody Transcription (RPT) by multiple ordinary listeners as a new method of prosody annotation that is both a valid and reliable way to obtain prosody transcriptions, allowing us to variability in the production and the perception of prosody in addition to the linguistic correlates of prosody. In the next chapters, various analyses will be performed to investigate such linguistic correlates of prosody as determined by multiple ordinary listeners as well as variability in prosody production and perception.

Chapter 3

The Distribution of Prosodic Scores by Phone Identity

3.1 Introduction

In this chapter, I examine whether the phone identity of the stressed vowel influences listeners' judgment on the presence/ absence of prosodic features. In other words, this chapter looks at the distribution of prosody scores by vowel phone category to see if prominence is more likely to be perceived on words with certain stressed vowels, or similarly, if prosodic boundaries are more likely to be perceived following syllables with certain vowels.

Many prior studies that investigate acoustic correlates of prosody in English examine only a few vowels from the phoneme inventory. For example, Cho (2005) compared acoustic and articulatory measures of the focused and non-focused versions of the vowel /ɑ/ and /i/. Beckman and her colleagues (1992; 1994) looked at the effects of accent and stress with the target vowels, /ɑ, ə/, and in his perception study, Kohler (2008) used one target word, *baba* to control the phonemic vowel category of interest. It is well known that there are differences among vowel phonemes in their 'intrinsic' intensity, duration and F_0 properties, and since these same properties are the primary correlates of prominence in English, it is important to look broadly across the vowel inventory to fully understand the effects of prosody on vowel production.

Findings from corpus studies of Greenberg and his colleagues (Greenberg et al., 2002, 2003; Hitchcock and Greenberg, 2001) also support the relevance of examining the effects of prosody by phone category. Looking at temporal properties of sponta-

neous phone conversations of American English with the Switchboard corpus, they found that vowel duration is primarily determined by vowel height and the more open and intrinsically long vowels, e.g., diphthongs and low and mid monophthongs are more likely to be fully stressed than the intrinsically short vowels including high diphthongs and monophthongs.

This thesis investigates prosody production and perception in a corpus study that encompasses all word classes, and which includes acoustic analyses from nearly the full inventory of vowel phonemes: 14 vowels of American English (every vowel in the American English phonemic system except the diphthong /ɔɪ/, of which there were insufficient tokens). In the following section, therefore, I examine whether there is any tight relationship between phone identity and prosodic features before further acoustic analyses in Chapter 4 and 5.

3.2 Distribution of prosody scores by phone

Table 3.1 summarizes the distribution of each vowel phone in the prosodically annotated speech materials analyzed in this study. This set of vowels includes only lexically stressed vowels in the distribution of perceived prominence, and only lexically stressed word-final vowels in the distribution of perceived boundary. Mean probabilistic P(rominence)- and B(oundary)-scores of each vowel phone are illustrated in Figure 3.1. As shown in Table 3.1 and Figure 3.1, each phone differs not only in terms of the likelihood of its occurrence in the corpus, but also in terms of its likelihood to receive a prominence or boundary marking in Rapid Prosody Transcription.

3.2.1 Distribution of prominence scores by phone

Mean probabilistic P-scores calculated over entire set of stressed vowel phonemes in the database range from 0.114 for /ɪ/ to 0.243 for /ɔ/. These values are for each vowel

Lexically stressed V	ɑ	æ	ʌ	ɔ	ɛ	ɜ ^ɹ	ɪ	i	ʊ	u	aʊ	ɑɪ	eɪ	oʊ
N	166	283	387	119	438	116	469	304	72	177	50	301	207	186
Lexically stressed word-final V	ɑ	æ	ʌ	ɔ	ɛ	ɜ ^ɹ	ɪ	i	ʊ	u	aʊ	ɑɪ	eɪ	oʊ
N	126	202	284	94	308	88	364	240	60	151	46	274	148	70

Table 3.1: The distribution of lexically stressed vowels and lexically stressed word final vowels

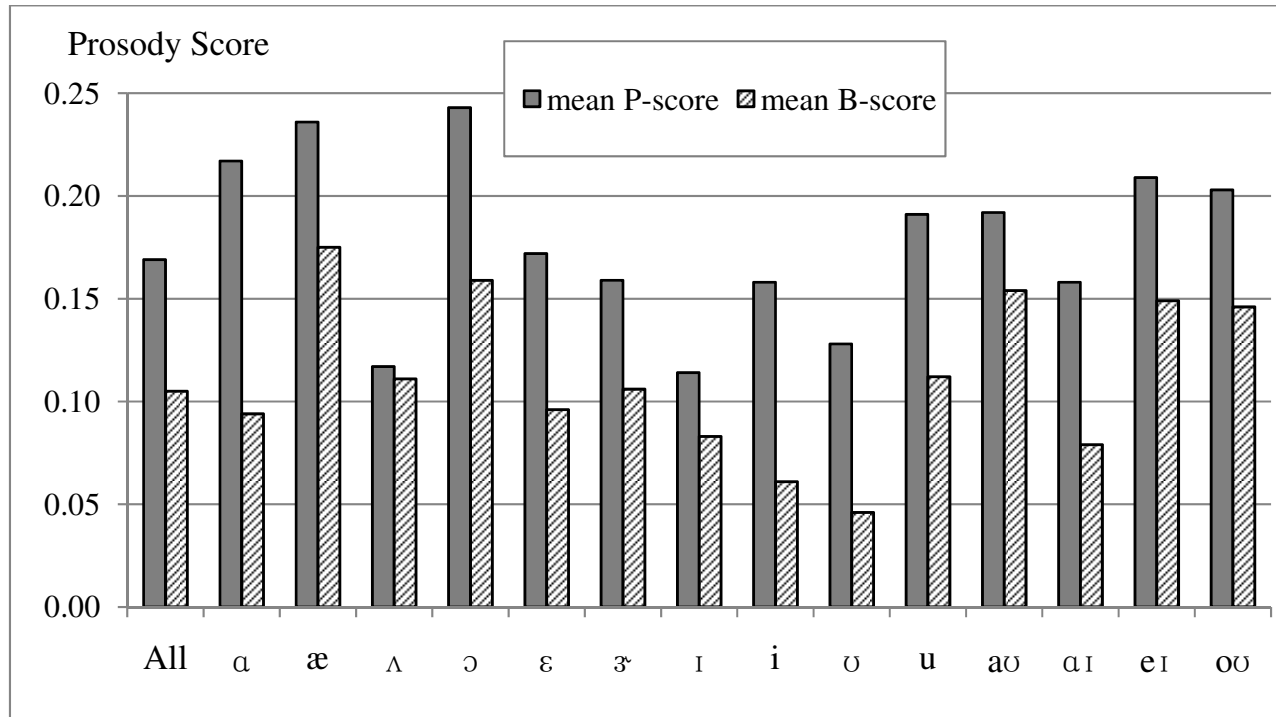


Figure 3.1: The distribution of the mean probabilistic P(rominence)-scores (solid) and B(oundary)-scores (oblique) of each phone

phone higher than the average B-scores, which range from 0.046 for /ʊ/ to 0.175 for /æ/, suggesting that across transcribers and over all speech excerpts, there are more words marked as prominent than there are words preceding a marked prosodic boundary.

One-way ANOVAs conducted to test for an effect of vowel identity (phoneme category) on P- and B-scores show significant differences. Subsequent post hoc tests further revealed that the vowel /æ/ has meaningfully higher P-scores than many vowels including /ʌ, ɑ, ε, ɪ, ʊ/, and that the vowels /ʌ, ɪ/ have statistically lower P-scores than many vowels including /ɑ, æ, ɔ, ε, eɪ, oʊ, u/. The diphthongs and open vowels /ɑ, æ, ɔ, au, eɪ, oʊ, u/, have relatively higher P-scores (P-scores > 0.190), while the vowels /ʌ, ɪ/ that are intrinsically short and often reduced, have the lowest P-scores (P-scores < 0.120). P-scores for the other vowels including a diphthong /ɑɪ/ and four vowels /ε, ɜ, i, ʊ/, are in between the other two subcategories (0.120 < P-scores < 0.190).

3.2.2 The distribution of boundary scores

Mean probabilistic B-score are not uniformly distributed over phones. There is, however, no distinct sub-group by phone identity for B-scores. The vowel /æ/ tends to have higher B-scores than the vowels /ɑ, ε, ɪ, i, ʊ/, and the vowel /i, ʊ/ has significantly lower B-scores than the vowels /æ, ɔ, eɪ/. Yet the mean differences of the average B-scores for the other vowels are not meaningfully large.

3.3 Conclusion and Discussion

Results show that mean probabilistic prosody scores are not evenly distributed over words including stressed vowel with different phone identity. The findings from the statistical analyses of P- and B-scores and the phonemic category of the vowels

demonstrate that depending on the phonemic vowel category, some vowels tend to be more prominent or to be more often perceived as followed by a phrase boundary than other vowels, providing the validation to examine the acoustic variation associated with prosodic features by phone. Especially, accordingly with findings from prior studies by Greenberg and his colleagues, the results indicate a systematic pattern between prominence and vowel identity, showing that long vowels including diphthongs and open vowels generally tend to be perceived as prominent than short vowels.

Findings from this study are, however, not entirely consistent with Greenberg and his colleagues' on the relation between vowel height and stress accent. First of all, in their study in the Switchboard corpus of spontaneous telephone dialogues, Hitchcock and Greenberg (2001) found that high vowels are much more likely to be unaccented. However, I find that not all words including high vowels that are intrinsically short in duration are more likely to be perceived as not prominent than non-high vowels. For example, high vowels /i, u/ have relatively high mean probabilistic P-scores and the vowel /i/ has almost the same mean P-scores as the vowel /ɜ:/ and the mean P-score of the vowel /u/ is higher than or almost the same as the mean P-scores of vowels /ɛ, ɜ, əv, aɪ/. Second of all, Hitchcock and Greenberg (2001) claimed that low vowels are much more likely to be fully accented than high vowels but my study shows that some words including non-high vowels that are not intrinsically short in duration are perceived as prominent. For example, a mid vowel /ʌ/ has the lowest P-scores among all stressed vowels. Furthermore, the relationship between vowel height and the likelihood of being prominent cannot directly be investigated without the consideration of word class. The examination of word class of the lexical items indicates the tight interaction between accentuation and word class, namely function and content words. In this study, it is found that the intrinsically long diphthong /aɪ/ has a low average P-score compared to some intrinsically short vowels /i, u/ as well as other intrinsically long vowels including other diphthongs and non-high

monophthongs. Looking closely at the lexical items, however, it is shown that a large portion of words with the diphthong /aɪ/ are function words. In particular, 70.1% (211 out of 301) of the tokens come from the first person pronominal forms ‘I’, ‘my’, and ‘mine’ (175) and a discourse marker ‘like’ (36), which are rarely used in pragmatic contexts as discourse new or focused, and which are likely to be reduced in most cases. Excluding these lexical items, the mean P-score of the diphthong, /aɪ/ is 0.301. These findings suggest that vowel identity is related to the likelihood to be perceived as prominent but other factors such as word class interplay with it in prominence perception. In the following two chapters, I will first look at the influence of prosodic features on acoustic variation and its contribution to prosody perception by each phone category.

Chapter 4

An Acoustic Investigation of Prosodic Prominence

4.1 Introduction

Chapters 4 and 5 examine the phonetic encoding of prosody in conversational speech through acoustic measures of the Buckeye corpus speech materials for which prosody transcriptions have been obtained. The goal is to understand what properties of the acoustic signal are correlated with the perception of prosodic prominence and boundary by ordinary listeners performing Rapid Prosody Transcription. Prior studies have demonstrated acoustic correlates of prominence and boundary based on (i) the analysis of speech (often read speech) produced in controlled experiments designed to elicit specific prosodic structures (Gussenhoven and Rietveld, 1988; Sluijter et al., 1995; Sluijter and van Heuven, 1995, 1996b; Turk and Sawusch, 1996; Turk and White, 1999 and many others), or (ii) the analysis of speech from corpora which are prosodically transcribed (Cole et al., 2007; Greenberg et al., 2003; Kim et al., 2006; Kochanski et al., 2005; Yoon et al., 2007 and many others). Collectively, these studies demonstrate that prosodic context is a significant source of phonetic variation affecting the suprasegmental properties of speech sounds (e.g., pitch, loudness, spectral modulation, and duration) as well as segmental properties (e.g., vowel formant patterns and consonant voicing). These findings from prior studies reveal consistent effects of prosodic context on words that occur at the initial and final edges of prosodic domains and on words that are assigned prominence. In this chapter and the next we examine acoustic evidence for effects of prosodic context based on the prosodic

annotation obtained from RPT. The specific aims of this inquiry are to identify which individual acoustic measures are correlated with listeners' perception of prominence and boundary in conversational speech, and to measure the contribution of each correlated measure in predicting listeners' perception of prosody. A further goal is to construct a statistical model of the influence of acoustic cues in listeners' perceptual response.

This chapter presents the findings from the acoustic study of prosodic prominence with correlation analyses of the relationship between perceived prominence, as measured by P-scores that encode the location and strength of perceived prominence, with acoustic measures. Results from multiple linear regression analyses are presented, and provide statistical models of the acoustic cues to prominence perception as determined by ordinary listeners. The following section will first review seminal works concerning the effects of prominence on the acoustic characteristics of speech, and the contribution of acoustic cues to the perception of prominence.

4.1.1 Fundamental frequency (F_0)

Pitch, or its acoustic correlate in fundamental frequency (F_0) is traditionally described as a primary cue for prominence in many languages, including American English (Beckman, 1986; Bolinger, 1958; Ladd, 2008; Liberman, 1975; Pierrehumbert, 1980; Roca and Johnson, 1999 among others). Previous production and perception studies have examined the relationship between F_0 variation and prosodic prominence in many languages: some of these studies have investigated the size of F_0 excursions in relation to prominence (Fry, 1958; Gussenhoven and Rietveld, 1988; Gussenhoven et al., 1997; Hermes and Rump, 1994; Lieberman, 1960, 1965; Liberman and Pierrehumbert, 1984), while other studies have explored the relationship between the shape of F_0 excursions and prominence production and perception (Erikson and Alstermark, 1972; Hermes and Rump, 1994; Rump, 1996; Shattuck-Hufnagel et al., 2004).

Many of the early production experiments investigate F_0 variation in relation to prominence on the basis of linguistic judgments determining the location of stress or accent (Cooper et al., 1985; Erikson and Alstermark, 1972; Lieberman, 1960). Later studies, in contrast, employed perceptual data obtained from controlled experiments. For instance, Lieberman (1965) performed transcription experiments with two trained linguists who transcribed prominence of the vowel /a/ with manipulated F_0 and amplitude in eight different versions of original utterances, each of which was claimed to represent different emotions. Lieberman (1965) found that transcription by linguists was dependent on pitch information rather than loudness. While Lieberman's study focused on a single perceived prominence in a phrase or utterance, Pierrehumbert (1979) conducted perception study in which she asked native English speakers to judge which stressed syllable of a multi-syllabic word in nonsense sentences had the higher pitch peak, while varying F_0 values on the last stressed syllable and eventually the degree of declination of F_0 . The results showed that even when F_0 of the second stressed syllable is lower than that of the first stressed syllable, the first and second syllables were judged as being equal in pitch, claiming that listeners are able to normalize for the declination of pitch over the course of an utterance when making prominence judgments.

Subsequent studies expanded on the findings of Pierrehumbert (1979) by investigating prominence perception in languages such as Dutch (Gussenhoven and Rietveld, 1988; Gussenhoven et al., 1997; Rietveld and Gussenhoven, 1985; Terken, 1991, 1994). For instance, in a series of perception studies with Dutch listeners, Terken (1991, 1994) manipulated the relative height of F_0 at various locations in a reiterant nonsense sentence in order to model the perception of prominence in contours with declining F_0 . Based on his findings, Terken concluded that listeners correct for baseline declination in their perception of prominence at various locations across a phrase. In a later study, Gussenhoven et al. (1997) tested the relation between the perception of two

prominences and their relative F_0 height with simple sentences originally produced by both a female and a male native speaker of Dutch. The results of this study corroborated earlier findings in showing that both the location of prominence within an utterance as well as the change of F_0 affect the perception of prominence, providing further evidence that in the perception of prominence, listeners rely on an abstract baseline F_0 rather than on a direct interpolation of raw F_0 values across successive peaks.

Other studies challenge the view that F_0 is the primary correlate of, and the most salient cue for, prominence (Heldner and Strangert, 1997; Kochanski, 2006; Kochanski et al., 2005; Fant et al., 1994; Sluijter and van Heuven, 1995, 1996b among others). Heldner and Strangert (1997) prepared two types of stimuli. In the first, the size of the F_0 rise of a phrase-medial, focused word was reduced, and in the second, the size of the F_0 rise of a phrase-medial, non-focused word was increased. Sentences were constructed to carry a narrow focus induced by a prompting question. As results, the stimuli presented to listeners provided conflicting F_0 information indicating the presence or absence of a focus: in the first, reduced F_0 in a focused word and in the second, increased F_0 in an unfocused word. It was found that listeners were insensitive to the manipulated F_0 pattern when presented with conflicting acoustic information. Neither the gradual addition of an F_0 rise on the non-focused word, nor the gradual reduction of F_0 rise on the focused word changed listeners' judgments on the presence or absence of prominence. They concluded that F_0 rise is neither necessary nor sufficient for the perception of focus, further claiming that F_0 movements are optional from the listener's point of view. More recently, in their large corpus studies, Kochanski (2006) and Kochanski et al. (2005) also claimed that F_0 plays at best a minor role as a correlate of prominence in production, or as a cue to the perception of prominence.

4.1.2 Other acoustic correlates of prominence

In addition to measures of F_0 , there are other segmental and suprasegmental acoustic measures shown in prior studies to be acoustic correlates of, and the acoustic cues for, prominence (e.g., duration, intensity, vowel formants, and spectral properties). However, studies do not agree on which acoustic measure or measures reliably cue prominence, and the question of how individual acoustic measures contribute as to the perception of prominence has not yet been well examined in the literature, though, duration and intensity are frequently reported as significant correlates of prominence.

It should be noted that in many early acoustic studies on prominence, there was no distinction between word- and sentence- or phrase-level prominence, either because the stimuli contained a list of isolated lexical items (e.g., Fry, 1955, 1958; Lieberman, 1960), or because the study looked at sentence or phrase-level prominence (phrasal accent) in addition to word-level prominence (lexical stress) (e.g., Cambier-Langeveld and Turk, 1999; Turk and Sawusch, 1996, 1997; Fant et al., 2000b,a). In one of the earliest acoustic studies on prominence, Fry (1955, 1958) measured duration and intensity as physical correlates of linguistic stress in English. He found that both duration and envelope amplitude ratios are relevant to listeners' stress judgments of disyllabic words, but that duration is more important than envelope amplitude in the perception of linguistic stress. Contrary to Fry (1955, 1958), Lieberman (1960) showed that peak envelope amplitude is the more relevant factor for listeners' syllable stress judgments. He suggested a schematized algorithm for syllable stress judgments on the basis of his acoustic findings. However, he stated that his findings do not clearly identify any single acoustic measure as about the single most important acoustic correlate of lexical stress.

In recent studies, many scholars have investigated how prosodic prominence influences the temporal implementation of speech, and how such temporal variation affects the perception of prosodic prominence in many languages. More specifically, stud-

ies have examined the effects of temporal expansion (durational lengthening) on the perception of prominence in many languages (Cambier-Langeveld and Turk, 1999 for Dutch and English; Eefting, 1991; Nootboom, 1972; Sluijter and van Heuven, 1995 for Dutch; Heldner and Strangert, 1997 for Swedish; Maekawa, 1997 for Japanese; Turk and Sawusch, 1996, 1997; Turk and White, 1999 for English). Nootboom (1972) explored the temporal effects of prominence in Dutch with reiterant nonsense words, varying the number of syllables, the positions of lexical stress, and the locations of a prominence-lending pitch accent. Results showed that all syllable nuclei in an accented word are lengthened. Sluijter and van Heuven (1995, 1996a,b,c) also investigated the effects of prominence-lending focal accents on syllables with reiterant nonsense words produced by Dutch speakers. The results indicated not only the effects of accentual lengthening of stressed syllables, but also some degree of temporal effect on unstressed syllables in an accented word. Heldner and Strangert (2001) explored the same question in Swedish. Instead of employing reiterant nonsense words, they used two simple sentences in which the locations of focal accents were elicited in a question-answer context. By comparing the relative durations of words, stressed syllables, and unstressed syllables, they found that when words are in focus, both word duration and stressed syllable duration are increased, but that the lengthening of focused words is mostly induced by the lengthening of stressed syllables. They also remark that the greatest variation in syllable and word duration arises from speaker variability.

English is another language in which researchers have explored the temporal effects of prominence. Beckman (1986) observes significant correlations between prominence and a combination of intensity and duration measures. Turk and her colleagues (Turk and Sawusch, 1996, 1997; Turk and White, 1999) studied the lengthening effects of prominence in English, and its strength as a cue to prominence relative to loudness information. For example, Turk and Sawusch (1996) manipulated duration

and intensity of reiterant two-syllable words each of which had a lexical stress at a different position in the word. It was found that there is a trading relation between duration and intensity. However, a linear regression model showed that durational information predicts prominence more consistently than loudness, and that loudness alone possesses little power for predicting prominence in English. Adopting gradient scales of prominence (R_s (syllable response predicted by a regression line), originally introduced in Fant and Kruckenberg, 1989) = 0, 10, 20, and 30) similar to Turk and Sawusch (1996), Fant and his colleagues (2000a; 2000b) investigated correlations between various acoustic measures including duration and listeners' prominence judgments. Their findings were consistent with those of Turk and Sawusch (1996) in finding that syllable duration is correlated with prominence judgments and is a good predictor of prominence.

Relatively few studies have examined spectral correlates of prominence (Heldner, 2003; Sluijter et al., 1997; Sluijter, 1995; Sluijter and van Heuven, 1996a,b,c). More recent studies have focused on sub-band spectral measures (SPHL-SPL in Fant et al., 2000a,b; spectral emphasis in Heldner, 2001a, 2003; spectral balance, $H1^*-H2^*$, $H1^*-A1$ and so forth in Sluijter and van Heuven, 1996a,b,c; Sluijter et al., 1997), whereas in earlier studies, overall intensity as a physiological correlate of loudness was tested as a relevant acoustic correlate of prominence (Fry, 1955, 1958; Lieberman, 1960; Lehiste and Fox, 1993). Sluijter and van Heuven's series of studies in Dutch are pioneering in this area. Adopting Glave and Reitveld's idea (1975) that vocal effort affects the intensity of speech spectra above 500 Hz but not below 500 Hz, Sluijter and van Heuven (1996b) employed band filtered intensities as a measure of "spectral balance" in four different frequency band regions: 0–500, 500–1000, 1000–2000, 2000–4000 Hz. Their main goal was to find the acoustic correlates of two hierarchically different levels of prominence, namely lexical stress and accent in Dutch. Among the results, they indicated that spectral balance is a stronger correlate of lexical stress than of accent,

and that its strength as a cue for prominence is comparable to that of duration. This study furthermore suggested that different acoustic measures are related to different levels of prominence. That is, duration and spectral balance are more likely to be reliable correlates of lexical stress while overall intensity is more likely to be a reliable correlate of accent. Later, they tested the perceptual relevance of duration, overall intensity, and spectral balance above 500 Hz as acoustic correlates of linguistic stress with a reiterant disyllabic nonsense word pair (Sluijter et al., 1997). It was shown that emphasized spectral balance over 500 Hz enhanced the perception of prominence, and that overall intensity has little influence on the perception of lexical stress. They further showed that the strength of sub-band intensity (spectral balance) as a cue for prominence is close to duration, but that in a reverberant environment, spectral balance tends to be better than duration as a perceptual cue to prominence.

Similar studies were conducted by Heldner (2001a,b, 2003). Heldner (2003) investigated the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. His measurements of spectral emphasis are quite different from those of Sluijter and van Heuven (1996b). Instead of setting up static cut-off frequencies, he used dynamic low-pass filters with a cut-off frequency set at 1.5 times the F_0 mean for each utterance. He compared spectral effects of prominence paradigmatically (between sentences) and syntagmatically (in its vicinity, within a sentence). The results of production and of automatic detection studies showed that both overall intensity and spectral emphasis are reliable acoustic correlates of prominence, and, furthermore, that spectral emphasis is a more reliable predictor of prominence. In his perception study, Heldner (2001b) tested the influence of spectral emphasis as a perceptual cue. The materials used in the study were (1) a read-aloud short story produced by a male native Swedish speaker in which accents were indicated by capitalization and (2) a synthesized story by a male mbrola synthesizer in which accents were marked by duration and F_0 . Listeners were asked to compare the prominence

and the naturalness of pairs of stimuli. The results did not, however, show any significant effects of spectral emphasis on the perception of prominence when implemented as focal accents or on the naturalness of speech. Yet, in his later acoustic and auto-detection studies (Heldner and Strangert, 2001; Heldner, 2003), he showed that spectral emphasis measured using a dynamic low-pass filter following the course of F_0 (detection rate: 75%) is a more reliable cue than overall intensity (detection rate: 69%) in the automatic detection of prominence—namely, focal accent.

Lastly, there also exist a small number of acoustic studies showing that vowel quality varies under prominence. Sluijter and his colleagues (Sluijter, 1995; Sluijter et al., 1995; Sluijter and van Heuven, 1996b; Sluijter et al., 1997) showed that both stress and accent affect the formant structures of vowels in Dutch. When vowels are not focused or unstressed, the first formant value is raised, which indicates that vowels are either reduced, or that they are produced with more mouth opening. It was also found that the second formant values are centralized when the vowel is in a non-focus condition. In his production study, van Bergem (1993) also investigated the effects of prominence (sentence accent and word stress) on both the steady-state and the dynamic formant structures of Dutch vowels. Like many other studies, production of sentence level prominence in specific locations was prompted by questions. It was shown that both sentence level and lexical prominence have significant effects on steady-state formant frequencies as well as on duration. Yet, it was further tested that the effect of sentence level prominence is less than that of lexical stress in a subsequent perception study. van Bergem (1993) tested the effects of sentence and lexical level prominence on vowel identification. When he presented listeners with segmented vowels extracted from three speakers in the production experiment, the listeners' vowel identification rate increased when the vowels were stressed and accented. This confirmed the relevance of word and phrasal prominence in the perception of vowel identity. Although there are additional studies examining effects of prosodic context

on formant values, most of these studies conduct acoustic analysis as a supplement to an articulatory analysis. More recently, in his articulatory and acoustic study, Cho (2005) showed that the accented vowels have distinct formant patterns which reflect the spatial displacement of the lips and the jaw. In producing a vowel in an accented (prominent) syllable, the jaw is more open regardless of vowel height, front vowels are more fronted, and back vowels are more back. Therefore, the accented high front vowel /i/ has higher F_2 s, and the accented low back vowel /ɑ/ has lower F_2 s, while both vowels have higher F_1 than the unaccented counter parts.

4.2 Acoustic analyses

4.2.1 Measurements

A variety of acoustic measures were extracted in order to explore the acoustic correlates of perceived prominence and their contribution to prosody perception. The acoustic measures analyzed in this study are: duration (ms), overall intensity (dB), bandpass filtered intensities (so-called spectral balance) in four frequency bands (dB), three measures of F_0 s (Hz), and formant frequencies (Hz). As described in Chapter 3, all acoustic measures were extracted from lexically stressed vowels so that lexical stress information remained constant and the effects of prosodic prominence could be isolated on top of lexical stress effects. The lexical stress information is hand labeled in consultant with the Merriam-Webster Online dictionary.

The phone and word level transcriptions that were originally created by automatic forced phone alignment and after by manual corrections include some misalignments. As pre-processing, I automatically checked time-alignment and if the misalignment between word and phone level transcription is greater than 5ms, then I manually corrected transcriptions in accordance with sound files.

4.2.1.1 Fundamental frequency (F_0)

Raw pitch values were automatically extracted at every 1 ms using the autocorrelation method in Praat (Boersma and Weenink, 2005) with two separate settings for male and female speakers for pitch floor (for male speakers 50 Hz, and for female speakers 75 Hz), and pitch ceiling (for male speakers 350 Hz, and for females 450 Hz). The extracted pitch values in the vicinity of the border of a consonant and a vowel (a 30 ms window) were smoothed by removing micro-perturbation with a median filter in which a midpoint value within a comparison window is replaced by a median value over pitch values at 13 points. Lastly, pitch values of all other voiceless parts such as stop closure or VOT were obtained by interpolating pitch values of the preceding and the following segments in order to obtain the pitch contour over the course of the whole utterance. After obtaining the pitch contour over a whole utterance, the pitch values were normalized within a 400 ms window. Then the following measures of F_0 were employed for the acoustic analyses: the maximum value of F_0 within each stressed vowel, F_0 at right edge of the stressed vowel, and the onset and the offset F_0 slope of the stressed vowel.

4.2.1.2 Other acoustic measures

Vowel duration was measured as follows. Vowel durations (in ms) from the onset to the offset of stressed vowels were automatically extracted from each word in the speech excerpts, with vowel onsets and offsets determined by the phone-level transcription that is published with the corpus and later manually corrected by myself. Mean RMS overall intensities were also automatically extracted from the stressed vowels, measured in Pascal as pressure level units and then converted to dB as in equation 4.1. The bandpass filtered intensities were obtained in four different frequency regions by using Hanning bandpass filters: 0–500, 500–1000, 1000–2000, and 2000–4000 Hz.

$$dB = 20 \log_{10} \frac{P_{RMS}}{P_{Ref}} \quad (4.1)$$

The formant values at the midpoint of the stressed monophthongs, and at 10% and 90% of the stressed diphthongs, were automatically extracted. 5 formant frequencies were traced with two different ceilings of frequencies according to gender (5000 Hz for male speakers, and 5500 Hz for female speakers) at every 10 ms time step with a 25 ms window.

4.2.2 Normalization of the acoustic measures

Ladefoged and Broadbent (1957) argued that when a speaker produces a vowel, the speaker conveys his/ her anatomical, physiological, and sociolinguistic information as well as phonemic information. These speaker-related features are treated as unwanted variations that a listener must eliminate or normalize in speech perception (Pols et al., 1973). In later studies, therefore, investigators have employed various normalization methods, and the validity of these methods has been compared. Adank et al. (2004) shows that Labanov’s z -normalization, as in equation 4.2, is the best normalization procedure and was therefore adopted for the normalization procedures of this study.

$$z_{ijk} = \frac{x_{ijk} - \bar{x}_{ij}}{s_{ij}} \quad (4.2)$$

where z_{ijk} represents the normalized value of the k^{th} actual acoustic measure, x_{ijk} , of the phone (j) extracted from i^{th} speaker, and \bar{x}_{ij} and s_{ij} represent the mean and the standard deviation of the phone, j , from i^{th} speaker.

Following procedures set forth by Labanov cited in Adank et al. (2004), all the acoustic measures including duration, overall and subband intensities, and formant values by phone and by speaker have been normalized in the present study. Normalization for F_0 measures was conducted in a local window (syntagmatically) and

not within phone categories (paradigmatically) as with the other acoustic measures, because of the expectation that phone-based variation in F_0 is less than variation due to the local speech context. F_0 measures were normalized within a 400 ms window within a speaker, a window size which is similar to that used by Kochanski and his colleagues (2005; 2006). In summary, while F_0 measures were normalized within-utterance, within a moving window with a fixed window size, all other acoustic measures were normalized by phone within a speaker.

4.3 Results

In this section, results are presented from correlation and linear regression analyses of the acoustic measures and P-scores.

4.3.1 How closely is each acoustic measure related with the perception of prosodic prominence by ordinary listeners?

Spearman's non-parametric correlation analyses were performed between perceived prominence and acoustic measures from stressed vowels. The results are summarized in Tables 4.1 and 4.2. First, P-scores ($p < 0.001$) are significantly correlated with all acoustic measures except the offset slope of F_0 , in an analysis that pools together all the stressed vowels, at a 95% confidence interval. For these comparisons, the F_2 measures are excluded because of opposite predictions about the effect of prominence on F_2 measures depending on the front/backness of a given vowel. Second, looking closely, the correlation coefficients of P-scores with acoustic measures are all comparable to one another, although vowel duration ($\rho = 0.262$, $p < 0.001$) is the strongest correlate. In other words, there is no single dominant acoustic measure in terms of its correlation with P-scores. Correlation coefficients for acoustic measures other than

vowel duration range from 0.095 ($p < 0.001$) for the bandpass filtered intensity in 0–500 Hz, to 0.187 ($p < 0.001$) for the bandpass filtered intensity in 500–1000 Hz. Third, perceived prominence positively correlates with all acoustic measures other than the slope of F_0 from the local peak of F_0 to F_0 at the right edge of the vowel. That is, as P-scores increase, the values of the vowel duration, the local maximum of F_0 , F_0 at the right edge of the vowel, the onset slope of F_0 , overall and bandpass filtered intensities, and F_1 all increase. The following sections present detailed results of the correlation analyses with P-scores and each acoustic measure, by vowel.

4.3.1.1 Effects of prosodic prominence on the duration of the lexically stressed vowel

Results of these analyses showed that perceived prominence is strongly correlated with vowel duration as summarized in Tables 4.1 and 4.2. Spearman’s non-parametric correlation analyses demonstrate that durations of all stressed vowels are positively correlated with P-scores. It was further shown that durations of all the stressed vowels other than the monophthongs / α / ($\rho = 0.080$, $p = 0.152$) and / v / ($\rho = 0.014$, $p = 0.452$), and the diphthong, / av / ($\rho = 0.222$, $p = 0.061$), are significantly correlated with P-scores. Positive correlations between P-scores and vowel duration reveal that in words that are perceived as prominent by ordinary listeners exhibit stressed vowels with longer duration than the stressed vowels of words that are not perceived as prominent.

As discussed above, significant correlations of P-scores with vowel duration were not found for all the lexically stressed vowels: the vowels / v / and / av / did not show significant correlations between vowel duration and P-scores. Yet, as shown in Table 3.1, the vowels / v / and / av / have the lowest token frequencies in the corpus, and this may be at least partly responsible for the non-significant correlation. It is possible that with a larger amount of data, significant correlations between P-

Vowels		All	ɑ	æ	ʌ	ɔ	au 10%	au 90%	aɪ 10%	aɪ 90%	ɛ
Duration	Coeff.	0.262	0.080	0.369	0.260	0.229	0.222		0.380		0.338
	<i>p</i>	< 0.001	0.152	< 0.001	< 0.001	0.006	0.061		< 0.001		< 0.001
F_1	Coeff.	0.173	0.106	0.226	0.357	0.280	0.163	-0.132	0.026	-0.037	0.281
	<i>p</i>	< 0.001	0.083	< 0.001	< 0.001	0.001	0.129	0.188	0.329	0.261	< 0.001
F_2	Coeff.	N/A	-0.185	-0.073	-0.147	-0.161	0.282	-0.303	-0.150	-0.023	-0.133
	<i>p</i>	N/A	0.007	0.110	0.002	0.040	0.024	0.019	0.005	0.348	0.003
Overall RMS intensity	Coeff.	0.140	0.144	0.067	0.105	0.046	0.222		0.229		0.163
	<i>p</i>	< 0.001	0.032	0.130	0.020	0.311	0.060		< 0.001		< 0.001
SB (0–500 Hz)	Coeff.	0.095	0.065	-0.022	0.028	-0.118	0.204		0.159		0.137
	<i>p</i>	< 0.001	0.202	0.354	0.293	0.101	0.078		0.003		0.002
SB (500–1000 Hz)	Coeff.	0.187	0.161	0.085	0.260	0.176	0.263		0.226		0.252
	<i>p</i>	< 0.001	0.019	0.076	< 0.001	0.027	0.033		< 0.001		< 0.001
SB (1000–2000 Hz)	Coeff.	0.159	0.201	0.146	0.262	0.043	0.280		0.236		0.228
	<i>p</i>	< 0.001	0.005	0.007	< 0.001	0.320	0.024		< 0.001		< 0.001
SB (2000–4000 Hz)	Coeff.	0.156	0.075	0.120	0.167	0.046	0.192		0.216		0.244
	<i>p</i>	< 0.001	0.169	0.022	< 0.001	0.310	0.091		< 0.001		< 0.001
$F_{0,max}$	Coeff.	0.143	0.194	-0.006	0.129	0.081	0.083		0.241		0.112
	<i>p</i>	< 0.001	0.006	0.460	0.005	0.192	0.284		< 0.001		0.010
Right F_0	Coeff.	0.172	0.156	0.150	0.191	0.214	-0.019		0.224		0.150
	<i>p</i>	< 0.001	0.022	0.006	< 0.001	0.010	0.447		< 0.001		0.001
Onset slope	Coeff.	0.097	0.253	0.253	0.071	0.135	0.217		0.197		0.075
	<i>p</i>	< 0.001	< 0.001	< 0.001	0.081	0.071	0.065		< 0.001		0.059
Offset slope	Coeff.	0.007	0.090	0.090	0.045	0.098	-0.107		-0.012		0.023
	<i>p</i>	0.296	0.065	0.065	0.191	0.145	0.230		0.420		0.315

Table 4.1: Summary of the Spearman’s non-parametric correlation analyses between P-scores and acoustic measures of each vowel (I). Spearman’s ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.

Vowels		ɜ	ei 10%	ei 90%	ɪ	i	ou 10%	ou 90%	ʊ	u
Duration	Coeff.	0.232	0.376		0.228	0.272	0.156		0.014	0.210
	<i>p</i>	0.006	< 0.001		< 0.001	< 0.001	0.016		0.452	0.003
F_1	Coeff.	0.331	0.062	0.031	0.190	0.115	0.043	-0.096	0.304	0.146
	<i>p</i>	< 0.001	0.190	0.333	< 0.001	0.023	0.282	0.098	0.005	0.026
F_2	Coeff.	-0.138	0.102	0.147	-0.014	0.188	-0.145	-0.074	-0.001	-0.164
	<i>p</i>	0.069	0.072	0.019	0.378	< 0.001	0.024	0.158	0.497	0.015
Overall RMS intensity	Coeff.	0.225	0.065		0.090	0.121	0.080		0.315	0.073
	<i>p</i>	0.008	0.178		0.025	0.018	0.140		0.004	0.166
SB (0–500 Hz)	Coeff.	0.181	0.043		0.064	0.111	0.106		0.282	0.057
	<i>p</i>	0.026	0.248		0.084	0.027	0.075		0.008	0.224
SB (500–1000 Hz)	Coeff.	0.311	0.048		0.152	0.106	0.050		0.340	0.159
	<i>p</i>	< 0.001	0.248		< 0.001	0.032	0.248		0.002	0.017
SB (1000–2000 Hz)	Coeff.	0.290	0.047		0.161	0.052	-0.043		0.352	0.023
	<i>p</i>	0.001	0.252		< 0.001	0.184	0.278		0.001	0.381
SB (2000–4000 Hz)	Coeff.	0.008	0.172		0.122	0.207	0.029		0.222	-0.049
	<i>p</i>	0.464	0.007		0.004	< 0.001	0.349		0.030	0.258
$F_{0,max}$	Coeff.	0.161	0.107		0.211	0.214	0.022		0.212	0.114
	<i>p</i>	0.042	0.062		< 0.001	< 0.001	0.384		0.037	0.066
Right F_0	Coeff.	0.240	0.158		0.192	0.197	0.141		0.260	0.113
	<i>p</i>	0.005	0.011		< 0.001	< 0.001	0.027		0.014	0.068
Onset slope	Coeff.	0.030	0.170		0.059	0.090	0.129		0.235	0.139
	<i>p</i>	0.377	0.007		0.099	0.058	0.040		0.023	0.032
Offset slope	Coeff.	-0.058	0.034		-0.006	-0.117	-0.027		-0.205	-0.095
	<i>p</i>	0.267	0.315		0.448	0.021	0.359		0.042	0.103

Table 4.2: Summary of the Spearman’s non-parametric correlation analyses between P-scores and acoustic measures of each vowel (II). Spearman’s ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.

scores of words containing these stressed vowels and vowel duration would arise. The correlation between P-scores and vowel duration is significant for the diphthongs /ov/ ($\rho = 0.156$, $p = 0.016$) and /aɪ/ ($\rho = 0.380$, $p < 0.001$), but there is no significant correlation for the open vowel /ɑ/, and in the latter case the lack of correlation cannot be explained due to the sparseness of the current data. Furthermore, it seems unlikely that the lack of a correlation between P-scores and duration for the open vowel /ɑ/ is due to a ceiling effect—this vowel is intrinsically long—because we don't see such ceiling effects for all long vowels, including diphthongs. The absence of this correlation for /ɑ/ is left here for future research.

4.3.1.2 Effects of prosodic prominence on the intensity measures of the lexically stressed vowel

Spearman's non-parametric correlation analyses reveal that overall intensity and sub-band intensities are also positively correlated with the P-scores for many stressed vowels. P-scores of all stressed vowels except the diphthong /ov/ are positively correlated with at least one intensity measure. Overall intensity is positively correlated with P-scores for 8 out of 14 stressed vowels, /ɑ, ʌ, aɪ, ε, ɜ, ɪ, i, u/, with correlation coefficients ranging from 0.090 ($p = 0.025$) for /ɪ/ to 0.315 ($p = 0.004$) for /u/. Among the measures of bandpass-filtered intensity, the subband intensity in 500–1000 Hz is significantly correlated with P-scores for 11 out of 14 stressed vowels (all except /æ, eɪ, ov/), while the correlation for subband intensity in 1000–2000 Hz is significant for 9 out of 14 vowels (all except /ɔ, eɪ, i, ov, u/), and in 2000–4000 Hz the correlation is significant for 8 out of 14 vowels (all except /ɑ, ɔ, av, ɜ, ov, u/). On the other hand, the bandpass filtered intensity in 0–500 Hz shows significant correlations with P-scores for only 5 out of 14 stressed vowels (all except /ɑ, æ, ʌ, ɔ, av, eɪ, ɪ, ov, u/). For these vowels the correlation coefficient (ρ) range from 0.106 ($p = 0.032$) with sub-band intensity in 500–1000 Hz for /i/, to 0.352 ($p = 0.001$) with sub-band

intensity in 1000–2000 Hz for /v/.

These results show that stressed vowels in words perceived as prominent by ordinary listeners have higher intensities than stressed vowels in non-prominent words. The results also demonstrate that the strength of the correlations between P-scores and the intensity measures is generally weaker than the strength of the correlations of P-scores with vowel duration. Looking at the results from the intensity measure in each subband, the intensity measure taken from the frequency band in 500–1000 Hz is most consistently significantly correlated with perceived prominence, followed by the frequency band in 1000–2000 Hz, and then 2000–4000 Hz. The subband intensity in 0–500 Hz does not show a significant correlation with P-scores with many vowels. The findings from Spearman’s non-parametric correlation analyses suggest that the bandpass filtered sub-band intensities in 0.5–2 kHz, which span the region of the first formant, are amplified as P-scores increase. The energy in the frequency band in 0–500 Hz, which is below the range of the first formant frequency for most vowels, is not affected by the presence or absence of perceived prosodic prominence. In sum, words that are perceived as prominent tend to have stressed vowels that are louder and that have enhanced subband intensities in mid-and high frequency regions, corresponding to the first and the second formant frequency bands for most vowels.

4.3.1.3 Effects of prosodic prominence on the fundamental frequency measures of the lexically stressed vowel

In the current study, the following F_0 measures were extracted: the local F_0 maximum, F_0 at the right edge of each target vowel, and the onset and the offset slope of F_0 within the target vowel. Results of Spearman’s non-parametric correlation analyses show that fundamental frequency measures are correlated with perceived prosody. First of all, P-scores in all stressed vowels other than the diphthong /av/ are significantly correlated with one of the F_0 measures, and the statistically significant correlation

coefficients (ρ) of P-scores with various measures of fundamental frequency are all positive. Additionally, comparing the correlation results of each F_0 measure with P-scores, the local maxima of F_0 are positively correlated with P-scores in many vowels (9 out of 14 except /æ, ə, eɪ, oʊ, u/), and the correlation coefficients that are statistically significant range from 0.112 ($p = 0.010$) for the local F_0 maximum of /ɛ/, to 0.241 ($p < 0.001$) for the local F_0 maximum of /ɑ/. The fundamental frequencies at the right edge of the stressed vowels show significant correlations with P-scores for a greater number of stressed vowels than the local F_0 maxima (12 out of 14 except /aʊ, u/), whose correlation coefficients range from 0.141 ($p = 0.027$, /oʊ/), to 0.260 ($p = 0.014$) for F_0 measured at the right edge of the vowel, /ʊ/. Regarding the slopes of F_0 within the stressed vowels, the slopes of F_0 measured from the onset to the local peak of F_0 , but not the slopes of F_0 measured from the local peak to the offset of F_0 (2 out of 14 except /ɑ, æ, ʌ, ə, aʊ, ɑɪ, ɛ, ɜ, eɪ, ɪ, oʊ, u/), are significantly correlated with P-scores in half of the stressed vowels (7 out of 14 except /ʌ, ə, aʊ, ɛ, ɜ, ɪ, i/).

In summary, many of the stressed vowels perceived as prominent have a higher F_0 maximum and a higher F_0 at the right edge of the vowel as well as a more positive onset slope of F_0 than non-prominent stressed vowels. Among many F_0 measures, the local peak of F_0 is the most reliable acoustic correlate of perceived prominence. Yet, in terms of the strength of the correlations as well as the number of the vowels showing a significant correlation, the fundamental frequency measures are less strongly correlated with P-scores than the vowel duration or the intensity measures in this corpus.

4.3.1.4 Effects of prosodic prominence on the formant frequency measures of the lexically stressed vowel

Spearman's non-parametric correlation analyses demonstrate that for many of the stressed vowels, P-scores are significantly correlated with first and second formant

values. Let us first discuss the results of the correlation analyses between P-scores and the first formant measures. Nine out of the fourteen lexically stressed vowels have higher F_1 values in words with high P-scores compared to the same vowels in words with lower P-scores. Notably, the vowels that show statistically significant correlations with F_1 are all monophthongs—all except the low back vowel /ɑ/ show significant positive correlations of P-scores with F_1 . No diphthong, on the other hand, shows significant correlations between P-scores and F_1 . The correlation coefficients (ρ) for all vowels are uniformly positive, although not all the correlation coefficients are statistically significant, and the correlation coefficients that are statistically significant range from 0.115 ($p = 0.023$) for /i/, to 0.357 ($p < 0.001$) for /ʌ/.

Unlike F_1 , the F_2 measures of some stressed vowels are positively correlated with P-scores, and other vowels have lower F_2 values than non-prominent stressed vowels. With monophthongs, the results show that 6 out of 10 stressed monophthongs demonstrate significant correlations with P-score and F_2 values. Looking closely, the front high vowel /i/ has a higher F_2 in words that are perceived as prominent ($\rho = 0.188$, $p < 0.001$), but other monophthongs, including /ɑ, ʌ, ɔ, ε, u/, have lower F_2 in words that are perceived as prominent than their counterparts in non-prominent words. The correlation coefficients (ρ) that are significant range from -0.185 ($p = 0.006$, /ɑ/) to -0.133 ($p = 0.002$, /ε/). As for diphthongs, the results show that some of the second formant values measured either at 10% (nucleus) or at 90% (offglide) of the vowel in all four diphthongs are significantly correlated with P-scores. The nuclei of two diphthongs /aɪ, oʊ/ tend to have lower F_2 ($\rho = -0.150$, $p = 0.005$ for /aɪ/ and $\rho = -0.145$, $p = 0.024$ for /oʊ/), but the nucleus of /aʊ/ has higher F_2 ($\rho = 0.282$, $p = 0.024$) when perceived as prominent compared with their non-prominent counterparts. The offglides of two diphthongs also show significant correlations with F_2 measures: the offglide of /aʊ/ has lower F_2 ($\rho = -0.303$, $p = 0.019$), but that of /eɪ/ has higher F_2 ($\rho = 0.147$, $p = 0.019$) when perceived as prominent compared with the corresponding

non-prominent vowels.

In sum, the first and the second formant values are significantly correlated with perceived prominence. F_1 values of stressed monophthongs tend to be uniformly higher in words that are perceived as prominent, while no diphthong shows any significant correlations between prominence and F_1 . Regarding F_2 , the front-most vowel, /i/ has a higher F_2 , while most back vowels and non-peripheral monophthongs have a lower F_2 in words perceived as prominent. Diphthongs also show significant correlations between P-scores and F_2 values measured either at 10% or at 90% of the vowel.

4.3.1.5 Summary of the findings from the Spearman's non-parametric correlation analyses

Findings from Spearman's non-parametric correlation analyses illustrate that the perception of prosodic prominence by ordinary listeners is closely associated with systematic changes in the acoustic signal including vowel duration, intensity measures, F_0 measures, and formant measures. In words that are perceived as prominent stressed vowels tend to have longer durations, higher intensities, higher fundamental frequencies, and formant structures that reflect more peripheral and more open articulations. In terms of the relative strength of each acoustic measure's correlation with P-scores, the results suggest that all the acoustic measures are comparably strong correlates of P-scores, although the first formant of the diphthongs is not a reliable P-score correlate. However, there is no single acoustic measure that is consistently correlated with perceived prominence across all stressed vowels, and there is no single lexically stressed vowel for which P-scores are significantly correlated with a single acoustic measure. In some vowels, including three diphthongs (/au, ei, ov/) and two monophthongs (/ɔ, u/), very few acoustic measures are significantly correlated with P-scores, while many or all the acoustic measures show significant correlations with P-scores

in other vowels.

4.3.2 How much do different acoustic measures contribute to the ordinary listeners' perception of prosodic prominence?

In the previous section, acoustic variation in the speech signal was shown to be associated with prosodic prominence as produced by the speaker and perceived by the ordinary listener. This section evaluates to what extent the combined changes in acoustic patterns contribute to listeners' perception of prosodic prominence, and, furthermore, to determine the contribution of individual acoustic measures to listeners' perception of prosodic prominence.

Since acoustic measures are possibly interrelated with one another, we first tested whether any acoustic information in the speech signal is redundant in signaling prosodic prominence, or whether acoustic measures co-vary synergistically to cue prominence perception. The total sum of the coefficient of determination (r^2) from a series of linear regression analyses between P-scores and a single acoustic measure was calculated to project the total sum of variation in listeners' perception of prosody which each acoustic measure can account for. A simple multiple linear regression analysis between P-scores and the combined set of acoustic measures models the total variation in listeners' perception of prosody as explained by acoustic measures. Three possible predictions can be established: If the set of acoustic measures are not interrelated and instead are independent from one another, then the total sum of r^2 from a series of linear regression between perceived prominence and each acoustic measure will be the same as r^2 from the linear regression between perceived prominence and all the acoustic measures together. If acoustic measures are collinear and any acoustic information in the speech signal is redundant, then the former should be greater than

the latter. On the other hand, if a subset of the acoustic parameters work together to signal prosodic prominence, then the former should be smaller than the latter.

The results show that the total sum of this variation is greater than the total variation obtained from simple multiple linear regression analyses, suggesting that acoustic measures as predictors of perceived prominence are interrelated, and that some information obtained from the acoustic measures is redundant for ordinary listeners. Only in the regression models of perceived prominence of two vowels, /ɔ, ov/, is the summation of variation explained by each single acoustic measure smaller than the total variation explained by all acoustic measures altogether. The results confirm that the individual acoustic measures are interrelated and contain redundant information for predicting prosodic prominence. The acoustic cues present in the vowels /ɔ, ov/ work together to signal prosodic prominence, and there is no vowel where acoustic measures are independent and not interrelated in signaling prosodic prominence. These findings further suggest that perceived prosodic prominence cannot be modeled by looking at the patterns of any single acoustic measure. Instead, modeling perceived prosodic prominence requires that the patterns of all the acoustic measures together in the speech signal be taken into account.

Since it is shown that the acoustic measures are interrelated with one another, multiple linear regression analyses were performed to model perceived prosodic prominence. Figure 4.1 summarizes the results of simple multiple linear regression analyses of perceived prominence and boundary, illustrating the total variation of prosody perception that can be explained from the combination of the acoustic measures. Overall, around 12% of prominence perception is explained on the basis of changes in the acoustic patterns. Looking at the multiple linear regression models by vowel, the acoustic variation in the combination of duration, intensity, fundamental frequency, and formants of the lexically stressed vowel can explain from around 15% (/ov/) to 40% (/av/) of the variation of listeners' prominence perception. Looking closely, the

total variation of prosodic prominence in the regression models of all vowels other than the vowel /av/ is smaller than 30%, suggesting that there must be other factors influencing ordinary listeners' perception of prosodic prominence.

Subsequent stepwise multiple linear regression analyses were performed to examine which acoustic measures should be included to best model perceived prominence, and how much each acoustic measure contributes to predicting perceived prominence. The results summarized in Figure 4.2 reveal that modeling the perception of prosodic prominence requires a large number of acoustic measures. A single acoustic measure is included in the regression models of perceived prosodic prominence of three vowels (intensity measure for /ɑ/, temporal measure for /ɔ/, and F_0 measures for /ov/), while perceived prominence of other vowels is modeled by a combination of more than two acoustic measures. In terms of the contribution of each acoustic measure to modeling perceived prominence there is no dominant or primary acoustic measure that appears as a significant cue across all stressed vowels. Looking closely, perceived prominence in all the vowels except /ɑ, ɔ, ov/ is modeled on the basis of a combination of more than one acoustic measure: intensity measures for /ɑ/, vowel durations for /ɔ/, and F_0 measures for /ov/. Vowel duration is included in the multiple linear regression models of P-scores in 12 out of 14 lexically stressed vowels excluding /ɑ, ov/, F_0 measures are included in 10 out of 14 vowels, excluding /ɑ, ɔ, ɜ, av/, intensity measures are included in 8 out of 14 vowels, excluding /æ, ɔ, ʊ, u, eɪ, ov/, and formant frequency measures are included in 7 out of 14 vowels, excluding /ɑ, ɔ, ɪ, u, av, eɪ, ov/.

To summarize, modeling perceived prosodic prominence requires information from a combination of acoustic measures. In terms of the relative strength of each acoustic measure as a cue for prosodic prominence, the contribution of each acoustic measure to the perception of prominence is comparable. Although there is no single dominant acoustic measure to model the perception of prosodic prominence, vowel duration

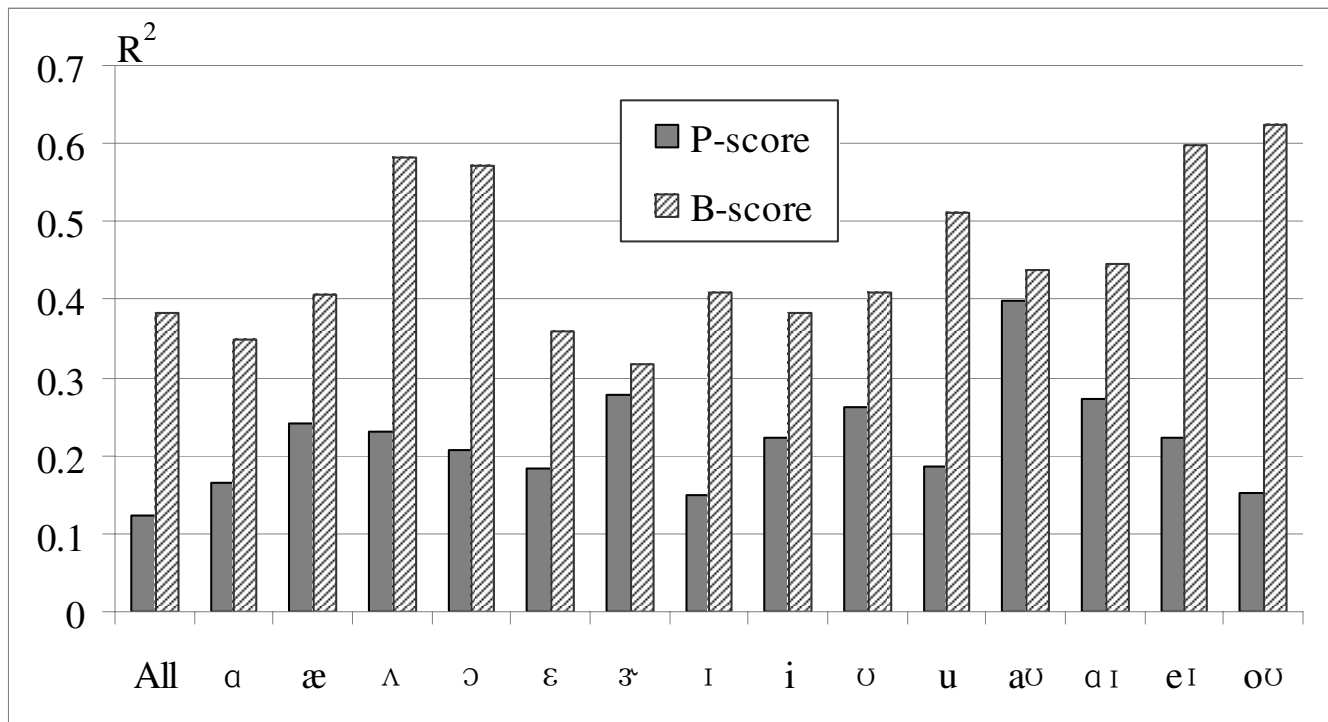


Figure 4.1: The distribution of the total variation (R^2) of the ordinary listeners' perception of prosodic prominence (oblique bars) and prosodic boundary (dotted bars)

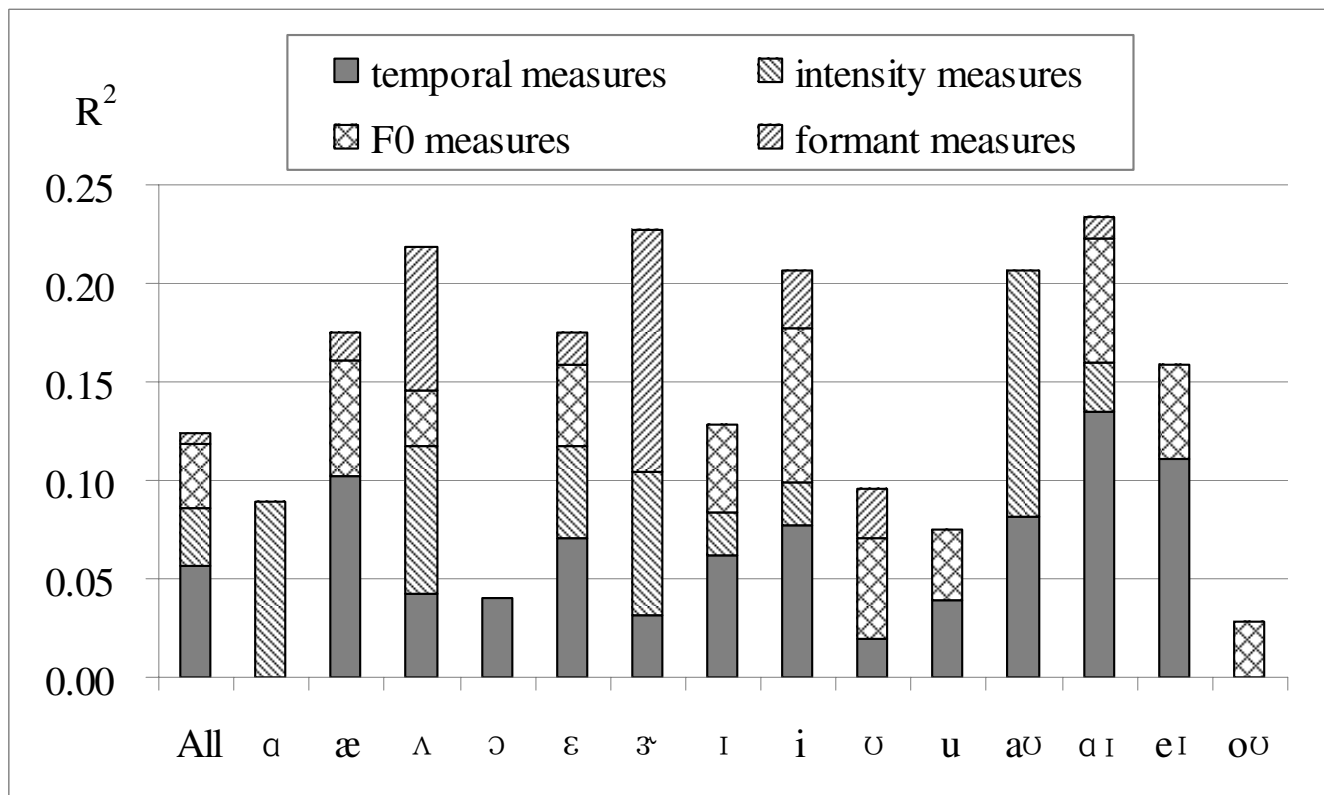


Figure 4.2: The distribution of the variation (R^2) in the ordinary listeners' response to prosodic prominence predicted by stepwise multiple linear regressions of the acoustic measures and P-scores

is included in the largest number of the regression models of P-scores, suggesting that vowel duration is a primary cue for prominence. Yet, there still remains much unexplained variability in ordinary listeners' responses to prosodic prominence, when only taking into account acoustic information in the speech signal. Therefore, the effects of other factors including syntactic structures, word frequency, word repetition in discourse, and default prosody must be examined to fully understand the nature of prosodic prominence, as will be discussed in later sections.

4.4 Summary and Discussion

The results from this study demonstrate that ordinary listeners' perception of prosodic prominence is correlated with and can further be modeled by systematic changes in the patterns of the acoustic characteristics of the lexically stressed vowels, and that these acoustic measures are interrelated with one another. The main acoustic characteristics of perceived prominence are as follows. In words judged to be prominent, all the acoustic measures significantly change in the direction of enhancing the phonetic characteristics of the lexically stressed vowels when compared to the phonetic characteristics of non-prominent counterparts. In other words in words perceived as prominent, lexically stressed vowels have longer vowel durations, greater overall and subband intensities, higher local peaks of fundamental frequency, and greater onset F_0 slopes. As for F_1 measure, although no effect of prosodic prominence on F_1 was found for the diphthongs, there were systematic pattern changes in regard to perceived prominence for monophthongs. All monophthongs except the open vowel /ɑ/ have higher F_1 values when heard as prominent than when not prominent, reflecting a more open vocal tract under prominence. However, the findings from Spearman's non-parametric correlation analyses between the F_2 measures and P-scores do not seem to be systematic at the first glance because F_2 values of one monophthong /i/

and two diphthongs /au, eɪ/ are positively correlated with perceived prominence, and many other monophthongs /ɑ, ʌ, ɔ, ɛ, u/ and three diphthongs /aʊ, ɔɪ, oʊ/ are negatively correlated with perceived prominence. Yet, looking closely, there are additional notable systematic findings. As for F_1 , it was shown that regardless of intrinsic vowel height, prosodic prominence identified by ordinary listeners is positively correlated with F_1 , suggesting that monophthongs have a more open vocal tract when perceived as prominent than when not. On the other hand, as for F_2 , the front high vowel /i/ has higher F_2 values, while other monophthongs have lower F_2 values. This suggests that three peripheral vowels /i, u, ɑ/ which are at the corner of the vowel space tend to be produced in an expanded vowel space in words perceived as prominent: the front vowel /i/ tends to be more front, the back high and low vowels, /u/ and /ɑ/ are likely to be more back. Other vowels that are not peripheral tend to be more back when heard as prominent. The findings about the correlations between formant values and P-scores in diphthongs showed that the phonetic distance from the nucleus to the offglide of the diphthongs in front/back dimension tends to increase under prominence. In summary, all the acoustic characteristics including segmental as well as suprasegmental features in spontaneous conversational speech of American English are more phonetically enhanced when identified as prominent by ordinary listeners than when they are not prominent.

Although there are strong systematic correlations between acoustic measures and perceived prominence, the results also show that there is no single acoustic measure that is significantly correlated with and is a cue for perceived prominence across all the lexically stressed vowels. Instead, prosodic prominence in spontaneous conversational speech of American English is signaled through changes in a combination of various acoustic measures, on the basis of which ordinary listeners' perception of prosodic prominence can be modeled.

In terms of the relative strength of the acoustic measures as the acoustic cor-

relates and cues for perceived prominence, the results show that vowel duration is most strongly and consistently correlated with prosodic prominence and most reliably acts as a cue for prosodic prominence, and that the other acoustic measures are also correlated with perceived prominence and contribute to listeners' perception of prosodic prominence. Even vowel duration does not emerge as a reliable acoustic correlate of prosodic prominence across all the lexically stressed vowels, and does not account for 100% of the variability in listeners' perception of prosodic prominence. Instead, the acoustic measures co-vary and interact with one another to signal prosodic prominence in spontaneous conversational speech of American English, and to guide ordinary listeners' identification of prosodic prominence. For example, all monophthongs other than the vowel /a/ have higher F_1 values in words perceived as prominent, reflecting more open articulation. This larger opening of the mouth for these monophthongs requires more time, subsequently lengthening vowel duration. Therefore, both the lengthened vowel durations and the increased F_1 values work together to help listeners to indicate prosodic prominence in many multiple linear regression models of perceived prominence for lexically stressed vowels. There is no model in which only formant measures, not vowel duration measures, are included for the regression models. This also explains why the correlation of P-scores of the open vowel /a/ with the vowel duration was found not to be significant. The vowel /a/ is intrinsically open with the lower jaw position, and thus does not have room for more lowering under prominence, which makes neither long vowel duration nor higher F_1 values a significant correlate of prosodic prominence and further an acoustic cue for prosodic prominence. For the vowel /a/, changes in the other acoustic measures including F_2 , intensity, and F_0 are related to the listeners' perception of prosodic prominence and stepwise linear regression demonstrates that changes in intensity measures, in particular subband intensities in 500–2000 Hz reliably account for the variability attested in the listeners' response to prosodic prominence.

On the other hand, as for diphthongs, vowel lengthening under prominence tends not to be related to mouth opening, but rather has to do with an increase of the phonetic distance between the nucleus and the offglide. The results demonstrate that all diphthongs are lengthened when heard as prominent, with the exception of the vowel /au/, and that all diphthongs increase the phonetic distance in the front/back dimension of the abstract vowel space when perceived as prominent, with the exception of the vowel /ou/. The vowel /au/ is longer when perceived as prominent with a marginal significance. In the vowel /ou/, there is a marginal lowering effect of F_1 values of the offglide ($\rho = -0.96$, $p = 0.98$), reflecting the increase of the phonetic distance between the nucleus and the offglide in the height dimension. Therefore, taking the findings altogether, it is suggested that the monophthongs and the diphthongs are governed by a different mechanism of vowel lengthening induced by prosodic prominence. As for monophthongs, vowel lengthening is due to larger mouth opening under prominence. On the other hand, for diphthongs, vowel lengthening results from an increase of transition time from the nucleus to the offglide in order to increase the phonetic distance between them.

The relationship between the formant and the intensity measures should also be noted. The overall intensity of the lexically stressed vowels was greater in words that are perceived as prominent than in words heard as non-prominent. In other words, prosodic prominence increases the energy of lexically stressed vowels. However, looking closely, the relative strength of the effects of prosodic prominence varies depending on the frequency bands of interest. The subband intensities in the mid and high frequency regions, 500–4000 Hz (11 out of 14 in 500–1000 Hz, 9 out of 14 in 1000–2000 Hz, and 8 out of 14 in 2000–4000 Hz) are significantly enhanced in a greater number of the lexically stressed vowels than subband intensities in the low frequency band (5 out of 14 in 0–500 Hz). The frequency regions in 500–2000 Hz are the frequency bands where the first (F_1) and the second (F_2) resonant frequencies for the majority

of vowels take place. This is especially apparent in 2 out of 3 vowels (/ou, eɪ/), where the subband intensity measures in 500–1000 Hz are not significantly correlated with P-scores. The mean F_1 values of male speakers are lower than 500 Hz, their F_2 values are higher than 1000 Hz, and in 4 out of 5 vowels (/ɔ, eɪ, i, ou/), where the subband intensity measures in 1000–2000 Hz are not significantly correlated with P-scores, the mean F_2 values of one gender are too close to or out of the cutoff frequency values. This strongly suggests, therefore, that the energy containing the frequency bands of the first and the second formants as well as the frequency values of the first and the second formants also increase to a greater extent under prosodic prominence than the energy in other frequency bands. However, in order to examine this relationship more closely, there are at least two things that must be taken into account for the selection of frequency bands for energy measurements: gender and phone type. In addition to the fact that different vowels have different formant values depending on their locations in the vowel space, females always have higher resonant frequencies due to the small size of their vocal tract, and the values of the first and the second resonant frequency are greatly influenced by the shapes of the articulators and the locations of the constriction in the vocal tract. However, further investigation will be reserved for another study.

It was also shown that the F_0 measures of many lexically stressed vowels are significantly correlated with and contribute to ordinary listeners' perception of prosodic prominence. Like other acoustic measures taken in the current study, the findings regarding the relationship between F_0 measures and prosodic prominence are in part compatible with the findings or descriptions from many prior studies (Cooper et al., 1985; Eady and Cooper, 1986; Ladd, 2008; Lieberman, 1960; 't Hart et al., 1990; Silverman et al., 1992; Welby, 2003, among many others). The local maxima and onset slope of F_0 of many of the lexically stressed vowels increase when perceived as prominent by ordinary listeners. At the same time, however, the findings of this

study are not consistent with those from many prior laboratory studies because other lexically stressed vowels do not show any significant correlations between P-scores and the F_0 measures, and ordinary listeners' perception of prosodic prominence is not modeled by changes in the patterns of the F_0 measures in the stepwise multiple linear regression models of many lexically stressed vowels.

The failure to find significant F_0 predictors of perceived prominence may be partly due to the limitation of the types of prosodic prominence examined and the measurement types and domains in the current study. In the original ToBI system of American English based on metrical-autosegmental phonology, there are 5 different types of pitch accents in terms of combinations of a low and a high tone targets: H*, L*, L*+H, L+H*, and H+!H*. The different target tones will result in contradictory predictions of the F_0 measures. For example, the local F_0 maxima will be greater if a word carries H* to encode prosodic prominence, and it will be smaller if a word carries L* to encode prosodic prominence than if a word does not carry any pitch accent. As a result, the raising effect of the local peak of F_0 associated with prosodic prominence (H*) will be canceled out or at least diminished by the lowering effect of the local F_0 maxima (L*). Although H* is the most frequent pitch accent type in American English (Ananthakrishnan and Narayanan, 2008; Ross and Ostendorf, 1996), around 20% of pitch accents are of other types including L*, L*+H, L+H*, and H+!H*. In RPT employed in the current study, the ordinary listeners indicated the presence or the absence of prosodic prominence, but did not distinguish among types of prosodic prominence. Therefore, it is possible that in the results of the current study, perceived prominence includes all different types of prosodic prominence.

Another limitation is related to the selected measurement domains. In the current study, all the F_0 measures within a lexically stressed vowel were extracted: the local peak of F_0 within a lexically stressed vowel, F_0 at the right edge of the vowel, the onset slope of F_0 from the onset to the peak of the vowel, and the offset slope of F_0

from the peak to the offset of the vowel. However, although it is true that the local peak of F_0 is realized within the nucleus of the accented syllable, namely the lexically stressed vowel within a word, it is possible that the F_0 peak occurs before or after the lexically stressed vowel. The rising and falling movements begin earlier and end later than those of lexically stressed vowels. One cannot exclude the possibility that if F_0 measures were taken over a larger domain, then the F_0 measures would show significant correlations with P-scores in a larger number of the lexically stressed vowels, and would account for the greater variability in the ordinary listeners' responses to prosodic prominence.

Admitting the limitations of the current study as discussed above, its findings are also consistent with those from other prior corpus studies (Kochanski et al., 2005; Kochanski, 2006; Silipo and Greenberg, 1999, 2000). In their studies employing a large corpus of natural speech (the IViE corpus in the Kochanski et al. study and the Switchboard corpus in Silipo and Greenberg's study), the researchers demonstrated that fundamental frequency plays little role in the automatic assignment or classification of prosodic prominence. Therefore, the role of fundamental frequency measures as a primary acoustic correlate of or a primary cue for prosodic prominence must be further examined in future work.

4.5 Conclusion

Findings from Spearman's non-parametric correlation and multiple linear regression analyses show that prosodic prominence in the spontaneous conversational speech of American English is signaled through changes in a combination of various acoustic measures, and that ordinary listeners are sensitive to such changes in the acoustic patterns of speech and use them as cues by which to reliably locate prosodic prominence. Yet, there is no single acoustic measure to explain the relationship between the per-

ception of prosodic prominence and the acoustic measures. There still remains large variability within ordinary listeners' responses to prosodic prominence that is unexplained on the basis of acoustic patterns; other linguistic and non-linguistic factors must be investigated to fully understand the nature of prosodic prominence.

Chapter 5

An Acoustic Investigation of Prosodic Boundary

5.1 Introduction

Prosodic structures comprise hierarchically structured phonological domains of different sizes (syllable, foot, phrase, utterance) with prominence relationships among them. The prosodic structure of an utterance is encoded in phonetic patterns that mark the edges of prosodic units and prominent elements within those units. The previous chapter discussed the phonetic nature of prosodic prominence in spontaneous conversational speech in American English, as it is perceived by listeners. The primary objective of Chapter 5 is to investigate the phonetic nature of prosodic phrasing. More specifically, this section will focus on how prosodic phrase boundaries are perceived by listeners, and on the nature of the acoustic properties that cue prosodic phrase boundaries. The ultimate goal of this study related to prosodic phrasing is to develop a model of the acoustic basis of prosodic phrasing using linear regression methods.

The consensus among phonologists is that there are many different levels of prosodic phrasing, which are hierarchically structured (Beckman and Ayers, 1997; Jun, 1993, 1998, 2003; Nespor and Vogel, 1983, 1986; Pierrehumbert, 1980; Selkirk, 1984 among many others). Yet, there remains disagreement on the number and types of prosodic phrases among researchers, and for different languages. Recent analyses of American English prosodic phrase structure, agree in proposing at least two levels of prosodic phrasing between the prosodic word level and the utterance level:

intermediate and intonational phrases (Beckman and Ayers, 1997; Beckman and Pierrehumbert, 1986; Pierrehumbert, 1980; Pierrehumbert and Hirschberg, 1990). This proposal is also reflected in the conventions of the ToBI transcription system (Beckman and Ayers, 1997).

The approach to prosodic analysis pursued here, based on prosody annotations derived from Rapid Prosody Transcription, no attempt is made to distinguish between levels of prosodic phrases. Transcribers are asked to locate prosodic phrase edges (“chunk boundaries”) between words according to their real-time auditory impression of an utterance, and are not asked to discriminate between higher or stronger junctures and lower or weaker ones (Appendix A.2.1). Like the acoustic analyses conducted for perceived prominence, the acoustic analyses of perceived prosodic boundaries are performed with B-scores, which are assigned to each word appearing in the transcripts based on the proportion of ordinary listeners who marked the word as followed by a prosodic boundary, out of the total number of listeners. Chapter 5 will summarize the findings from correlation and regression analyses of prosodic boundaries and their associated acoustic measures.

5.1.1 Fundamental frequency (F_0)

As described above, in the widely accepted ToBI standard (Beckman and Ayers, 1997), there are two levels of phrasing above the prosodic word level in American English, which is defined according to the specific tunes of a speech utterance: an intermediate (ip) and intonational phrase (IP). Both phrases are marked by particular F_0 contours; the intermediate phrase is marked by a simple high (H) and low (L) tone “phrase accent”, while the intonational phrase is marked by an additional high (H) and low (L) “boundary” tone. In other words, by definition, prosodic phrases, and, in particular, the edges of both intermediate and intonational phrases are demarcated by F_0 contours derived from the specified phrase accents and boundary tones. Unlike the

ToBI standard, the IPO system, discussed in Ladd (2008), operates with the notions of “tune” and “relative prominence” defined by pitch contours (rising-falling for tune, and strong-weak and weak-strong for relative prominence). Ladd describes speech utterances as prosodically structured and demonstrates that these prosodic structures determine the distribution of tune and relative prominence in utterances. In other languages including Korean and Japanese, prosodic analysis in terms of hierarchically organized prosodic phrases, similar to English’s ToBI, is proposed. That is, in such languages, prosodic phrasing is described as marked by pitch contours at phrase edges. Beckman and Jun (1996) proposes K-ToBI for Korean and Venditti (2006) proposes J-ToBI for Japanese, with prosodic or intonational phrases demarcated by systematic phonetic patterns including delimitative tones and shifts in pitch range. They further propose that there are two levels of prosodic phrasing in both languages, the accentual phrase (AP) and intonational phrase (IP).

The abstract phonological prosodic phrases are phonetically implemented- in most cases with edge-marking F_0 contours. Prior studies of prosody have directly investigated how F_0 is manifested in relation to prosodic phrasing, whether F_0 variation is a cue for prosodic phrasing, and whether F_0 patterns reflecting differences in the level of prosodic phrase boundary, if any exist, can be utilized in phrase classification or automatic phrase detection in many languages (Aguilar et al., 2009; Bruce et al., 1993; Carlson et al., 2005; Carlson and Swerts, 2003a,b; Kim et al., 2006; Ferrer et al., 2002; Wagner, 2010). In a series of perception experiments conducted in Swedish, Bruce et al. (1993) showed that a deep F_0 valley in the downward trend of F_0 functions to signal a phrase boundary, and that listeners rely on F_0 as well as on segmental duration to determine phrase boundaries. Carlson and Swerts (2003a) also evaluated the role of F_0 features such as the median F_0 value of the last 50 ms voiced region of a word, as perceptual cues to upcoming prosodic breaks in a perception study in Swedish. They found a significant though small correlation effect between F_0 mea-

tures and boundary strength as judged by listeners, suggesting that F_0 measures cue not only for the presence or absence of a prosodic break but also the relative strength of it.

In their cross-linguistic study of automatic phrase boundary detection, Vicsi and Szaszak (2006) found that in both Hungarian and Finnish, the time series of fundamental frequency together with energy results in the best detection rate for phrase boundaries, while syllable length does not largely influence such detection. Likewise, Aguilar et al. (2009) showed that the F_0 values measured in the last sonorant before a prosodic break and the F_0 difference between the local maximum F_0 in the stressed syllable and F_0 measured immediately before a prosodic break can differentiate the levels of a prosodic break. However, the contribution of the F_0 measures relatively smaller than the contribution of temporal measures including silent pause duration and word-final syllable duration in Catalan. Further evidence for this comes from a study by Leemann et al. (2009), where they looked at whether different types of prosodic phrases (continuing vs. terminating phrase) differ in terms of F_0 contours in four different dialects of Swiss German.

Finally, there are also a great number of studies investigating the relationship between prosodic boundaries and various F_0 measures in American English. Chavarria et al. (2004) examined whether an F_0 drop can differentiate the intermediate phrase boundary from the higher intonational phrase boundary in spontaneous conversational speech, but did not find any significant effect. In another corpus study, however, Kim et al. (2006) compared various F_0 measures at the intermediate phrase boundary (ip) with those at the intonational phrase boundary (IP) in the Switchboard corpus of spontaneous conversational speech and the Boston University Radio News corpus, where professional FM radio news announcers read news stories. In both corpora, prosodic phrase boundaries are transcribed following the ToBI standard. It was found that the end F_0 values at the higher prosodic boundary (IP, L-L%) are

significantly lower than at the lower prosodic boundary (ip, L-).

The role of F_0 features as cues to prosodic phrasing was also investigated in numerous studies of sentence processing and prosodic disambiguation (Kang and Speer, 2002; Kjelgaard and Speer, 1999; Venditti, 2006, among others). In these studies, researchers manipulated F_0 features and silent pauses to signal prosodic phrase boundaries. For example, in Kjelgaard and Speer (1999), semantically ambiguous sentences were constructed which were disambiguated by prosodic phrase boundaries at one of two locations, and then presented to listeners to examine whether listeners interpret the sentence on the basis of the acoustic cues for prosodic phrasing, including a falling F_0 contour (L-L%) and silent pause. The results show that listeners do utilize acoustic cues to prosodic boundary in disambiguating such sentences.

Although findings from these studies yield significant insight into the relationship between prosodic phrasing and variation in F_0 measures, there are also some limitations to this kind of research. In studies in which prosodic phrases are labeled based on the ToBI standard, any significant results might be an artifact of prosodic labeling, namely, of the explicit reliance of the trained expert labelers on specific falling or downward pitch contours when judging the presence or absence of a prosodic phrase boundary and its level. Furthermore, studies investigating the role of prosodic phrasing in sentence processing typically manipulate the presence or absence and the duration of silent pauses as well as F_0 values in the stimuli presented to listeners, and do not usually attempt to control any possible lengthening effects in the vicinity of a prosodic boundary. To the extent that properties other than F_0 , including final rime duration or silent pause duration, contribute to listeners' perception of boundaries in a natural speech setting, these studies fail to capture the full complexity of prosodic processing. The role of F_0 measures as correlates of and as cues for prosodic phrase boundary will be revisited in the current study.

5.1.1.1 Pause

The presence of a pause after a prosodic phrase has long been considered a major correlate of prosodic phrasing (e.g., Ferreira, 1993; Hansson, 2003) and a major cue for prosodic phrases. In fact, the presence of such a pause was employed to manipulate the location of prosodic phrases in many prior studies of language processing (Kang and Speer, 2002; Kjelgaard and Speer, 1999; Kraljic and Brennan, 2005; Watson and Gibson, 2004). In other previous research, the relation between the presence/absence of a silent pause and its length on one hand, and prosodic phrasing on the other was investigated and the role of silent pauses in boundary perception was additionally evaluated (e.g., Aguilar et al., 2009 in Catalan; Carlson and Swerts, 2003a; Fant et al., 2003; Heldner and Megyesi, 2003; Horne et al., 1995; Strangert and Heldner, 1995; Strangert and Helnder, 1995 in Swedish; Krivokapic, 2007 in English; Sanderman, 1996 in Dutch; Lin and Fon, 2009; Yang, 2007 in Mandarin Chinese). For example, Horne et al. (1995) examine to what extent the amount of final lengthening and the duration of a silent pause are related to the perception of the relative strength of prosodic boundaries. The results show that as the relative strength of a perceived prosodic boundary increases (prosodic words < prosodic phrases < prosodic utterances), the duration of a silent pause increases, and, moreover, that the silent pause duration is most strongly tied to higher boundaries—namely, prosodic phrases and utterances, not prosodic words. Silent pause information is also utilized for the development of automatic speech recognizers (e.g., Bulyko and Ostendorf, 2001; Ogata et al., 2009). For example, Bulyko and Ostendorf (2001) achieved a 95.8% rate for automatic phrase boundary detection just by employing silent pause duration, which was implemented as a prosody prediction module of the speech synthesizer of a travel planning system.

The presence of a silent pause and its duration is, however, determined by a number of factors in addition to prosodic structures: syntactic and discourse structures,

speech rate, etc. (Gee and Grosjean, 1983; Krivokapic, 2007; Watson and Gibson, 2004 for syntactic structures and the length of prosodic phrases, Smith, 2004 for discourse structure, Trouvain and Grice, 1999 for speech rate, and Fant et al., 2003 for speaker). Watson and Gibson (2004) showed that the number of phonological phrases in the syntactic phrase preceding and following a word juncture, as well as the length of the flanking syntactic phrases, all affect the occurrence of an intonational phrase boundary at that location, and these factors are also strongly correlated with the occurrence of a silent pause. In a recent study by Krivokapic (2007), the likelihood that the presence of a silent pause is affected by prosodic structures and syntactic phrase length was investigated. For test items which vary in terms of the length and internal prosodic structure of the intonational phrases (IP) preceding and following the silent pause, the results showed that both the complexity of the internal structure and the length of the IP affect pause duration. Regarding speech rate, Trouvain and Grice (1999) found that when reading aloud, speakers increase the number of pauses in their utterance when slowing down, while the number of pauses is reduced when speakers speed up. The findings from Fant and his colleagues' (2003) demonstrate greater variability in pause duration within sentences than between sentences.

Another complication arises from the fact that the presence of a silent pause is neither a sufficient nor a necessary factor to signal the presence of a prosodic phrase. According to Yoon et al. (2007), only 40.6% (984 out of 2423) of phrase boundaries including both intermediate and intonational phrase boundaries are followed by a silent pause. The remaining 59.4% (1439 out of 2423) of phrase boundaries are not followed by a silent pause in the Boston University Radio News corpus. In a recent study by Aguilar et al. (2009), it was shown that the presence of a silent pause combined with the duration of the final syllable is the most relevant factor when determining whether a word is followed by a prosodic break or not. Aguilar further showed that the duration of a silent pause above a certain threshold (452

ms), only serves to discriminate among the levels of prosodic breaks. In other words, the presence of a silent pause which is shorter than 452 ms does not influence the discrimination of prosodic break levels. Therefore, further investigation of this topic may reveal the role of the presence or absence of a silent pause and its duration in the production and the perception of prosodic phrase boundaries and the interaction with other acoustic measures like F_0 and word duration.

5.1.1.2 Duration

It has been determined that the duration of at least some portion of a word is influenced by the presence/absence of a prosodic phrase boundary (e.g., Berkovits, 1994 in Hebrew; Heldner, 2003; Horne et al., 1995 in Swedish; Chavarria et al., 2004; Ferreira, 1993; Kim et al., 2006; Klatt, 1975, 1976 in English; Nakai et al., 2009 in Northern Finnish; Ueyama, 1999 in Japanese). Klatt (1975, 1976) found that stressed vowels in word-final syllables of phrase-final words are significantly longer in duration (about 30%, 40 ms) than vowels in words that are not phrase-final, and stressed vowels in non-final syllables of non-phrase-final words have the shortest duration. In a more recent controlled production study of Athenian Greek, Kainada (2007) compared the duration of pre-boundary vowels before five different levels of prosodic boundaries and found that durations increase as the boundary strength increases. Another controlled laboratory study found similar findings in Finnish. With phonemic length contrasts, Finnish has been claimed not to show preboundary lengthening (Nakai et al., 2009). Nakai et al. (2009) observed that the utterance-final consonant and vowels are significantly lengthened.

Boundary-related lengthening is also observed in studies of speech corpora. Kim et al. (2006) found that boundary vowel duration in a syllable preceding an intonational phrase boundary (L-L%) is significantly longer than in a syllable preceding an intermediate phrase boundary (L-) in both the Switchboard corpus of spontaneous

conversational speech of American English and the Boston University Radio News Corpus. Using the Boston Corpus, Yoon et al. (2007) further demonstrated that there is cumulative preboundary lengthening of the vowel as a function of the level of prosodic boundary. That is to say, the duration of word-final vowels is positively correlated with the level of the following prosodic phrase boundary. Such variation in duration before a prosodic phrase boundary was further shown to be utilized for automatic boundary detection and classification (e.g., Wagner, 2010).

There are several points of analysis that must be considered when examining durational measures as acoustic correlates of and as acoustic cues to prosodic phrase boundary. First of all, the domain of measurement should be carefully chosen. Many prior studies showed that phrase-related lengthening is not restricted to a certain segment within a word, but spreads to a larger domain, e.g., a syllable or a word immediately before a boundary, or even a syllable or a word that is not adjacent to a boundary (e.g., Beckman and Edwards, 1990; Berkovits, 1993a,b, 1994; Byrd, 2000; Byrd et al., 2006; Cambier-Langeveld et al., 1997; Crystal and House, 1988a,b; Nakai et al., 2009; Turk and Shattuck-Hufnagel, 2000, 2007; Wightman et al., 1992). For example, in their early study, Wightman et al. (1992) found that boundary-related lengthening does not spread evenly over a syllable or a word, but is instead mainly concentrated in the rhyme of the final syllable before a prosodic phrase boundary. Yet, results regarding the domain of final lengthening, and, in particular, to what extent boundary-related lengthening stretches, are somewhat conflicting. In Hebrew, Berkovits (1993a,b, 1994) also found effects of distance from the edge of a prosodic phrase on the duration of a word-final syllable, showing that the lengthening effect associated with the finality of utterance is progressive. That is, the coda consonant of the word-final syllable is lengthened to a greater extent than the preceding vowel. However, she also found that the location of lexical stress functions as an anchor to determine where boundary-related lengthening begins. If the final syllable of a

disyllabic word carries a lexical stress, then boundary-related final lengthening is confined to the final syllable, and the penultimate syllable does not show any significant lengthening effect. Yet, if the lexical stress is located in the first syllable of a disyllabic word, then lengthening begins from the lexically stressed syllable. Related findings from a production and perception study of Dutch are reported by Cambier-Langeveld et al. (1997) who show that boundary-related final lengthening is confined to the final syllable, and specifically, to the rhyme. However, they also found that although boundary-related final lengthening begins at the rhyme of the word-final syllable in most cases, final lengthening begins in the penultimate syllable when the final syllable contains only a schwa.

In a series of articulatory studies with the simple disyllabic word *dodo*, Byrd and her colleagues (Byrd, 2000; Byrd and Saltzman, 1998, 2003; Byrd et al., 2006) investigated the domain of boundary-related lengthening. Findings from these studies indicate that preboundary lengthening resulting from the longer opening movement of a preboundary consonant in the CV syllable is localized near a boundary, and the magnitude of the lengthening effect decreases as a function of distance from the edge of an utterance boundary. More recently, Turk and Shattuck-Hufnagel (2007) also found boundary-related lengthening in English, but with two independent targets of final lengthening domains: final syllables and lexically stressed syllables. However, their results are not entirely consistent with progressive lengthening because in the majority of cases the coda consonant of the final syllable did not show significant lengthening effects to a greater extent than the nucleus of that syllable.

Another complication arises from the fact that boundary-related lengthening is also observed in the post-boundary condition (Byrd and Rigg, 2008; Byrd and Saltzman, 2003; Byrd et al., 2006; Cho and Keating, 2001; Fougeron, 2001; Keating et al., 2003, among others). For example, Cho and Keating (2001) and Fougeron (2001) found that the length of the initial consonant of the first syllable following a prosodic

boundary increases as boundary strength increases. Byrd et al. (2006) demonstrates post-boundary lengthening in the closing durations of the onset consonant of the first syllable of the word **dodo** from all subjects as well as in the opening durations from two subjects, and further found a consistent shortening effect for the second and third consonant (*dodo knocking*) compared to the first consonant. Boundary-related lengthening after a prosodic boundary and its inverse relationship with the distance from a prosodic boundary edge was also observed in Byrd and Rigg (2008) and Kri-vokapic (2007). When reviewing the findings from the studies discussed above, it is shown that not only segments immediately before and after a prosodic boundary but also segments somewhat remote from a prosodic boundary are lengthened, and that boundary-related lengthening effects diminish as the distance from a prosodic boundary increases. It is important to select an appropriate domain for the durational measurement in relation to a prosodic boundary.

Lastly, final lengthening is often accompanied by and interacts with a following silent pause (Ferreira, 1993; Horne et al., 1995; Kainada, 2007). Ferreira (1993) claims when the duration of a word is relatively short, then the following pause is long and the sum of word and pause duration remains approximately equal regardless of prosodic contexts. Horne et al. (1995) also proposes that there is a trading relationship between segment duration and following silent interval duration, showing that segment duration is negatively correlated with silent interval duration at lower ranked boundaries. More recently, Kainada (2007) also found a compensatory relation between the degree of final lengthening and the duration of a pause following the lengthened word. Therefore, it is important to investigate whether the elongated duration of a segment or word under investigation and the silent pause following it contain redundant information that in signaling a prosodic boundary, or whether the contribution of final lengthening to the perception of a prosodic boundary is independent of that of a silent pause.

Although few studies make use of acoustic measures other than F_0 , segmental duration, and silent pauses as acoustic correlates of prosodic boundaries, other studies have shown that voice source and intensity measures are also influenced by a prosodic phrase boundary (Klatt and Klatt, 1990; Redi and Shattuck-Hufnagel, 2001 for voice quality measures, Chavarria et al., 2004; Choi et al., 2005; Kim et al., 2006 for voice quality and intensity). For example, Redi and Shattuck-Hufnagel (2001), employing two corpora of American English, first restricted the lexical items under investigation to those locations where all speakers produced a full intonational phrase boundary, and then marked four different types of glottalized events based on visual inspection of speech wave form, spectrogram, and F_0 contours, and auditory inspection: aperiodicity, creak, diplophonia, and glottal squeak. They found that normal speakers exhibit a glottalized voice quality in the vicinity of an intonational phrase boundary in American English. However, there is large variation in the types and the extent of glottalization that each speaker exhibits. In an automatic boundary detection study by Choi and her colleagues (2005), various voice sources measures including harmonic structure (end value, slope, and convexity of h1-h2) and spectral tile measures (end value, slope, and convexity of h1-a1, h1-a3, and a1-a3) at boundaries were found to be good indicators for boundaries. As for intensity measures, no study has found any significant influence of the presence/ absence of a prosodic phrase boundary on intensity measures, to my knowledge.

5.2 Acoustic analyses

As discussed in Chapter 2, boundary transcriptions obtained through Rapid Prosody Transcription in real time were pooled over all listeners. Each word in the transcribed speech materials was assigned a B-score representing the proportion of transcribers who perceived that word as followed by a boundary. Mean B-scores varied by phone-

mic category as shown in Figure 3.1. The following acoustic analyses were performed on each phone (as displayed in Table 3.1).

5.2.1 Measurements

The following acoustic measures were extracted from the word-final lexically stressed vowels: vowel duration (ms), overall rms intensity (dB), bandpass filtered intensities (i.e. spectral balance) in four different frequency bands (dB), four measures of F_0 (the local maximum of F_0 , F_0 at the right edge (Hz), and the onset and the offset F_0 slopes), formant frequencies (F_1 and F_2 ; Hz), and silent pause. Only the word-final lexically stressed vowels were used for these acoustic measures because prior research indicated that the lexically stressed vowel is one of the targets of boundary-related lengthening (Turk and Shattuck-Hufnagel, 2007; Berkovits, 1993a,b, 1994). This also allows us to hold any potential influence of lexical stress constant and to make parallel comparisons with prosodic prominence effects. All acoustic measures were extracted and normalized in the same way as for the prominence analyses in section 4.2.1. In addition, silent pauses longer than 20 ms were extracted, discarding silent pauses shorter than such duration.

5.3 Results

This section shall summarize findings from Spearman’s non-parametric correlation and multiple linear regression analyses of B-scores, indexing the relative strength of a perceived prosodic boundary, with the acoustic measures.

5.3.1 How closely is each acoustic measure related with the perception of prosodic boundary?

Results from Spearman’s non-parametric correlation analyses of perceived boundary with the acoustic measures from the word-final stressed vowels are summarized in Tables 5.1 and 5.2. Looking at the correlation analyses between acoustic measures and B-scores pooled over all vowels, the results show that many of the acoustic measures are significantly correlated with B-scores. For the same reason as in the analysis of perceived prominence as discussed in section 4.2.1, the F_2 measure was excluded from the analysis of perceived boundary. The presence of a silent pause following a word is the strongest correlate of B-scores ($\rho = 0.458$, $p < 0.001$), followed by vowel duration ($\rho = 0.378$, $p < 0.001$), local peak of F_0 ($\rho = 0.169$, $p < 0.001$), right edge F_0 ($\rho = -0.115$, $p < 0.001$), first formant ($\rho = 0.106$, $p < 0.001$), subband intensity in 0–500 Hz ($\rho = -0.069$, $p < 0.001$), overall rms intensity ($\rho = -0.066$, $p = 0.001$), and offset slope of F_0 ($\rho = 0.039$, $p = 0.026$), in order. On the other hand, the subband intensities in mid and high frequency regions and the onset slope of F_0 are not significantly correlated with B-scores.

Some acoustic measures including the presence of a silent pause following a word, vowel duration, the local maximum of F_0 , the first formant, and the offset slope of F_0 are positively correlated with B-scores while other acoustic measures, including intensity measures (overall rms intensity and subband intensity in 0–500 Hz) and F_0 measured at the right edge of the vowel, are negatively correlated with B-scores. In terms of the size of correlation coefficients, B-scores are more strongly correlated with two temporal measures—namely vowel duration and the silent pause duration—than with any other acoustic measure. The following section presents the results of correlation analyses between B-scores and each acoustic measure by vowel in detail.

Vowels		All	ɑ	æ	ʌ	ɔ	au 10%	au 90%	ai 10%	ai 90%	ε
Duration	Coeff.	0.378	0.067	0.505	0.484	0.355	0.420		0.313		0.437
	P	< 0.001	0.228	< 0.001	< 0.001	< 0.001	0.002		< 0.001		< 0.001
F_1	Coeff.	0.105	-0.345	0.347	0.213	0.063	0.023	-0.295	-0.034	-0.177	0.160
	p	< 0.001	0.001	< 0.001	< 0.001	0.272	0.440	0.026	0.289	0.002	0.002
F_2	Coeff.	N/A	-0.108	0.008	-0.014	0.064	0.294	-0.017	0.062	0.275	0.020
	p	N/A	0.061	0.457	0.409	0.271	0.024	0.456	0.155	< 0.001	0.361
Overall RMS intensity	Coeff.	-0.070	-0.307	-0.196	-0.136	-0.067	0.260		-0.120		0.039
	p	0.001	< 0.001	0.003	0.011	0.261	0.041		0.024		0.249
SB (0–500 Hz)	Coeff.	-0.072	-0.255	-0.194	-0.168	-0.047	0.275		-0.120		0.055
	p	< 0.001	0.002	0.003	0.002	0.326	0.032		0.023		0.166
SB (500–1000 Hz)	Coeff.	0.003	-0.316	-0.175	0.002	-0.072	0.176		-0.119		0.038
	p	0.446	< 0.001	0.006	0.489	0.244	0.120		0.024		0.253
SB (1000–2000 Hz)	Coeff.	-0.022	-0.345	-0.146	-0.022	0.002	0.189		-0.051		0.086
	p	0.133	< 0.001	0.019	0.357	0.492	0.105		0.201		0.066
SB (2000–4000 Hz)	Coeff.	-0.036	-0.349	0.149	-0.047	-0.007	0.223		0.216		0.140
	p	0.035	< 0.001	0.017	0.214	0.474	0.068		< 0.001		0.007
$F_{0,max}$	Coeff.	0.175	0.134	0.141	0.116	0.139	0.251		0.193		0.166
	p	< 0.001	0.067	0.023	0.025	0.091	0.046		0.001		0.002
Right F_0	Coeff.	-0.120	-0.085	-0.091	-0.156	0.093	-0.228		-0.147		-0.130
	p	< 0.001	0.067	0.098	0.004	0.186	0.063		0.007		0.011
Onset slope	Coeff.	0.014	-0.127	0.117	0.068	-0.042	0.110		-0.035		-0.026
	p	0.241	0.078	0.048	0.126	0.343	0.233		0.281		0.323
Offset slope	Coeff.	0.036	-0.176	0.121	0.130	0.227	-0.143		0.012		0.019
	p	0.035	0.024	0.043	0.014	0.014	0.172		0.420		0.370
Pause	Coeff.	0.458	0.457	0.496	0.535	0.467	0.296		0.493		0.437
	p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.023		< 0.001		< 0.001

Table 5.1: Summary of the Spearman’s non-parametric correlation analyses between B-scores and acoustic measures of each vowel (I). Spearman’s ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.

Vowels		ɜ	ei 10%	ei 90%	ɪ	i	ou 10%	ou 90%	u	u
Duration	Coeff.	0.499	0.453		0.365	0.314	0.441		0.200	0.275
	<i>p</i>	< 0.001	< 0.001		< 0.001	< 0.001	< 0.001		0.063	< 0.001
F_1	Coeff.	0.111	0.247	-0.029	0.117	0.106	0.115	0.046	0.217	0.255
	<i>p</i>	0.151	0.001	0.366	0.013	0.051	0.091	0.297	0.048	0.001
F_2	Coeff.	-0.109	-0.035	0.095	0.016	0.189	-0.075	-0.089	0.019	-0.266
	<i>p</i>	0.156	0.338	0.129	0.381	0.002	0.195	0.153	0.444	< 0.001
Overall RMS intensity	Coeff.	0.050	-0.194		-0.002	0.007	-0.217		0.231	-0.167
	<i>p</i>	0.323	0.009		0.486	0.455	0.006		0.038	0.020
SB (0–500 Hz)	Coeff.	0.061	-0.203		-0.015	0.008	-0.185		0.214	-0.182
	<i>p</i>	0.285	0.007		0.391	0.450	0.016		0.050	0.013
SB (500–1000 Hz)	Coeff.	0.108	-0.039		0.112	0.005	-0.172		0.276	0.049
	<i>p</i>	0.158	0.319		0.016	0.472	0.023		0.017	0.276
SB (1000–2000 Hz)	Coeff.	0.089	-0.043		0.017	0.071	-0.108		0.030	-0.107
	<i>p</i>	0.205	0.302		0.373	0.138	0.105		0.409	0.096
SB (2000–4000 Hz)	Coeff.	-0.086	-0.131		0.039	0.083	-0.135		0.095	-0.259
	<i>p</i>	0.212	0.057		0.229	0.100	0.059		0.236	0.001
$F_{0,max}$	Coeff.	0.226	0.232		0.117	0.109	0.252		0.135	0.228
	<i>p</i>	0.017	0.002		0.013	0.047	0.002		0.152	0.002
Right F_0	Coeff.	-0.012	-0.081		-0.140	-0.094	-0.248		-0.127	-0.138
	<i>p</i>	0.455	0.162		0.004	0.073	0.002		0.166	0.045
Onset slope	Coeff.	0.085	0.007		-0.004	-0.026	0.039		0.092	-0.018
	<i>p</i>	0.215	0.468		0.473	0.347	0.327		0.243	0.412
Offset slope	Coeff.	-0.045	-0.051		-0.010	0.121	-0.091		0.084	0.112
	<i>p</i>	0.337	0.269		0.427	0.030	0.136		0.262	0.085
Pause	Coeff.	0.339	0.529		0.497	0.326	0.497		0.434	0.402
	<i>p</i>	0.001	< 0.001		< 0.001	< 0.001	< 0.001		< 0.001	< 0.001

Table 5.2: Summary of the Spearman’s non-parametric correlation analyses between B-scores and acoustic measures of each vowel (II). Spearman’s ρ coefficients for significant correlations are shown in shaded cells, with corresponding p -values. For diphthongs, formant measures are sampled at a location 10% into the duration of the vowel and again at the 90% location.

5.3.1.1 Effects of prosodic boundary on the duration of word-final lexically stressed vowels

Results from Spearman's non-parametric correlation analyses indicate that B-scores representing the locations of a prosodic boundary and its relative strength are strongly correlated with the duration of word-final lexically stressed vowels as seen in Tables 5.1 and 5.2. All the word-final stressed vowels (except /a, u/) show a significant positive correlation between vowel duration and B-scores, for which correlation coefficients (ρ) range from 0.275 (/u/, $p < 0.001$) to 0.505 (/æ/, $p < 0.001$). These results show that word-final lexically stressed vowels are longer when perceived as followed by a prosodic boundary.

Findings from correlation analyses of B-scores and acoustic measures are similar to those from the correlation analyses of P-scores and acoustic measures in that most of the vowels under investigation are significantly lengthened as prosody scores increase. Compared with the correlation coefficients in the P-score analyses ($\rho = 0.156 \sim 0.380$), the correlation coefficients of B-scores and vowel duration ($\rho = 0.275 \sim 0.505$) are in general large. Looking closely, the correlation coefficient of B-scores for each vowel is consistently higher than that of the P-scores for the corresponding vowel, and there is only one correlation coefficient ($\rho = 0.275$ for /u/), which is smaller than 0.300, while more than half of the correlation coefficients for P-scores and vowel duration are smaller than 0.300.

5.3.1.2 Effects of the presence of a silent pause on the perception of a prosodic boundary

The results from Spearman's non-parametric correlation analyses indicated that B-scores are significantly correlated with the presence of a word-final silent pause longer than 20 ms as shown in Tables 5.1 and 5.2. All the words containing the lexically stressed vowel in Tables 5.1 and 5.2 are followed by the presence of a silent pause, and

their correlation coefficients range from $\rho = 0.296$ ($p = 0.023$) for /av/ to $\rho = 0.547$ ($p < 0.001$) for /ov/.

5.3.1.3 Effects of prosodic boundary on overall and subband intensity measures of the word-final lexically stressed vowel

Spearman's non-parametric correlation analyses were also performed for B-scores and the intensity measures including overall and subband intensities in four frequency regions (0–500, 500–1000, 1000–2000, and 2000–4000 Hz). The results show that among four different intensity measures, both overall intensity and subband intensity in 0–500 Hz are the most reliably and strongly correlated with B-scores across vowels: 8 (/ɑ, æ, ʌ, aɪ, eɪ, u, av, v/) and 9 (/ɑ, æ, ʌ, aɪ, eɪ, ov, u, av, v/) out of 14 word-final lexically stressed vowels, respectively. The corresponding correlation coefficients range from $\rho = -0.397$ ($p < 0.001$) for /ɑ/ to $\rho = 0.260$ ($p = 0.041$) for /av/ and from $\rho = -0.255$ ($p = 0.002$) for /ɑ/ to $\rho = 0.275$ ($p = 0.032$) for /av/ in order. Other intensity measures including subband intensities in 500–1000, 1000–2000, and 2000–4000 Hz are not consistently significantly correlated with B-scores for the following: 5 (/ɑ, æ, aɪ, ɪ, v/), 2 (/ɑ, æ/), and 5 (/ɑ, u, æ, aɪ, ε/) out of 14 word-final lexically stressed vowels.

The results from Spearman's non-parametric correlation analyses further indicate several differences in the cues for the perception of prosodic prominence vs. cues for boundary perception. In terms of kinds of intensity measures that are significantly correlated with P- and B-scores, in addition to overall intensity, B-scores are mainly correlated with subband intensity in the low frequency band (0–500 Hz), while P-scores are correlated with subband intensities in the mid and high frequency bands (500–2000 Hz). Contrary to the findings that overall and subband intensities in the mid and high frequency bands increase as P-scores increase, overall and subband intensities in the low frequency band decrease as B-scores increase in most word-final,

lexically stressed vowels. It was further shown that when subband intensity in 0–500 Hz is significantly correlated with B-scores, overall intensity always demonstrates a significant correlation with B-scores.

5.3.1.4 Effects of prosodic boundary on F_0 parameters of word-final lexically stressed vowel

Spearman’s non-parametric correlation analyses also show that perceived boundary is significantly correlated with changes in F_0 patterns for many of the word-final lexically stressed vowels. Among the F_0 measures included in the current study, both the local F_0 maximum and the F_0 at the right edge of the vowel are the most reliably correlated with B-scores across all vowels. The local peak of F_0 is significantly correlated with B-scores in 11 out of 14 word-final, lexically stressed vowels, whose correlation coefficients range from $\rho = 0.116$ ($p = 0.025$) for /ʌ/ to $\rho = 0.251$ ($p = 0.046$) for /av/. The F_0 measured at the right edge of the word-final lexically stressed vowels is significantly correlated with 6 out of 14 vowels, whose correlation coefficients range from $\rho = -0.264$ ($p = 0.011$) for /ov/ to $\rho = -0.130$ ($p = 0.014$) for /ε/. Yet, the onset and the offset slope of F_0 is significantly correlated with B-scores in only very few vowels (1 vowel with the onset F_0 slope, /æ/ and 3 vowels with the offset F_0 slope, /ɑ, æ, i/). More interestingly, the results showed that B-scores are positively correlated with the local F_0 maximum, while they are negatively correlated with the F_0 at the right edge of the word-final, lexically stressed vowel, suggesting that when a word is perceived as followed by a prosodic boundary, the local F_0 peak is higher, but the F_0 measured at the right edge of the vowel is lower, which may result from the locations of nuclear prominence within a prosodic phrase. The interaction between nuclear prominence and prosodic boundary in the realization of F_0 measures will be discussed in section 5.4 in detail.

Comparing these results from correlation analyses of B-scores and F_0 measures

with those from correlation analyses of P-scores and F_0 measures, there are several interesting findings to be noted. First of all, both P- and B-scores are significantly, positively correlated with the local F_0 maximum. In other words, as a word is perceived as prominent or as followed by a prosodic boundary, the local F_0 peak within the lexically stressed vowel increases. There are also some distinct differences in the signaling of a prosodic boundary from the signaling of prosodic prominence. When ordinary listeners perceive a word as prominent, the maximum F_0 (5 out of 14) and the slope of F_0 from the onset to the peak (7 out of 14) within a vowel increases in many lexically stressed vowels, but neither F_0 at the right edge nor the offset slope of F_0 is significantly correlated with P-scores. To the contrary, F_0 at the right edge of the vowel is a good cue for a prosodic boundary, but the onset slope of F_0 is not. In sum, prosodic prominence is cued by the rising slope of F_0 at the beginning of the vowel, while prosodic boundary is signaled through the lower F_0 at the right edge of the vowel.

5.3.1.5 Effects of prosodic boundary on formant structure of word-final lexically stressed vowels

The results from Spearman's non-parametric correlation analyses show that the formant values of stressed word-final vowels are also influenced by the presence of a prosodic boundary. As for F_1 , the first formant measures of 10 out of 14 word-final, lexically stressed vowels are significantly correlated with B-scores. More specifically, the vowel midpoint F_1 measures in 6 monophthongs (/æ, ʌ, ε, ɪ, ʊ, u/) are positively correlated with B-scores, while the correlation coefficient with B-scores and F_1 is negative for the vowel /ɑ/. Diphthongs also show sporadic significant correlations with B-scores and F_1 values: negative correlations with F_1 in the offglide of two diphthongs (/av, aɪ/), and a positive correlation with F_1 in the nucleus of one diphthong (/eɪ/). Turning next to F_2 , the F_2 measures of 5 out of 14 word-final, lexically

stressed vowels are significantly correlated with B-scores: the high front vowel /i/ has higher F_2 s, and the high back vowel /u/ has lower F_2 s when perceived as followed by a prosodic boundary. Three diphthongs (the nucleus of one diphthong /av/ and the offglide of two diphthongs /aɪ, oʊ/) also show significant correlations with B-scores and F_2 values.

5.3.1.6 Summary of the findings from the Spearman's non-parametric correlation analyses

Findings from Spearman's non-parametric correlation analyses illustrate that the perception of prosodic boundary by ordinary listeners is closely associated with systematic changes in the acoustic signal, in particular, temporal measures including vowel duration and silent pauses. In other words, the word-final, stressed vowels when heard as preceding a prosodic boundary tend to have longer vowel durations and are often followed by a silent pause longer than 20ms. In addition, when perceived as followed by a boundary, word-final, lexically stressed vowels have higher local F_0 peaks but lower F_0 measures at the right edge. In terms of intensity measures, word-final stressed vowels tend to have a reduced overall intensity and reduced subband intensity in 0–500 Hz. However, subband intensities in the other bands and formant measures are not affected by the presence of a prosodic boundary. Although formant values do not show systematic correlations with B-scores, F_1 measures in many vowels tend to be higher while F_1 measures in the open vowel is lower as B-scores increase and F_2 values in few vowels are significantly correlated with perceived boundary.

5.3.2 To what extent do different acoustic measures contribute to listeners' perception of prosodic boundary?

This section presents an analysis of the combined contribution of acoustic cues to boundary perception, and analyses of the contribution of individual acoustic measures to the perception of prosodic boundary. These results are based on the results from multiple stepwise linear regressions of perceived prosodic boundary.

The possible redundancy among the acoustic cues for prosodic boundary was evaluated by comparing the total sum of the coefficient of determination (r^2) from a series of linear regression analyses between B-scores and a single acoustic measure with the coefficient of determination from a simple multiple linear regression analysis between B-scores and the acoustic measures altogether. The results show that the total sum of this variation is always greater than the total variation obtained from simple linear regressions, suggesting that acoustic measures as predictors are interrelated, and that some information in the acoustic signal is redundant. Similar to the findings from perceived prominence, these findings further suggest that patterns of combined acoustic measures should be taken into account in order to model prosodic boundary perception.

Figure 4.1 illustrates the total variation of boundary perception that is accounted for on the basis of changes in acoustic cues. Pooling together the boundary markings from all the words, over 38% of the variation in boundary perception can be explained by acoustic information in the speech signal. Looking at the multiple linear regression models by individual vowel, the acoustic variation in the combination of temporal measures (vowel duration and silent pause), intensity measures (overall and subband intensities in four frequency bands), fundamental frequency measures (local peak of F_0 , F_0 at the right edge, and the onset and the offset slope of F_0),

and formant measures (F_1 and F_2) of the word-final, lexically stressed vowel can account for between 32% (for /ɜ:/) to 63% (for /ou/) of the variation found in listeners' boundary perception. In comparisons with the multiple linear regression models of prosodic prominence as perceived by listeners, the multiple linear regression models of boundary perception account for a greater proportion of the variability; all the regression models of perceived prosodic boundary predict more than 30% of listeners' response to prosodic boundaries, while only one regression model of perceived prosodic prominence can account for such a proportion.

Subsequent stepwise multiple linear regression analyses were performed to investigate which acoustic parameters contribute most in the statistical model of boundary perception, and how much the individual acoustic parameters contribute to predicting boundary perception. Contrary to regression models of perceived prosodic prominence, the results summarized in Figure 5.1 reveal that the best statistical model of boundary perception relies heavily on temporal measures and the contribution of other acoustic measures is not significant in most of the regression models. In other words, perceived boundary is primarily modeled by temporal measures including vowel duration and the presence of a subsequent silent pause, while modeling prominence perception requires a large number of acoustic parameters. Acoustic measures other than vowel duration and silent pause are included in the statistical model of boundary perception for very few word-final lexically stressed vowels (intensity measures for four vowels (/ɑ, æ, u, eɪ/), F_0 measures for five vowels (/ɑ, ɪ, aɪ, eɪ, ou/), and formant measures for two vowels (/i, aɪ/), but these acoustic measures also contribute at best very little to boundary perception.

In summary, the findings from simple multiple linear regression analyses reveal that a great proportion (32–63%) of the variation in boundary perception can be explained on the basis of acoustic information, contrary to the situation with perceived prominence (15–40%). Yet, in terms of the number of acoustic measures as predic-

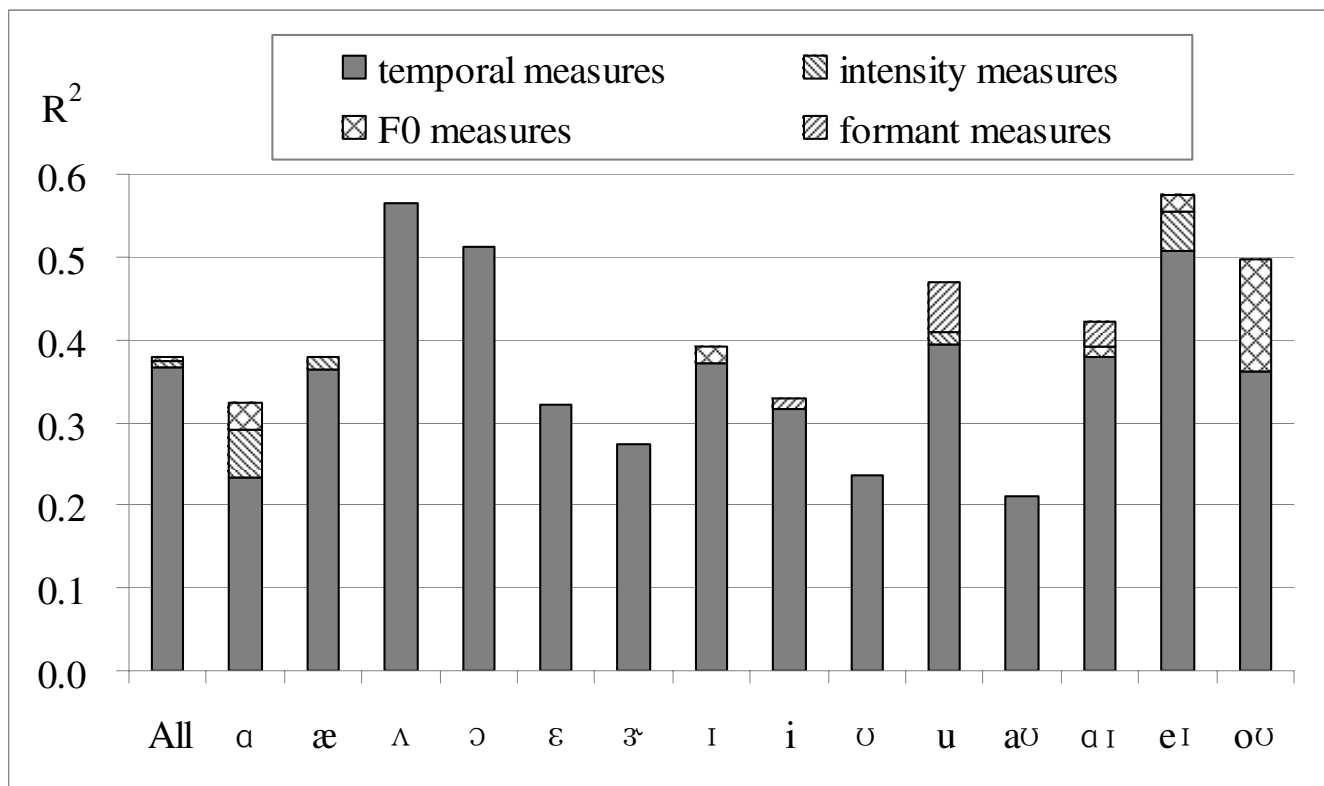


Figure 5.1: The distribution of the variation (r^2) in the ordinary listeners' response to prosodic boundary predicted by stepwise multiple linear regressions of the acoustic measures and B-scores

tors in statistical models of prosody perception, the model of prominence perception requires information from a combination of acoustic parameters including vowel duration, intensity, F_0 , and formant measures, while the model of boundary perception is primarily dependent on temporal measures: the elongation of vowel duration and the presence of a silent pause following a word.

5.4 Summary and Discussion

The findings from the current study show that listeners' perception of prosodic boundary is correlated with the acoustic patterns of word-final, lexically stressed vowels. Among all the acoustic measures, speakers use primarily the temporal properties of speech to signal prosodic boundaries, e.g., elongation and pausing. Other acoustic parameters are also shown to be varied in pre-boundary position. When followed by a prosodic boundary, word-final, lexically stressed vowels are longer than their phrase-medial counterparts, and are often followed by a silent pause longer than 20 ms. In addition, when a word is perceived as preceding a prosodic boundary, the overall intensity and the subband intensity in 0–500 Hz tend to be lower, and F_0 at the right edge of the vowel also tends to decrease before a prosodic boundary, while the local F_0 maximum increases.

These results point to some similarities in the phonetic implementation of prosodic boundary and prosodic prominence: lexically stressed vowels tend to have longer vowel, higher local F_0 maximum, and higher F_1 (suggesting a larger mouth opening and/or lower tongue position). At the same time, the results also reveal that prosodic boundary is signaled through lower intensities (overall and in the low frequency band, 0–500 Hz) and lower F_0 at the right edge of the vowel, and is often followed by a silent pause longer than 20 ms. To the contrary, prosodic prominence is realized through higher intensities (overall and in the mid and high frequency bands,

500–2000 Hz), higher onset slope of F_0 , and formant values that reflect more peripheral tongue position in the front/back dimension. In sum, taking into account all the findings from the acoustic investigation of prosodic prominence and boundary, it is demonstrated that, under prosodic prominence lexically stressed vowels enhance their phonetic characteristics (including segmental and suprasegmental characteristics) in all acoustic dimensions. However, before a prosodic boundary, these word-final, lexically stressed vowels are phonetically reduced in most of acoustic dimensions except temporal characteristics.

One can attempt to interpret these findings about acoustic variation associated with prosodic prominence and boundary from an articulatory perspective. The findings from the acoustic investigation of prosodic prominence are consistent with predictions of the task-dynamic model of speech production, based on the theory of articulatory phonology augmented with a model of prosody in terms of articulatory gestures (Byrd and Saltzman, 1998, 2003; Saltzman et al., 2008). In the task-dynamic production model of prosodic prominence, prosodic prominence is associated with temporal expansion, resulting from the activation of temporal modulation gestures (μ_T) and with spatial expansion, resulting from the activation of the spatial modulation of constriction gestures of the articulators (μ_S) that increase the spatial target parameters of constriction gestures. These enhanced spatio-temporal arrangements of the articulators by -gestures under prosodic prominence are reflected in the acoustic domain as the acoustic enhancement of segmental and suprasegmental characteristics, which are shown in the findings from the present study, in that the lexically stressed vowels when perceived as prominent have a longer duration, higher intensity and F_0 , and a dispersed and amplified formant structure.

This study also demonstrates robust acoustic evidence for the reduction or shrinking of speech sounds in all acoustic domains other than the temporal domain, which is not predicted by the task-dynamic model of prosody production. In the task-dynamic

model of prosodic boundary production, temporal modulation gestures (π) are activated, and subsequently slow down articulatory gestures in the vicinity of prosodic boundaries, which can account for the temporal stretching of speech elements near a prosodic boundary. Yet, the articulatory model does not propose any particular spatial effects on gestures due to proximity to a prosodic boundary. As a consequence, under the task-dynamic model of boundary production, there is no prediction for the acoustic effects reflecting the particular spatial coordination of the articulators before a prosodic boundary. The results from the current study also provide evidence that word-final, lexically stressed vowels have a reduced subband intensity in the low frequency region in addition to an overall reduced intensity and lowered F_0 at the right edge of the vowel, confirming that speech elements before prosodic boundaries are acoustically reduced and less distinct.

The findings from the current study are consistent with those from some recent articulatory studies, suggesting that there must be different underlying articulatory mechanisms associated with lengthening. For example, examining μ -beam x-ray data, Goldstein and his colleagues (Cole et al., 2008) report that before a prosodic boundary, the duration of gestural release is longer with lower stiffness, while the duration of gestural formation is longer under prosodic prominence. The findings from the current acoustic and other articulatory studies suggest that a prosody gesture other than the π -gesture should be introduced to account for acoustic and articulatory reduction before a prosodic boundary. This added gesture may reflect subglottal and supraglottal aerodynamic factors in addition to the spatiotemporal coordination of the articulators. This proposal for an additional gesture to model prosody production builds on recent work to incorporate aerodynamic and laryngeal features into articulatory research such as McGowan and Saltzman (1995).

Although there are a handful of studies in which the relationship between changes in aerodynamic factors and speech variation that stems from prosodic context are ex-

amined, it has been observed that the downtrend of F_0 movement in the course of an utterance is closely correlated with a downtrend in subglottal pressure (Collier, 1975; Lieberman, 1967; Ohala, 1990; Strik and Boves, 1994, 1995; Titze, 1989; Fant et al., 1996, 1997 among others). For example, Lieberman (1967) pointed out that a decrease of the open quotient of glottal waveforms may directly result from the increase of subglottal air pressure. In addition, Collier (1975) also looked at the physiological implementation of prosodic phenomena, including intonation and prominence, and found that a decrease of subglottal pressure is roughly correlated with an utterance final drop of F_0 as well as a gradual decrease of F_0 over the course of an utterance and the increase of subglottal pressure with the rising F_0 contour due to accentuation. Among the most recent studies, the majority of which were carried out by Fant and his colleagues at the Royal Institute of Technology in Sweden to examine the relations between subglottal pressure and the F_0 realization of accent (Fant et al., 1996, 1997; Liljencrants et al., 2000), few studies have investigated the relationship between subglottal pressure fall and falling F_0 at the end of a prosodic phrase (Atkinson, 1978; Herman et al., 1996). For example, Herman and her colleagues (1996) examined how final lowering of F_0 is attributed to a declination in subglottal pressure and demonstrated that sentences with a low boundary tone (L in the paper following the ToBI standards) are strongly bound to steeper declination of subglottal pressure. The amplitude of speech elements has also been shown to be affected by aerodynamic factors (Hanson, 1997; Tanaka and Gould, 1983). Tanaka and Gould (1983) found that subglottal pressure is correlated with vowel intensity by simultaneously measuring vocal intensity and subglottal pressure during the production of sustained vowel /a/. In a more recent study by Hanson (1997), in which supraglottal pressure along with the acoustic signal was recorded during the production of a reiterant utterance, it was also shown that speakers manipulate subglottal pressure to control vowel amplitude. These findings from prior aerodynamic studies are consistent with those from the

current study by which it was shown that word-final, lexically stressed vowels have lower intensity (overall and subband in 0–500 Hz) and F_0 at the right edge of the vowel, suggesting that the spatiotemporal configuration of the vocal tract above the glottis, air pressure, and flow in the vocal tract, as well as trans- and sub-glottal air flow and pressure should be taken into account simultaneously in models of prosody production.

The final finding that requires further discussion is that the local F_0 maximum of many of the word-final, lexically stressed vowels is significantly positively correlated with B-scores, which is contradictory to the predictions from the articulatory and aerodynamic accounts, by which it is claimed that speech elements before a prosodic boundary should be reduced in the segmental and suprasegmental acoustic dimension, except in the temporal dimension. However, looking closely, many words just before a prosodic boundary carry prosodic prominence. In other words, a phrase final word is a nuclear prominence bearing word in many cases. As described by prior research (Calhoun, 2006; Ladd, 2008; Pierrehumbert and Beckman, 1988), the right-most prominence within a phrase is the strongest and the most phonetically distinct prominence, and is structurally determined. Therefore, it is possible that the final word within a phrase is perceived as prominent, which would account for its positive correlation with the raised local F_0 maximum. The possibility that the location of a prosodic boundary is correlated with the location of prosodic prominence and if so, that nuclear prominence is distinct from prenuclear prominence in terms the acoustic cues to prominence will be examined in future research.

After confirming that there are acoustic marks of prosodic boundary on a word-final, lexically stressed vowel, the statistical multiple linear regression models of perceived prosodic boundary are established as shown in Figure 4.1 and Figure 5.1. Compared with the multiple linear regression models of perceived prosodic prominence, the statistical models of perceived prosodic boundary account for a larger variability

of listeners' boundary perception. The regression model of perceived prominence for only one vowel achieves a prediction of over 30% of listeners' perception, while the regression models of perceived prosodic boundary account for over 40% of vowels. In light of the findings from Spearman's non-parametric correlation analyses of B-scores and the acoustic measures, the regression models of boundary perception for most vowels include only temporal measures (vowel duration and subsequent silent pauses) as predictors, with other acoustic measures included for only a small number of vowels. Their contribution to the perception of prosodic boundary, if any, is very limited, which may stem from indistinctiveness of reduced acoustic measures. In other words, the only salient cues for prosodic boundary available to listeners are the elongated duration of speech sounds preceding the boundary, approximated as vowel duration in the current study, and a perceptible silent pause following it.

5.5 Conclusion

Findings from Spearman's non-parametric correlation analyses and multiple linear regression analyses indicate that American English speakers signal the location of prosodic boundaries through temporal parameters in spontaneous conversational speech, and listeners attend to these acoustic cues use them to reliably locate prosodic boundary. Comparing the acoustic implementation of prosodic prominence with that of prosodic boundary, the present study also reveals that the underlying production mechanisms of prosodic boundary are different from those of prosodic prominence production. In the production of prosodic prominence in spontaneous conversational speech of American English, speakers employ various acoustic parameters to enhance the acoustic salience of speech elements under prominence. On the other hand, in the production of prosodic boundary, speakers primarily manipulate the temporal characteristics of speech elements in the vicinity of a prosodic boundary, with shrinkage of

all other acoustic characteristics. Listeners attend to the enhancement of many acoustic parameters in the perception of prosody in spontaneous conversational speech of American English.

Chapter 6

Prosodic Effects on the Temporal Structure of Monosyllabic CVC Words

6.1 Introduction

Among various acoustic correlates of prosody, phone duration is a primary cue signaling both prosodic prominence and boundary. This chapter looks further into the durational effects of prosody, to consider durational effects on subsyllabic structure of monosyllable CVC words. Prior studies demonstrate that both prosodic phrase boundary and prominence affect the temporal properties of words and syllables through boundary- and prominence-related lengthening (e.g., Cambier-Langeveld and Turk, 1999; Crystal and House, 1988a,b; Turk and Sawusch, 1996, 1997; Turk and White, 1999 for prosodic prominence and Beckman and Edwards, 1990; Byrd et al., 2006; Cambier-Langeveld, 1997; Cambier-Langeveld et al., 1997; Klatt, 1975; Wightman et al., 1992). Many studies particularly investigate the domain of lengthening due to prosodic prominence and boundary. For instance, in an early study, Crystal and House (1988a; 1988b) indicate that speech segments found under and before stress are longer than corresponding segments that are neither stressed nor before stress. Turk and Sawusch (1996; 1997) find that accentual lengthening effects within a word (i.e., lengthening due to phrasal stress) begin with an accented syllable and extend through at least one unstressed syllable following the accented syllable. Cambier-Langeveld and Turk (1999) show that syllables within an accented word are indeed lengthened, but that the degree of lengthening within a word varies as a function of position. Overall, they find that adjacent syllables to the right of the accented

syllable are lengthened more than adjacent syllables to the left of an accented syllable, in addition to the lengthening of the accented syllable. These findings suggest that as the distance from the accented syllable increases, the size of the lengthening effect decreases. Turk and White (1999) present similar findings, whereby accentual effects on duration spread onto syllables that are not adjacent to an accented syllable. In sum, findings from prior studies have demonstrated that when a word is accented, not only are accented syllables lengthened, but so are syllables that are not adjacent to the accented syllables, with a lessening of the effect as the distance from the accent increases.

Other studies have investigated the domain of boundary-induced lengthening effects, though the findings are somewhat contradictory. In an early study, Klatt (1975) finds that phrase-final syllables are lengthened. Later, employing a speech corpus of 35 pairs of phonetically-similar but syntactically ambiguous sentences read by professional news announcers of American English, Wightman and his colleagues (1992) report that segmental lengthening induced by prosodic boundary does not spread evenly over a syllable or a word, but is restricted to the rhyme of the pre-boundary syllable.

In addition to English, there are many languages where prior studies indicate prosodic boundary-related lengthening effects in the vicinity of prosodic boundary. Contrary to findings from Wightman et al. (1992), Cambier-Langeveld and her colleagues (Cambier-Langeveld, 1997; Cambier-Langeveld et al., 1997) show that in Dutch, regardless of stress position and of the depth of prosodic boundary, final lengthening is not confined only to a word's final segment or final rhyme, but that the amount of lengthening is greatest in the final segment and gradually decreases as the distance from that boundary increases. In Hebrew, effects of distance from the edge of a prosodic phrase on the duration of a word-final syllable are reported by Berkovits (1993a,b, 1994), showing that lengthening effects associated with the final-

ity of an utterance are progressive, with coda consonants in word-final position longer than the preceding vowel. However, her findings also indicate an interaction between final lengthening and the location of lexical stress. If the final syllable of a disyllabic word carries lexical stress, then boundary-related final lengthening is confined to the final syllable, and the penultimate syllable does not show a significant lengthening effect. If the lexical stress is, on the other hand, located in the first syllable of a disyllabic word, then lengthening begins from the lexically stressed syllable. This study demonstrates that the domain of final lengthening is a function of both the distance from an utterance boundary and the location of lexical stress.

More recently, in their series of articulatory studies, Byrd and her colleagues (Byrd, 2000; Byrd and Rigg, 2008; Byrd and Saltzman, 1998, 2003; Byrd et al., 2006) have investigated boundary-related lengthening and its domain. Their results indicate that speech segments both before and after a boundary are lengthened, and that this effect decreases as the distance from the edge of an utterance boundary increases. The recent study by Turk and Shattuck-Hufnagel (2007) confirms earlier findings for boundary lengthening and also shows evidence of boundary-related lengthening in potentially non-contiguous multiple domains that include the final syllable and the rhyme of the main-stress syllable. These results are similar to findings from Berkovits (1993a,b, 1994).

The studies discussed above examine durational effects of prosody in read speech elicited in a laboratory, where punctuation or other text marking devices are used to evoke the desired prosodic structures. Among the few studies to examine the temporal encoding of prosody in spontaneous speech, Aylett and Turk (2004) report prominence-induced lengthening of syllables in spontaneous speech from a Map Task corpus. Their analysis considers the number of phones in the syllable, but does not report prosodic lengthening effects on sub-constituents of the syllable, or on individual phones. Greenberg et al. (2003), on the other hand, examine lengthening

effects at the phone level due to prominence (‘stress accent’ in the terminology of their study), by analyzing spontaneous speech data from the Switchboard Corpus of Telephone Conversations in American English. Findings from this study indicate that the magnitude of accentual lengthening is largest in the nucleus of an accented syllable, with a smaller effect on onset consonants, and no significant effect on coda consonants.

The current study expands upon the approach of Greenberg et al. (2003), asking whether prosodic prominence and boundary exert similar effects on the temporal structure of subcomponents of monosyllabic CVC words in spontaneous speech. First, like Greenberg et al. (2003), the current study examines the effect of prominence on the temporal structure of monosyllabic CVC words and boundary effects, and looks at the interaction of prosodic prominence and boundary in their effects on temporal measures. Second, whereas Greenberg and colleagues identify ‘stress accent’ based on the transcriptions of two trained, expert transcribers, the analysis here is based on prosody annotation from ordinary listeners.

6.2 Acoustic analyses

The data analyzed here are the same as for the acoustic analyses presented in Chapters 4 and 5, using the RPT method described in Chapter 2.

6.2.1 Acoustic measurements

Using the word and phone transcriptions available from the Buckeye Corpus, 771 monosyllabic CVC words containing a lexical stress were isolated from a set of 54 speech excerpts. That is, all multisyllabic words, and all monosyllabic words that do not contain a lexically stressed vowel-e.g., function words and frequently reduced words-and whose syllable shape is not CVC were removed from the dataset. Prior

studies (Berkovits, 1993a,b, 1994; Crystal and House, 1988a,b; Klatt, 1975; Turk and Shattuck-Hufnagel, 2007) indicate the contextual effects on prominence- and boundary-related lengthening. Looking at only monosyllabic CVC words carrying a lexical stress, therefore, I can eliminate some sources of contextual effects although the variability in contexts cannot be completely removed. The total word duration as well as the durations of the onset, nucleus, and coda of each monosyllabic CVC word were measured from the time-aligned phone transcriptions. The relative proportions of the onset, nucleus, and coda within the monosyllabic word were calculated by dividing the duration of each subsyllabic component of a word by the total duration of the word containing that subsyllabic component as in equation 6.1. The duration of each subcomponent was also normalized by phone within speaker as in equation 4.2. In sum, four temporal measures of the monosyllabic CVC words were included in the statistical analyses: raw durations, durational ratio, and z -normalized durations of the onset, nucleus, and coda and raw durations.

$$\text{Ratio of O, N, C} = \frac{\text{Duration of subsyllabic component (O, N, or C)}}{\text{Total duration of the word (O + N + C)}} \quad (6.1)$$

6.3 Results

In this section, I summarize findings from multiple linear regression analyses of P- and B-scores with the various durational measures of speech segments—the total raw durations of the monosyllabic CVC words and the durations of each subsyllabic component (onset, nucleus, and coda). A model of the internal temporal structure of the monosyllabic CVC words in relation to prosodic structure is further proposed.

6.3.1 How does prosodic prominence influence the temporal characteristics of monosyllabic CVC words?

Before examining prosodic prominence effects on the internal temporal structure of the monosyllabic CVC words, the scatter plot in Figure 6.1 illustrates that word durations increase along with the increase of prosodic prominence scores (P-scores) as indicated by ordinary listeners. Looking closely, the raw durations of each subsyllabic component of the monosyllabic CVC words are also positively correlated with P-scores, as shown in Figure 6.2. These results demonstrate that, in addition to the total durations of the monosyllabic CVC words, duration measures of the onset, nucleus, and coda of the monosyllabic CVC words are all longer in words that are more likely to be perceived as prominent, and the same durations are all similar to one another in words that are more likely to be perceived as not prominent by ordinary listeners. Yet, as shown in Figure 6.2, although the durations of each subcomponent of the monosyllabic CVC words are significantly positively correlated with P-scores, the strength of the correlations between the durational measures and P-scores varies as a function of the position within a word: $\rho = 0.339$ ($p < 0.001$) for the onset duration, $\rho = 0.496$ ($p < 0.001$) for the nucleus duration, and $\rho = 0.167$ ($p < 0.001$) for the coda duration. In order to measure the strength of the correlation between these temporal measures of the monosyllabic CVC words and P-scores, linear regression analyses of P-scores and the durational measures were thus performed.

The results of the linear regression analyses are summarized in Figure 6.3, revealing several noticeable findings. First, to some extent, the variability of perceived prominence can be explained by changes in the temporal patterns of the monosyllabic CVC words. Among the various durational measures, the onset and the nucleus durations-whether normalized or not-account for a greater portion of the variability found in ordinary listeners' perception of prosodic prominence than does the coda duration, which contributes less to explaining the variability in P-scores. In partic-

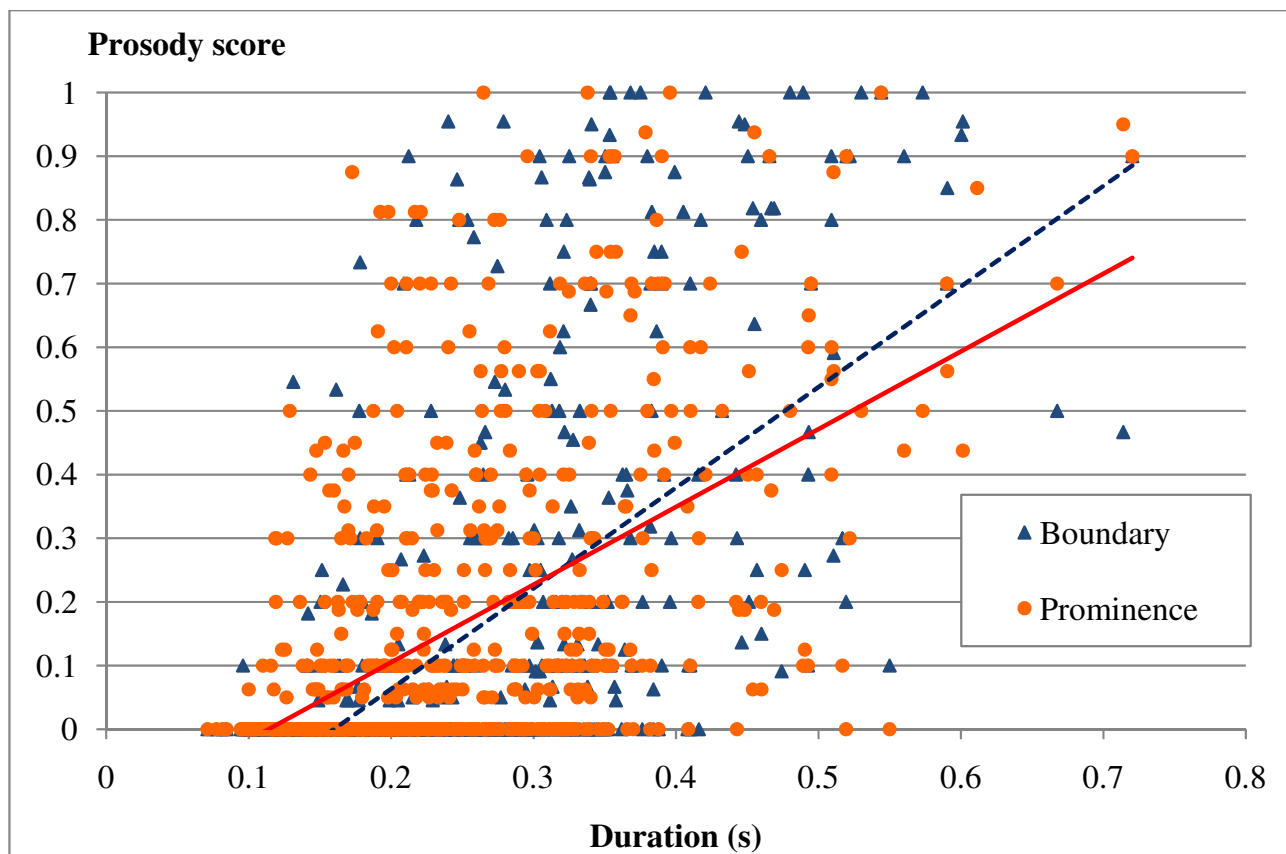


Figure 6.1: Scatterplot with regression lines between Prosody scores (P- and B-scores) and the raw durations (in seconds) of the monosyllabic CVC words

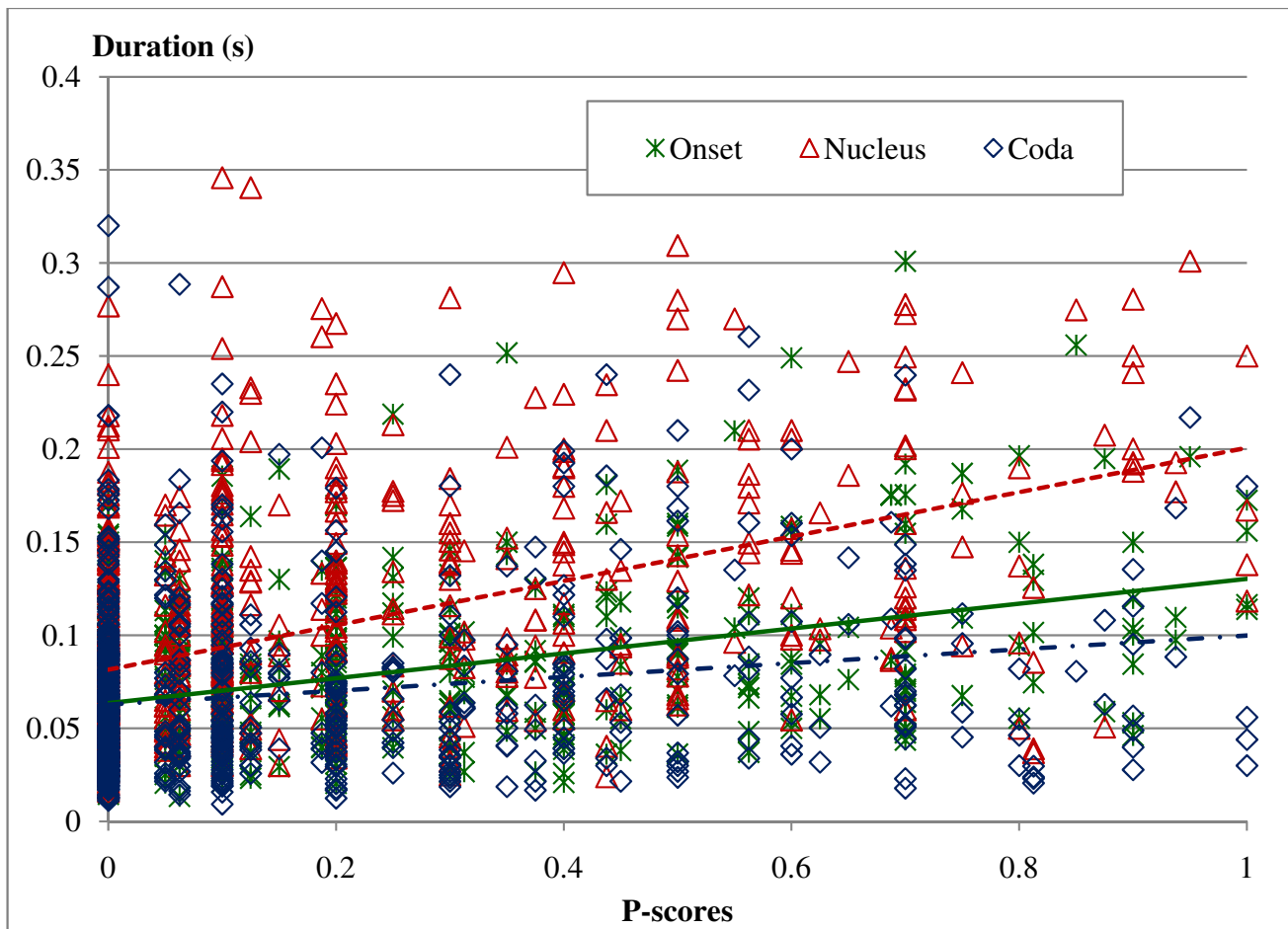


Figure 6.2: Scatterplot with regression lines between P-scores and the raw durations of the onset (—*—), nucleus (- - - Δ - -), and coda (- · - \diamond - · -)

ular, changes in nucleus durations account for the greatest amount of variability in P-scores. Second, comparing the measurements of raw durations with z -normalized durations, the raw durations better account for P-score variability, suggesting that ordinary listeners are sensitive to the raw durations of subcomponents, rather than to durations normalized relative to the mean and standard deviation of duration of each phone in the language at large. In addition, comparing the r^2 values of the models predicting P-scores from individual components of the syllable with those from a multiple linear regression of word duration as a whole, the former provides a better prediction rate for ordinary listeners' perception of prosodic prominence than the latter: r^2 is 0.275 with word duration as a single regressor and 0.304 with all three subsyllabic components as separate regressors. Last of all, stepwise multiple linear regression models demonstrate that coda duration does not play any role in explaining the variability of listeners' perception of prosodic prominence: in both simple and stepwise multiple linear regression models, 30.4% of variation in perceived prominence is explained from the raw onset, nucleus, and coda durations of monosyllabic words while 25.6% is predicted from the z -normalized onset, nucleus, and coda durations, and coda duration is not included as a predictor variable under either regression model.

6.3.2 How does prosodic boundary influence the temporal characteristics of the monosyllabic CVC words?

The total word durations of monosyllabic CVC words are positively correlated with perceived prosodic boundary scores (B-scores) as seen in Figure 6.1. That is, the raw durations of the monosyllabic CVC words are longer when ordinary listeners perceive the word as followed by a prosodic boundary. Similar to prominence effects on the internal temporal structure of the monosyllabic CVC words, Figure 6.4 demonstrates that the raw durations of all the subcomponents of the monosyllabic words are also

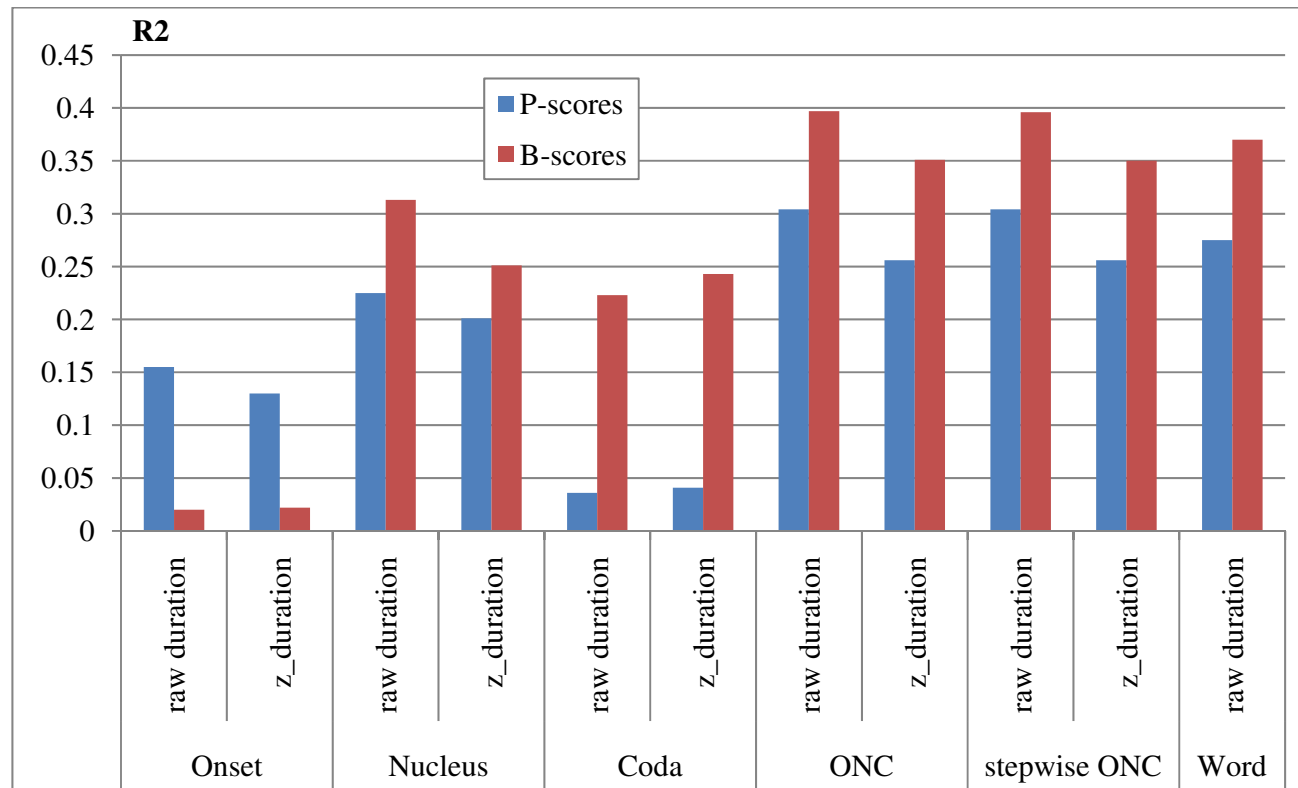


Figure 6.3: The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic prominence (oblique bars) and prosodic boundary (dotted bars), explained from the various durational measures of the monosyllabic CVC words

positively correlated with B-scores, indicating that the durations of the onset, nucleus, and coda of the monosyllabic CVC words increase when perceived as followed by a prosodic boundary by ordinary listeners and durational measures are all similar to one another when they are not likely to be perceived as followed by a prosodic boundary. Yet, in terms of the strength of the correlations between B-scores and the subcomponental durations, the duration of the nucleus of monosyllabic CVC words is strongly correlated with P-scores ($\rho=0.506$, $p < 0.001$), followed by the coda ($\rho=0.428$, $p < 0.001$) and the onset durations ($\rho=0.166$, $p < 0.001$) in order.

Linear regression analyses of perceived boundary were performed with the durational measures of the onset, nucleus, and coda of monosyllabic CVC words, including the raw word durations and the raw and z -normalized durations of the subcomponents of the monosyllabic words in order to see how strongly the perceived boundaries are correlated with those temporal measures of the monosyllabic CVC words. As summarized in Figure 6.3, a considerable proportion of the total variation of ordinary listeners' perception of prosodic boundary is taken into account on the basis of the durational measures of the monosyllabic words. The portion of the total variation (r^2) of ordinary listeners' perception of prosodic boundary that can be modeled by the raw durations of the monosyllabic CVC words is 0.370. After, comparing the regression models of B-scores based on durational measures with those of P-scores based on the same durational measures, the first finding to note is that more variation in perceived prosodic boundary is accounted for based on changes in the durational parameters—whether durational parameters are z -normalized or not, compared with the variation in prominence perception explained by the durational parameters. Furthermore, in terms of the magnitude of each subcomponent's contribution in the regression models, while listeners' response to prosodic prominence is mostly predicted by the durational measures of the onset and nucleus of the monosyllabic CVC words, the variability of perceived prosodic boundary is primarily explained by the durational measures of

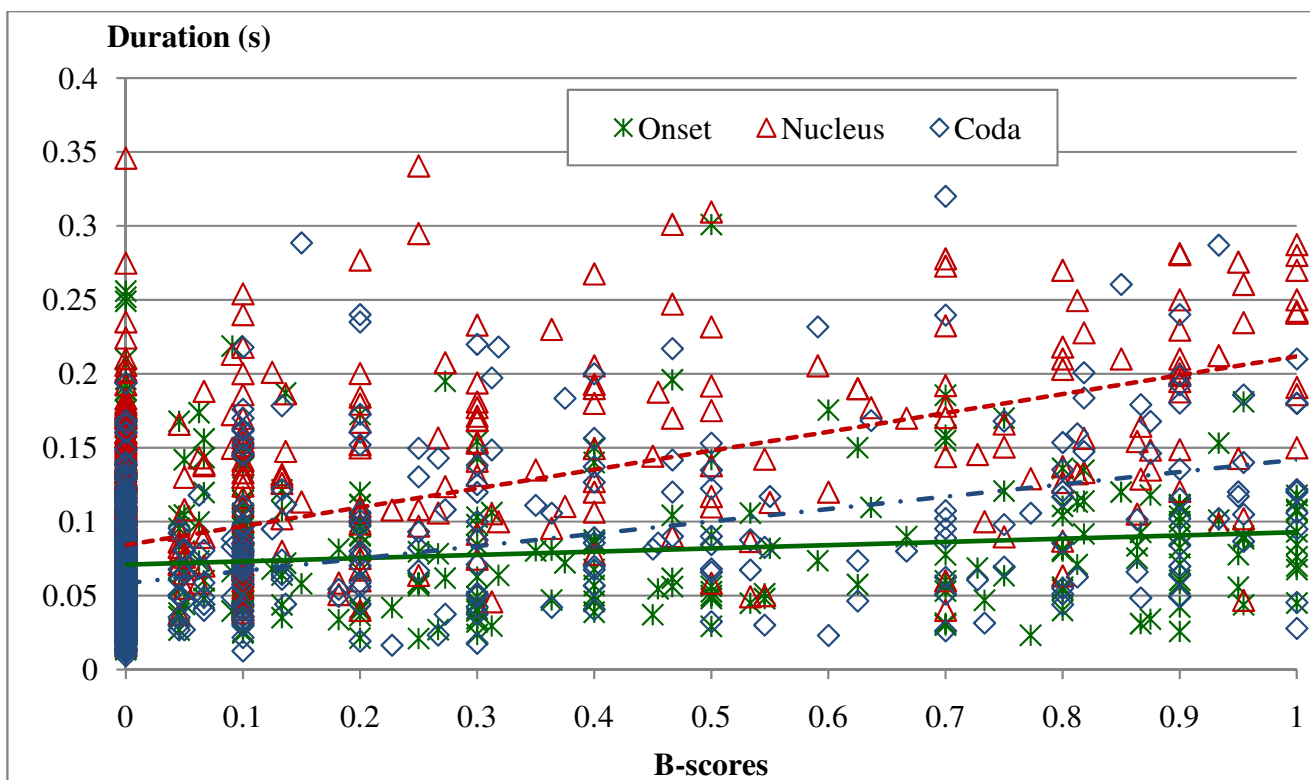


Figure 6.4: Scatterplot with regression lines between B-scores and the raw durations of the onset (—*—), nucleus (- - Δ - -), and coda (- · - \diamond - · -)

the nucleus and coda of the monosyllabic words: with raw durations, $r^2 = 0.02$ (onset), 0.313 (nucleus), and 0.223 (coda) and with z -normalized durations, $r^2 = 0.022$ (onset), 0.251 (nucleus), and 0.243 (coda).

Looking next at the contribution of the onset, nucleus, and coda, individually, to the variation in B-scores, it is the nucleus and the coda durations-whether they are normalized or not-that account for the greater amount of variability of ordinary listeners' perception of prosodic boundary, while the onset duration contributes only very little. In particular, as stated above, changes in the nucleus durations account for the greatest variability of listeners' perception of prosodic boundary, followed by the coda and the onset duration in order. Compared with the z -normalized durations, the raw durations provide a better model of listeners' perception of prosodic boundary, accounting for a greater portion of variance in listeners' responses. The results also show that, comparing the multiple linear regression model of boundary perception as predicted by word duration as a whole with a model that includes the individual durations for onset, nucleus, and coda as separate regressors, the latter accounts for a greater portion of the variation in listeners' perception of prosodic boundary than the former. The fact that there is little difference between r^2 values of the stepwise multiple linear regression models with onset, nucleus, and coda durations as predictors and the r^2 values of simple multiple linear regression models of the same durational measures suggest that the onset duration plays a little role in explaining the variability of listeners' perception of prosodic boundary. For example, the simple multiple linear regression model of B-scores with the raw durations of the onset, nucleus, and coda of the monosyllabic CVC words explains 39.7% of the variation found in listeners' boundary perception, while the stepwise multiple linear regression model accounts for 39.6% of the variation. That is, adding onset duration as a separate predictor to the regression model of boundary perception does not improve the model.

6.3.3 How do prosodic features influence the internal temporal structure of the monosyllabic CVC words in spontaneous conversational speech of American English?

Based on the findings presented above from Spearman's correlation and multiple linear regression analyses of durational measures and prosody scores, in this section I summarize the findings in relation to the prosodically influenced shaping of the internal temporal structure of the monosyllabic CVC words. Figures 6.5 and 6.6 demonstrate the relative effects of prosody on the temporal structure of the subcomponents of the monosyllabic CVC words. As P-scores increase, the ratio of nucleus duration to overall duration in monosyllabic CVC words increases, but the ratio of the onset levels off and the coda ratio decreases. As for the effect of prosodic boundary, when a word is perceived as preceding a prosodic boundary, the ratio of nucleus duration to the total duration of a monosyllabic word increases, while the ratio of the coda duration over the total duration of a monosyllabic CVC word remains almost the same and the onset durational ratio decreases.

6.4 Summary and Discussion

The current study examines whether prosodic context affects the internal temporal structures of the subcomponents of monosyllabic CVC words, and if the effects are uniform on the onset, nucleus, and coda. The current study investigates whether the temporal effects of prosodic prominence and boundary, as the two major sources of lengthening, are similar on the subsyllabic structures of monosyllabic CVC words. This study also examines to what extent temporal variation in monosyllabic CVC words by itself can take into account the variability of perceived prosodic promi-

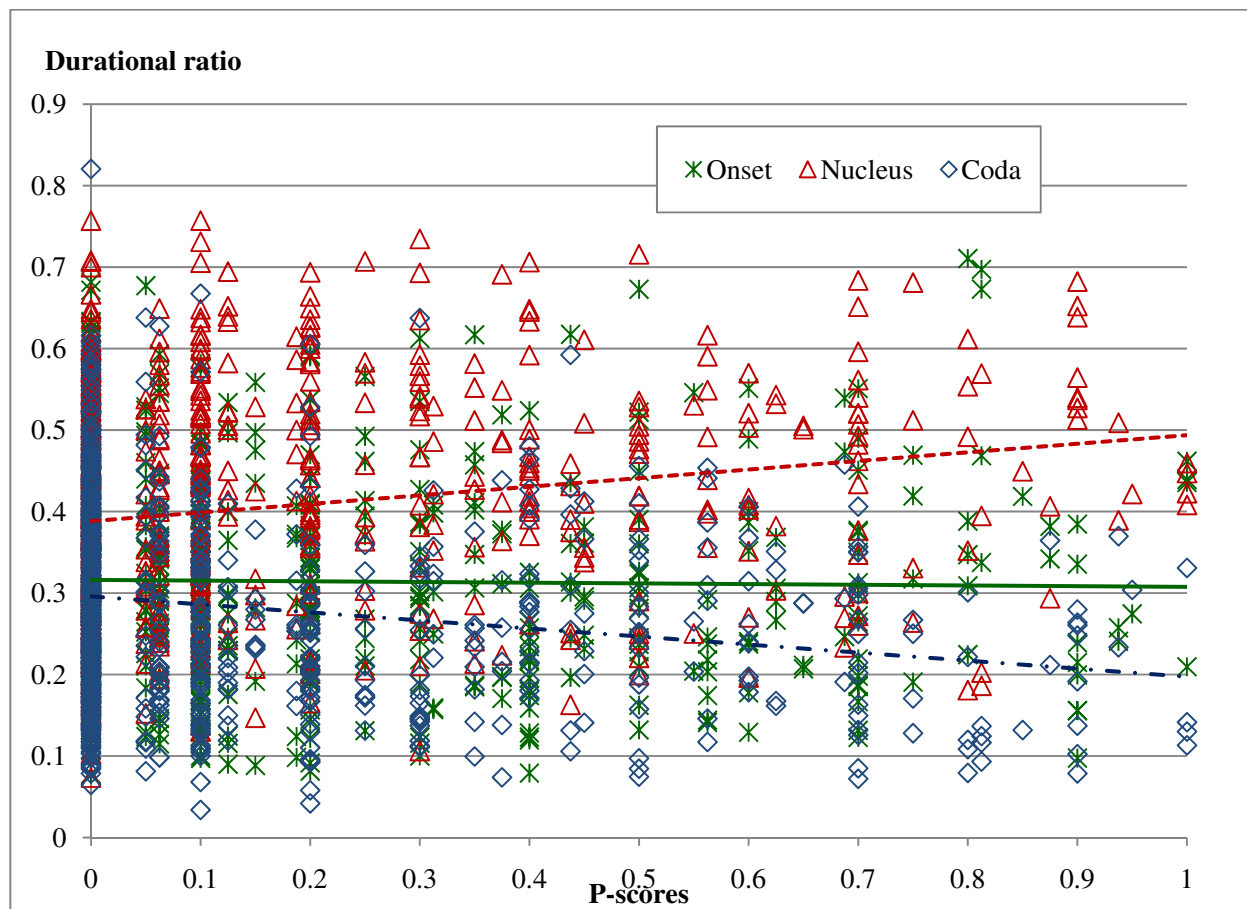


Figure 6.5: Scatterplot with regression lines between P-scores and the ratio of the onset (—*—), nucleus (---△---), and coda (—·—◇—·—) duration to syllable duration in the monosyllabic CVC words

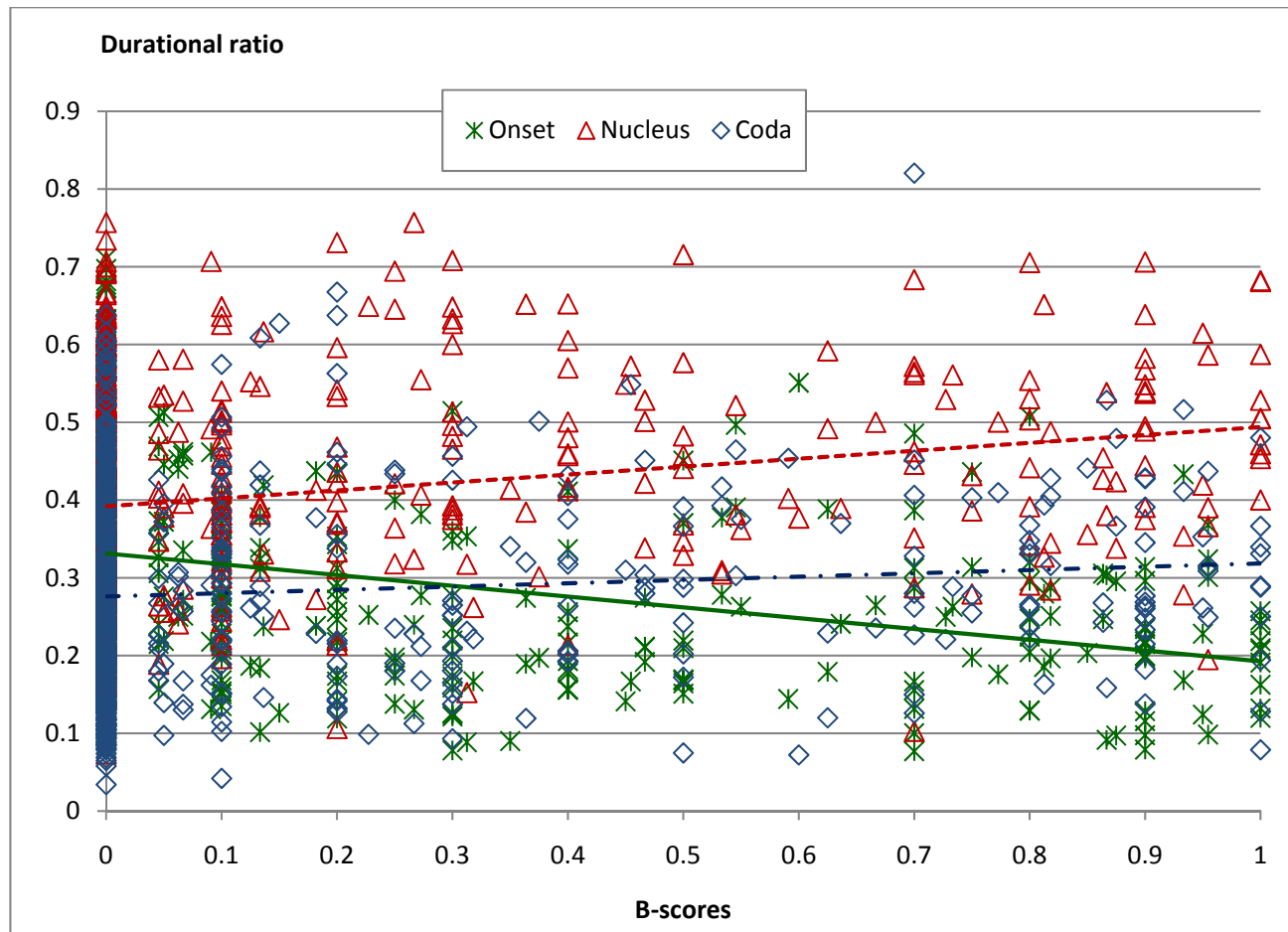


Figure 6.6: Scatterplot with regression lines between B-scores and the ratio of the onset (—*—), nucleus (- - - Δ - - -), and coda (- · - \diamond - · -) duration to syllable duration in the monosyllabic CVC words

nence and boundary, further evaluating which durational measures are better cues for prosodic prominence and boundary.

The findings from this study show that the durations of monosyllabic CVC words tend to be longer for words that listeners perceive as prominent or followed by a boundary. Moreover, when a word is perceived as prominent or followed by a boundary, the raw and z -normalized durations of all the subcomponents of the monosyllabic CVC words increase. Yet, in terms of the magnitude of temporal effects of prosodic context, the findings from the current study reveal that raw durations are better predictors of perceived prosody, including both prosodic prominence and boundary, compared to z -normalized durations. Due to data sparseness, in this study I did not z -normalize word durations of the monosyllabic CVC words, and z -normalized durations of the overall duration of monosyllabic words were excluded from the comparisons. These findings suggest that when ordinary listeners make a judgment on the locations of prosodic prominence and boundary, they need not rely on duration measures assessed relative to the duration patterns that a particular speaker produce for individual phones. Instead, ordinary listeners perceive prosodic prominence and boundary when the absolute raw duration of a word is long, and are relatively insensitive to the phonetic identity of all the subcomponents of the word. It is also possible that prosodic effects on duration measures of monosyllabic words override temporal variation resulting from the types of phone. For example, when perceived as prominent or followed by a prosodic boundary, the duration of the front high vowel /i/, which is intrinsically short, is longer than the duration of the low back vowel /ɑ/ in a word that is not prominent or phrase-final. Furthermore, these two possibilities are not incompatible and further examination is required.

It was also shown that the magnitude of temporal effects of prosodic context is not uniform over all syllable positions within monosyllabic CVC words. Regarding the effects of perceived prosodic prominence as a function of position within a word,

nucleus duration shows the greatest increase due to prominence, followed by the durations of the onset and of the coda, respectively. These findings are consistent with those of Greenberg et al. (2003), confirming that the durations of subcomponents of the syllable are affected by prosodic prominence in spontaneous speech, as it is perceived by ordinary listeners. Similarly, regarding the effects of perceived prosodic boundary, nucleus duration again shows the largest lengthening effect. Contrary to the effects of prosodic prominence, codas showed a greater lengthening effect than onsets when perceived as preceding a prosodic boundary.

The asymmetric effects of prosodic structure on temporal patterns observed here suggest that the underlying production mechanisms of prosodic prominence are different from those of prosodic boundary, as proposed in the previous sections (Chapters 4 and 5). In other words, under prosodic prominence, speakers actively lengthen the duration of the onset and nucleus of a word, the components which arguably play important roles in word identification, and this lengthening is greater than the lengthening effect on coda duration. Yet, before a prosodic boundary, the duration of the nucleus and coda more greatly increases, likely reflecting a slowing down of articulator movement. Such slow-down appears to be strongest for lexically stressed vowels. However, it cannot be well explained why the nucleus but not the coda duration lengthens the most from the production mechanisms proposed in this study and moreover, the findings from the current study are somewhat contradictory to those from prior studies in which it was found that the segment nearest to the boundary is lengthened to the largest degree, and the magnitude of lengthening by boundary decreases as the distance from a boundary increases. This difference might be due to the nature of the speech materials used in the current study, which are excerpted from conversational speech. The number and the location of prosodic prominences and boundaries may differ in spontaneous vs. scripted speech materials, and the frequency of words that are both prominent and phrase-final, which is common in our

materials, may also be different from that found in scripted laboratory materials. To further explore this matter, we will have to analyze a larger corpus from which all words with high P-scores and high B-scores are excluded.

The findings that the multiple linear regression model with the total duration of a word as a single predictor accounts for less variability in prosody perception than the regression model with the duration of each subsyllabic component as a separate predictor suggests that word duration is not processed as one percept but rather as three separate percepts in the perception of prosodic features. In other words, listeners do not attend to lengthening of word duration as a whole, but are sensitive to lengthening of each subsyllabic component within the monosyllabic CVC word when they assess the presence or absence and the location of prosodic features in spontaneous conversational speech. For example, if one subsyllabic component of a word is extraordinarily lengthened, then listeners are likely to hear that word as prominent or as followed by a boundary even if the duration of that word is not much lengthened.

How does prominence- and boundary-related lengthening modulate the temporal structures of the syllable of the monosyllabic CVC words? The finding that prosodic effects on the internal temporal structure of the monosyllabic CVC words are not uniform over all syllable positions and that prosodic effects most greatly affect the durations of the nucleus shows that both prosodic prominence and boundary influence the temporal implementation of the monosyllabic CVC words, and further that the effects are primarily restricted to the nucleus. The models of prosodic effects on the internal temporal structure of the monosyllabic CVC words are illustrated in Figure 6.7 and Figure 6.8. As shown in Figure 6.7, the durations of each subsyllabic component within a monosyllabic CVC word are almost equal when they are not prominent—the onset, nucleus, and coda each take up around 30% of the overall word duration. On the other hand, the overall word duration and the duration of

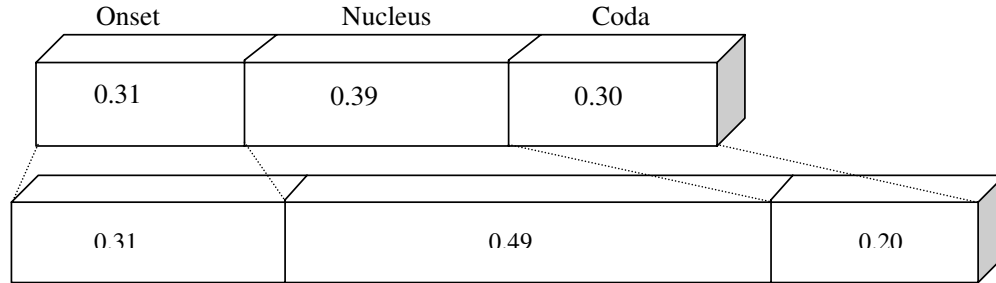


Figure 6.7: Schematic representation of temporal structure of the monosyllabic CVC word: non-prominent word (P-score = 0, top) vs. prominent word (P-score = 1, bottom)

each subsyllabic component of the monosyllabic CVC words are lengthened when prominence scores increase. Yet in terms of durational ratio of the onset, nucleus, and coda, the proportion of the nucleus duration increases to almost 50% of the word duration, consuming the coda portion. The relative duration of the onset remains the same. Figure 6.7 displays a schematic representation of the internal temporal structure of monosyllabic CVC words when perceived as followed by a prosodic boundary, compared with the temporal structure when they are perceived as in phrase-medial position. Similar to prominence effects, the presence of a prosodic boundary lengthens the overall duration and each subsyllabic component's duration of the monosyllabic CVC words. Looking closely, the durations of onset, nucleus, and coda relative to total syllable duration are almost equal for words that are perceived as not preceding a boundary. The relative duration of the nucleus of the syllable greatly increases up to almost 50%, taking over the onset proportion in a word perceived as followed by a prosodic boundary by ordinary listeners. The proportional duration of the coda remains unchanged in words that are followed by a boundary.

How much variation in ordinary listeners' perception of prosody can be accounted for on the basis of the temporal patterns of monosyllabic CVC words? As illustrated in Figure 6.3, the current study concludes that the variation in durational parameters of

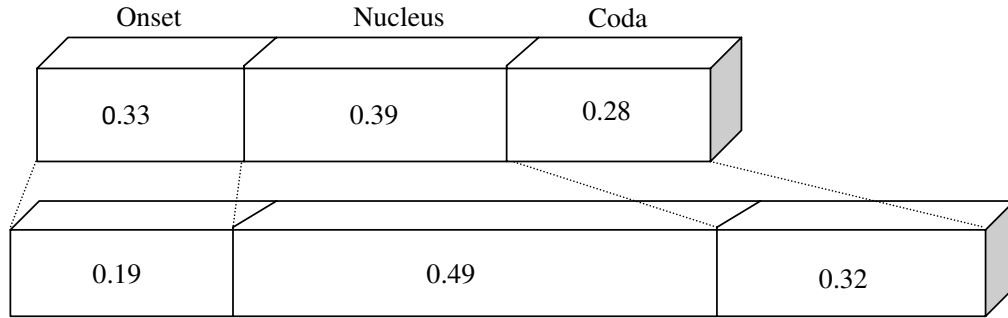


Figure 6.8: Schematic representation of temporal structure of the monosyllabic CVC word: phrase-medial word (B-score = 0, top) vs. phrase-final word (B-score = 1, bottom)

monosyllabic words can itself predict around 30% of listeners’ perception of prosodic prominence and around 40% of boundary perception. For perceived prominence, only the duration of the onset and nucleus of a syllable within a monosyllabic CVC word plays an important role, while the nucleus and coda duration cues the occurrence of a prosodic boundary. Thus, we find that the internal temporal structure of the monosyllabic CVC words is manipulated by prosodic context, and these changes in the temporal structure of a word guide listeners to perceive the location of prosodic prominence and boundary in spontaneous conversational speech of American English.

6.5 Conclusion

Findings from Spearman’s correlation and multiple linear regression analyses of prosody scores with the various durational measures of a word and its subcomponents show that prosodic features modulate the temporal structures of the subcomponents of the monosyllabic CVC word in everyday conversational speech of American English. Ordinary listeners are sensitive to the temporal patterns of onset, nucleus, and coda structures in their perception of prosody in everyday conversational speech. More specifically, listeners perceive prosodic features by integrating temporal information

from each subsyllabic component of a word, not by responding to word duration as a whole. In terms of the internal subsyllabic structure of the monosyllabic CVC word in relation to prosodic structure, when words are perceived as prominent, the proportional duration of the nucleus increases while that of the coda decreases; when words are perceived as followed by a boundary the proportional duration of the nucleus increases while that of the onset decreases.

Chapter 7

How do Ordinary Listeners Perceive Prosodic Features? Syntagmatic VS. Paradigmatic Comparison

7.1 Introduction

The current study corroborates findings most prior studies of “laboratory” speech (e.g., simple sentences, read speech) in demonstrating that prosody is encoded in the acoustic signal in suprasegmental properties and also in the properties that encode phonological segmental features. Ordinary listeners attend to these acoustic cues to prosodic form in online realtime tasks of prosody perception. The current chapter examines how ordinary listeners glean necessary acoustic information from long stretches of speech in order to perceive the presence or absence of prosodic features and their locations in spontaneous conversational speech. This study further asks whether the mechanisms that underlie the perception of prosodic prominence are different from those of boundary perception. First, this study poses the following questions about the acoustic information that ordinary listeners attend to when perceiving prosodic features: (1) Do ordinary listeners rely on changes in the raw measures of acoustic features? Or, (2) do they identify the presence or absence of prosodic features based on changes in acoustic measures in relation to a baseline? That is, do ordinary listeners rely on the raw acoustic measures or on the normalized ones in prosody perception? Additionally, the current study further explores what acoustic baseline if any ordi-

nary listeners refer to: (3) Do such listeners normalize acoustic information relative to each speaker’s phonetic space? That is, do they utilize stored speaker-specific acoustic information in prosody perception-performing paradigmatic comparisons of acoustic measures among tokens of phones belonging to the same phone class? Or (4) do they perceive prosodic features by comparing acoustic information in relation to the local surrounding context of the utterance? That is, do they identify the presence or absence as well as the locations of prosodic features by attending to acoustic changes in the local context regardless of their phone class-performing syntagmatic comparison?

To date, there have been no studies directly investigating the normalization domains that ordinary listeners employ in the perception of prosodic features. In most prior laboratory studies, acoustic and articulatory materials are either controlled to reduce or eliminate variation due to phone identity, or measures are normalized within phone categories, comparing normalized measures across prosodic categories (e.g., Beckman and Edwards, 1994; Beckman et al., 1992; Byrd, 2000; Byrd and Saltzman, 1998, 2003; Cho, 2002, 2005, 2006, 2008; Cooper et al., 1985; Eady and Cooper, 1986; Heldner, 2001a; Turk and Shattuck-Hufnagel, 2007; Sluijter and van Heuven, 1995, 1996b; Turk and Sawusch, 1996, 1997; Turk and White, 1999; van Bergem, 1993; Wightman et al., 1992). For example, in a series of production studies examining the influence of focal accents on duration and F_0 (Cooper et al., 1985; Eady and Cooper, 1986), the production of the same sequences of words as well as the locations of focal accents were controlled. Turk and Shattuck-Hufnagel (2007) guided participants to produce the same strings of words with the desired prosodic structure by inserting a parenthesis as an indication of the locations of phrase boundaries. In their articulatory studies, Beckman and her colleagues (1992; 1994), Byrd and her colleagues (2000; 2003), and Cho (2005; 2006) investigated articulatory variation attributed to the presence (or absence) as well as the location of prosodic features by locating tar-

get nonsense words having a very limited set of phones, e.g., /ɑ, ə/ in Beckman and her colleagues (1992; 1994), *baba* in Byrd and her colleagues (2000; 2003) and /a, i/ in Cho (2005, 2006) at the target locations. In sum, a major body of literature in the acoustic and the articulatory implementation of prosody has compared acoustic and/or articulatory characteristics of the same phone under different prosodic environments after prompting participants to produce the target words with the desired prosodic structure.

A large group of perception studies employing laboratory speech or corpus materials have also controlled the phone types under investigation (e.g., Cambier-Langeveld et al., 1997; Fry, 1958; Gussenhoven et al., 1997; Hermes and Rump, 1994; Kohler, 2008; Lehiste and Fox, 1993). For example, Gussenhoven et al. (1997) manipulated the F_0 contours of a given sentence, *Amanda gaat naar Malta.*, in their investigation of the role of F_0 declination as a cue for prominence. In Kohler (2008), the duration and F_0 of a synthesized word, *baba* (obtained by replacing the second *ba* with the first *ba* from a natural production of a series of the simple word, *baba*) were manipulated. Even in corpus studies, acoustic variation associated with prosodic features has been investigated by phone category or by sentence across speakers (e.g., Carlson and Swerts, 2003a; Kim et al., 2006; Yoon, 2010). To sum up, no matter what kind investigation was performed (production, perception, or corpus studies), the phonetic characteristics associated with prosody have traditionally been paradigmatically investigated, assuming that it is not appropriate to compare phonetic measures from phones that have intrinsically different phonetic characteristics, which might either diminish or interfere with prosodic effects.

The current study begins with a linguistically naïve question: How do ordinary listeners perceive prosodic features in everyday communication? Do they need to normalize acoustic parameters when perceiving prosodic features? Do they employ the same comparison domains for the perception of prosodic prominence as for boundary?

In answering these questions, I consider two possible ways that a listener may identify the prosodic structure of a given utterance: through paradigmatic and syntagmatic comparison. In paradigmatic comparisons, ordinary listeners perceive prosodic features by referring to phonetic variation relative to a speakers' phone-specified phonetic space. In syntagmatic comparisons, on the other hand, ordinary listeners identify prosodic features by referring to changes in phonetic parameters in the local context. Taking into account the fact that, in everyday communication, a listener is rarely given multiple opportunities to listen to a particular sequence of words, and that, given such short time in the task of perception, a listener may not be able to establish a speaker specific phonetic space, we may hypothesize that ordinary listeners will identify prosodic features by attending to changes in the patterns of phonetic parameters in the local context.

An alternative hypothesis is also possible, in light of the observed differences in prominence and boundary perception. Findings presented in earlier chapters, based on Spearman's correlation and stepwise multiple linear regression analyses, suggest that the underlying mechanisms of prominence production are different from those of prosodic boundary production. Prosodic prominence is signaled through changes in the patterns of a combination of acoustic measures including duration, intensities in mid- and high-frequency bands, F_0 , F_1 and F_2 , while prosodic boundary is signaled by temporal changes, in particular by lengthened segmental duration and the presence of a silent pause. Moreover, when a silent pause is present after a word, the perception of prosodic boundary relies primarily on that cue. These findings suggest that listeners must employ different perception domains for prosodic prominence than for boundary. That is, in the perception of prosodic prominence, listeners must attend to changes in acoustic measures in the local domain, while in boundary perception, they look for the presence of a silent pause after a word and are only attentive to other acoustic parameters if such a pause is not present. If this is the case, then

prominence perception may well be different from boundary perception. Variation in continuous acoustic parameters should be noticed in order to perceive prosodic prominence, while, on the other hand, the presence of a silent pause, if any, is a primary cue for prosodic boundary, and listeners do not need to attend to acoustic information from other parameters in boundary perception.

In sum, the central goal of the present study is to evaluate how ordinary listeners perceive prosodic features in realtime tasks of prosody annotation in spontaneous conversational speech of American English, and, further, whether the underlying perceptual mechanisms for prosodic prominence are the same as those for prosodic boundary. Acoustic measures normalized paradigmatically and syntagmatically, as well as raw acoustic measures, are evaluated in order to determine which one best accounts for ordinary listeners' responses to prosodic features.

7.2 Normalization

The acoustic measures included in the present study are as follow: vowel duration, intensities (overall and subband in 0–500, 500–1000, 1000–2000, 2000–4000 Hz), and local F_0 maximum. These acoustic measures were all extracted from lexically stressed vowels for prominence analysis as well as from word-final, lexically stressed vowels for boundary analysis. As discussed below, the extracted acoustic measures were normalized in two different ways by employing Labanov's z -normalization of which the validity was tested for acoustic investigation in Adank et al. (2004).

7.2.1 Paradigmatic normalization

In paradigmatic normalization, acoustic measures are compared with the same kinds of acoustic measures from the same lexically stressed vowels produced by a single speaker, hypothesizing that ordinary listeners utilize stored information from each

speaker’s acoustic space for each phone. This normalization assumes that speaker-specific and phone-specific acoustic information is accumulated through exposure to a sufficient amount of speech to establish the speaker’s phone space and patterns of phonetic variation. Additionally, it is assumed that, when perceiving prosodic features in everyday conversations, listeners make decisions on the presence and the locations of prosodic features based upon their memory representations of the acoustic information from each phone, in particular for the lexically stressed vowels in this study, and based on those stored acoustic values, a threshold may be established for each phone when identified as prominent or in the context of a prosodic boundary. Another possibility is long-term speaker adaptation if a speaker’s speech is long enough to allow the listener to adapt to the speaker-specific characteristics of each vowel. In both cases, ordinary listeners will identify prosodic features, relying on phone-specific acoustic information.

Based on the assumptions above, each acoustic measure was z -normalized by phone within each speaker, using the mean and standard deviation over all instances of each vowel in the combined excerpts from a single speaker. The log-transformed acoustic measures were also z -normalized as in equation 7.1. Another way to paradigmatically normalize acoustic measures is the gamma-normalization shown in equation 7.2. As a result, three different kinds of paradigmatically normalized acoustic measures have been prepared: z -normalized forms of raw and log-transformed acoustic measures and gamma-normalized acoustic measures.

$$z_{ij}^{log} = \frac{\log x_{ijk} - \overline{\log x_{ij}}}{std(\log x_{ij})} \quad (7.1)$$

where $\log x_{ijk}$ is the log-transformed k -th actual acoustic value of the j -th phone from the i -th speaker, $\overline{\log x_{ij}}$ and $std(\log x_{ij})$ is the mean and the standard deviation of log-transformed acoustic value of j -th phone from i -th speaker, in order, and z_{ij}^{log} is the z -normalized value of the log-transformed k -th actual acoustic measure, x_{ijk} .

$$\gamma_{ijk} = \frac{x_{ijk}}{\overline{x_{ij}}} \quad (7.2)$$

where x_{ijk} is the k -th actual acoustic value of the j -th phone from the i -th speaker, $\overline{x_{ij}}$ is the mean acoustic value of j -th phone from i -th speaker, and γ_{ij} is the gamma-normalized value of the k -th actual acoustic measure, x_{ijk} .

7.2.2 Syntagmatic normalization

In comparison with paradigmatic normalization, syntagmatic normalization is performed in the local context. That is, acoustic measures from a target phone are compared with corresponding acoustic measures from its neighboring phones regardless of their phone class. This normalization assumes that it is not necessary for ordinary listeners to store all acoustic information for each phone when identifying prosodic features in everyday communication. It is instead assumed that ordinary listeners attend to changes in the patterns of acoustic parameters in the local context in perceiving prosodic features. On the basis of these assumptions, it is hypothesized that a listener may not need to hear a large amount of speech or to establish a speaker-specific inventory of a specific vowel, syllable, or word, but rather that listeners make an instantaneous decision of the presence or absence of prosodic features, responding to acoustic variation in a small comparison domain, which moves over the course of speech. In short, listeners are sensitive to relative changes of acoustic measures in a sequence of incoming sounds when compared with neighboring sounds in the perception of prosody.

The study of syntagmatic normalizations in this study is innovative because, to my knowledge, there is no prior study in which the size of the prominence processing domain is explored. In exploratory syntagmatic normalization, the appropriate size of normalization domain (3 adjacent vowels, 5 adjacent vowels, 3 adjacent words and

3 adjacent stressed vowels) was tested to see whether the syntagmatic normalization is better than or at least comparable to the paradigmatic normalization in terms of its power to explain prominence perception by ordinary listeners. Moreover, if this is the case, the domain size that will be the best fit for the perception of prosodic prominence will also be discussed. Later, two syntagmatic normalization methods with a different window size (5 adjacent vowels and 3 adjacent stressed vowels) are evaluated in order to see which normalization method can better account for variability in listeners' perception of prosodic boundary.

7.3 Results

In the following sections, findings from simple and stepwise multiple linear regression analyses of P-scores and various acoustic measures are reported, followed by findings from regression analyses for B-scores. The acoustic measures cited below include raw values as well as paradigmatically- and syntagmatically-normalized values with different size comparison windows.

7.3.1 Prosodic prominence

A number of the total variations in P-scores accounted for based upon raw acoustic measures and those that have been normalized are summarized in Figure 7.1. Comparing multiple linear regression models of perceived prominence with raw acoustic measures, or paradigmatically- or syntagmatically-normalized acoustic measures, it is revealed that the syntagmatically-normalized acoustic measures generally better account for the variation in listeners' perception of prosodic prominence (r^2 ranges from 0.146 in a window of 3 adjacent words to 0.203 in a window of 5 adjacent vowels) than do the paradigmatically-normalized acoustic measures (r^2 ranges from 0.096 when gamma-normalized to 0.117 when z -normalized with log-transformed acoustic

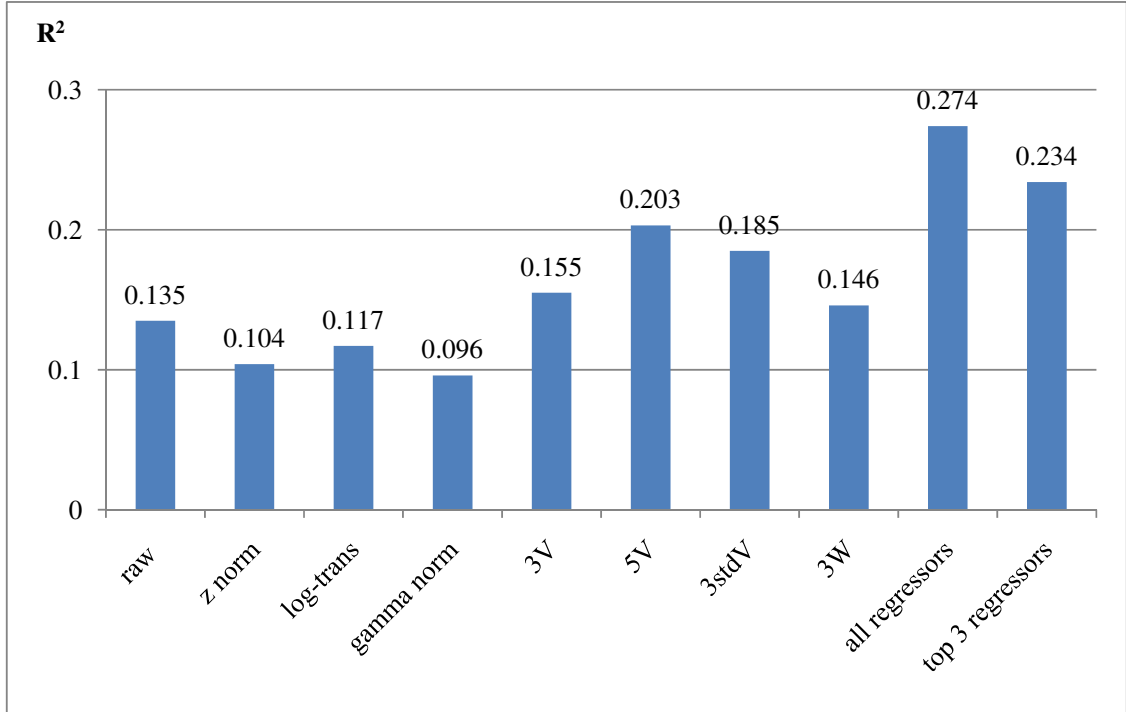


Figure 7.1: Distribution of the total variations (r^2) of the ordinary listeners' perception of prosodic prominence, explained from the acoustic measures: raw, paradigmatically (z -, $z(\log)$ -, and γ -), and syntagmatically (window size: 3 adjacent vowels, 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words) normalized acoustic measures from left to right.

measures). Looking closely, the best regression model of perceived prominence was obtained by employing acoustic measures normalized in a dynamic window of 5 adjacent vowels ($r^2 = 0.203$). The second best model was created when the acoustic measures were normalized in a dynamic window of 3 adjacent stressed vowels.

When putting all acoustic measures (including all raw and paradigmatically- and syntagmatically-normalized vowel duration, overall and subband intensities, and local F_0 maximum) into one simple multiple linear regression model of perceived prominence, 27.4% of the total variation (r^2) in P-scores is accounted for. From all 7 raw acoustic measures and their 21 corresponding paradigmatically- and 28 syntagmatically-normalized acoustic measures, the 3 top predictors were selected:

(1) overall intensity normalized in a dynamic window of 5 adjacent vowels; (2) raw vowel duration; and (3) local F_0 maximum normalized over a 3 adjacent stressed vowels. Using only these top three predictors in a new multiple linear regression model, 23.4% of the variability in P-scores is explained. That is, the top 3 acoustic parameters can account for 85.4% of the total variation (23.4%/27.4%) that can be explained by all the acoustic measures together. Upon further analysis, the contribution of each acoustic measure is as follows: overall intensity normalized in a dynamic window of 5 adjacent vowels (10.4%), raw vowel duration (9.4%), and local F_0 maximum normalized over a 3 adjacent stressed vowels (3.6%).

Subsequent stepwise multiple linear regression analyses were performed as summarized in Figure 7.2. The results indicate that, first, raw vowel duration on its own accounts for the second largest amount of variation ($r^2 = 0.103$) of perceived prominence among all vowel duration measures. That is, slightly over 10% of the variation in ordinary listeners' response to prosodic prominence is accounted for based solely upon the absolute duration of lexically stressed vowels. Secondly, the contribution of overall intensity measure drastically increases when it is normalized relative to the local context. As shown in Figure 7.2, the raw measure and paradigmatically normalized measures of overall intensity do not contribute much to cueing prosodic prominence. Yet, when employing syntagmatic normalization, overall intensity plays an important role in prominence perception; in particular, when overall intensity measures are normalized with a dynamic window of 5 adjacent vowels or 3 adjacent words, the largest variation in P-scores from overall intensity is taken into account ($r^2 = 0.104$ in a 5 vowel domain and $r^2 = 0.092$ in a 3 word domain). In fact, overall intensity normalized in a dynamic window of 5 adjacent vowels is the most powerful cue for perceived prominence among all the acoustic measures employed in this study. Lastly, the results also demonstrate that the role of local F_0 maximum can be evidenced only when it is normalized relative to its local context—namely, in a window

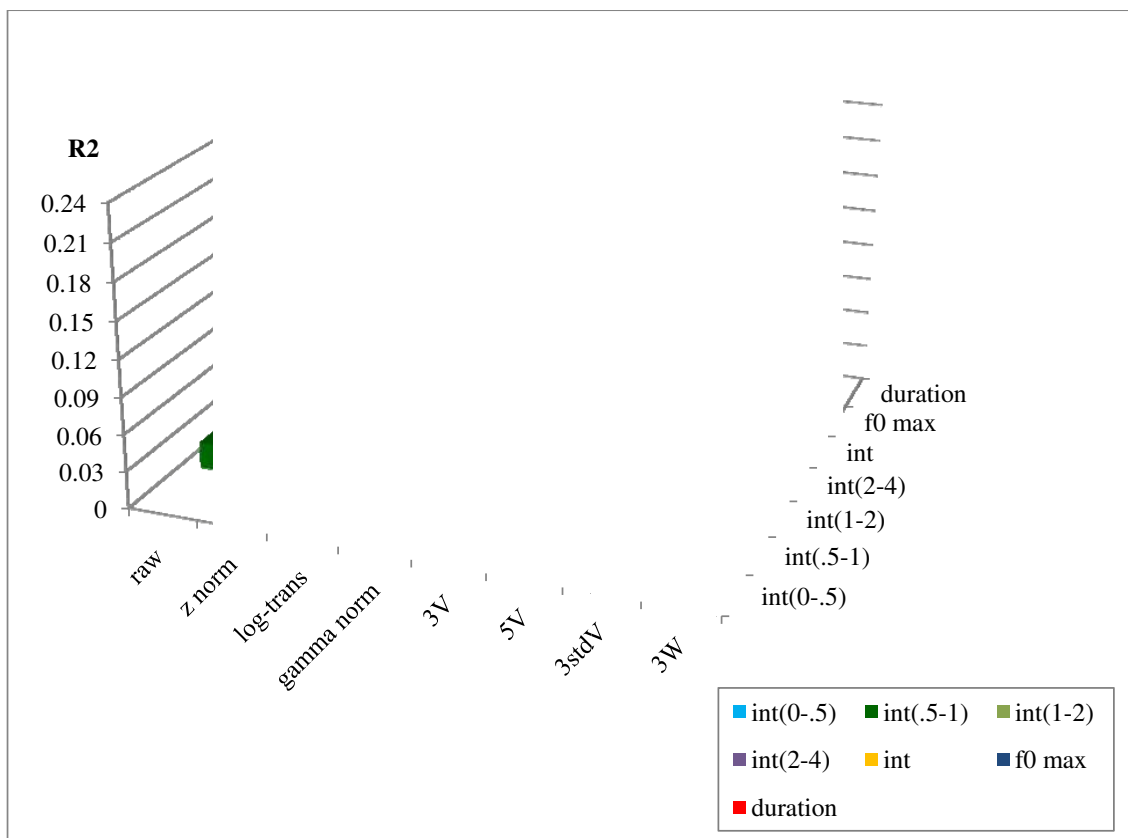


Figure 7.2: Contribution of each acoustic measure (duration, local $F_{0,max}$, overall intensity, subband intensities in 0–500, 500–1000, 1000–2000, and 2000–4000 Hz) to predicting ordinary listeners’ response to prosodic prominence: with raw acoustic measures, z -normalized acoustic measures by phone, z -normalized log-transformed acoustic measures by phone, gamma-normalized acoustic measures by phone, syntagmatically normalized acoustic measures over a dynamic window of 3 adjacent vowels, 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words, in order from left to right.

of three adjacent vowels ($r^2 = 0.08$) and in a window of 3 adjacent stressed vowels ($r^2 = 0.092$).

In summary, simple and stepwise multiple linear regression analyses indicate several notable findings: (1) Over 27% of the variability in listeners' responses to prosodic prominence is predicted on the basis of all the acoustic measures including raw, paradigmatically-, and syntagmatically-normalized ones, and 3 out of the total 56 acoustic measures explain most of variation in listeners' responses to prosodic prominence: 24.3% of the variation (85.4% of the total variation, 0.234/0.274). (2) In general, syntagmatically-normalized acoustic measures can better account for the variation in perceived prominence than raw or paradigmatically-normalized acoustic measures. (3) As a single cue for prosodic prominence, overall intensity syntagmatically normalized in a dynamic window of 5 adjacent vowels is the best predictor of prominence perception, followed by raw vowel duration, and local F_0 max in a 3 adjacent stressed vowels. (4) As for vowel duration, no matter what forms of normalization (raw vs. syntagmatic vs. paradigmatic) is employed for regression analysis, it is evidenced that vowel duration plays a consistently important role in the perception of prosodic prominence. (5) Overall intensity and local F_0 maximum are sensitive to different ways of normalization and the size of the comparison window: only when syntagmatically normalized, overall intensity and F_0 max contribute more to cueing prosodic prominence.

7.3.2 Prosodic boundary

Figure 7.3 summarizes the results of simple and stepwise multiple linear regression analyses, illustrating a number of the total variations between B-scores and different forms of acoustic measures including raw and paradigmatically- and syntagmatically-normalized vowel duration, overall intensity, subband intensities in 0–500, 500–1000, 1000–2000, and 2000–4000 Hz, and local F_0 maximum. In terms of variation in per-

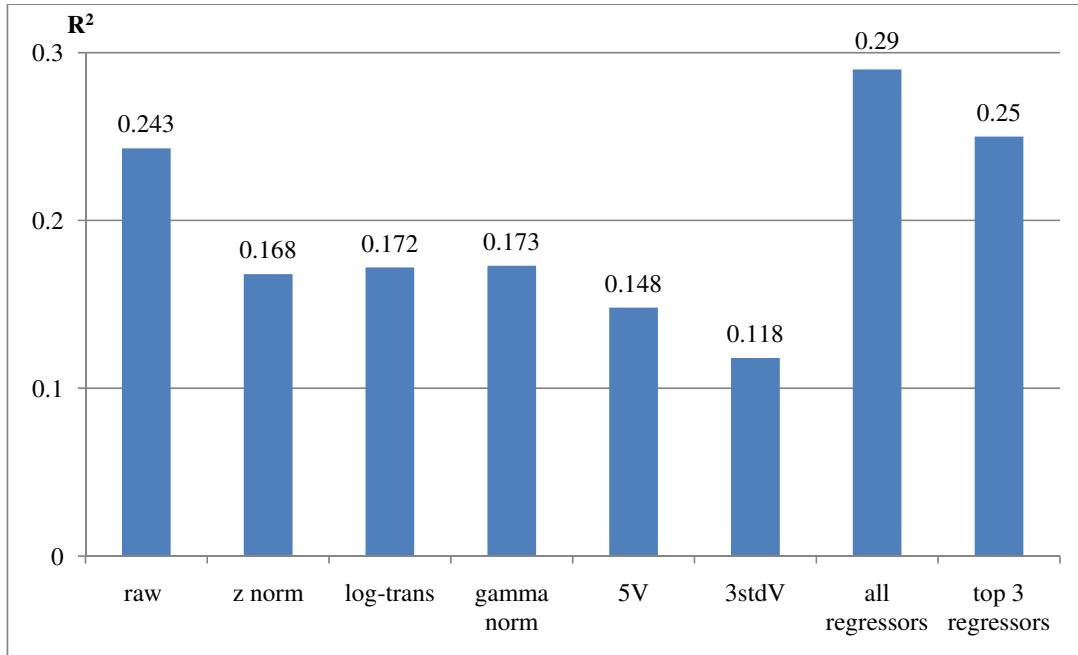


Figure 7.3: Distribution of the total variations (r^2) of the ordinary listeners' perception of prosodic boundary, explained from the acoustic measures: raw, paradigmatically (z -, $z(\log)$ -, and gamma-), and syntagmatically (window size: 5 adjacent vowels and 3 adjacent stressed vowels) normalized acoustic measures from left to right.

ceived boundary, the multiple linear regression model of B-scores with raw measures of vowel duration, intensity measures, and local F_0 maximum explains the largest portion of the variation ($r^2 = 0.243$) in B-scores. In terms of normalization methods, the paradigmatically-normalized acoustic measures ($r^2 = 0.168$ by z -normalization, $r^2 = 0.172$ by z -normalization after log-transformation, and $r^2 = 0.173$ by gamma normalization) are always better able to account for variation in B-scores compared to syntagmatically-normalized ones ($r^2 = 0.118$ in a dynamic window of 3 adjacent stressed vowels and $r^2 = 0.148$ in a 5 adjacent vowel window).

Figure 7.3 also shows that 29% of the variation (r^2) in B-scores is taken into account based upon all the acoustic parameters including raw and paradigmatically- and syntagmatically-normalized acoustic measures. Yet, upon closer investigation, out of

7 raw, 21 paradigmatically normalized, and 14 syntagmatically normalized acoustic measures, the top 3 acoustic measures explain around 25% (86.2%, 0.250/0.290) of variation in B-scores: raw vowel duration (23%), overall intensity normalized in a 5 adjacent vowel window (1.4%), and z -normalized subband intensity in 0–500 Hz (0.7%). Subsequent stepwise multiple linear regression analyses indicate further findings regarding the relationship between B-scores and various acoustic measures as illustrated in Figure 7.4. Regardless of what forms of the acoustic measures are included for regression analyses, vowel duration is the only primary acoustic cue for prosodic boundary. That is, no acoustic parameter contributes more to signaling the presence or absence and the locations of prosodic boundary for ordinary listeners than does vowel duration. More specifically, raw vowel duration accounts for the largest variation ($r^2 = 0.23$) in listeners' perception of prosodic boundary, followed by paradigmatically-normalized and syntagmatically-normalized vowel durations. Comparing the results from these two different normalization methods, vowel durations that are paradigmatically-normalized in three different ways generally account for a greater amount of the variation in boundary perception: The total variation (r^2) of perceived boundary with paradigmatically-normalized vowel duration ranges from 0.144 (z -normalization) to 0.154 (z -normalization after log-transformation), while syntagmatically-normalized vowel durations explain 9.9% (3 adjacent stressed vowels) and 12% (5 adjacent vowels) of the variability in ordinary listeners' response to prosodic boundary.

Following are the summarized results of simple and stepwise multiple linear regression analyses. (1) The best regression model of perceived boundary is obtained when employing raw measures of acoustic parameters. (2) Paradigmatically-normalized acoustic measures generally account for larger variation in B-scores than do the syntagmatically normalized ones. (3) As for the contribution of each acoustic measure, vowel duration is the major predictor of perceived boundary, regardless of the vowel

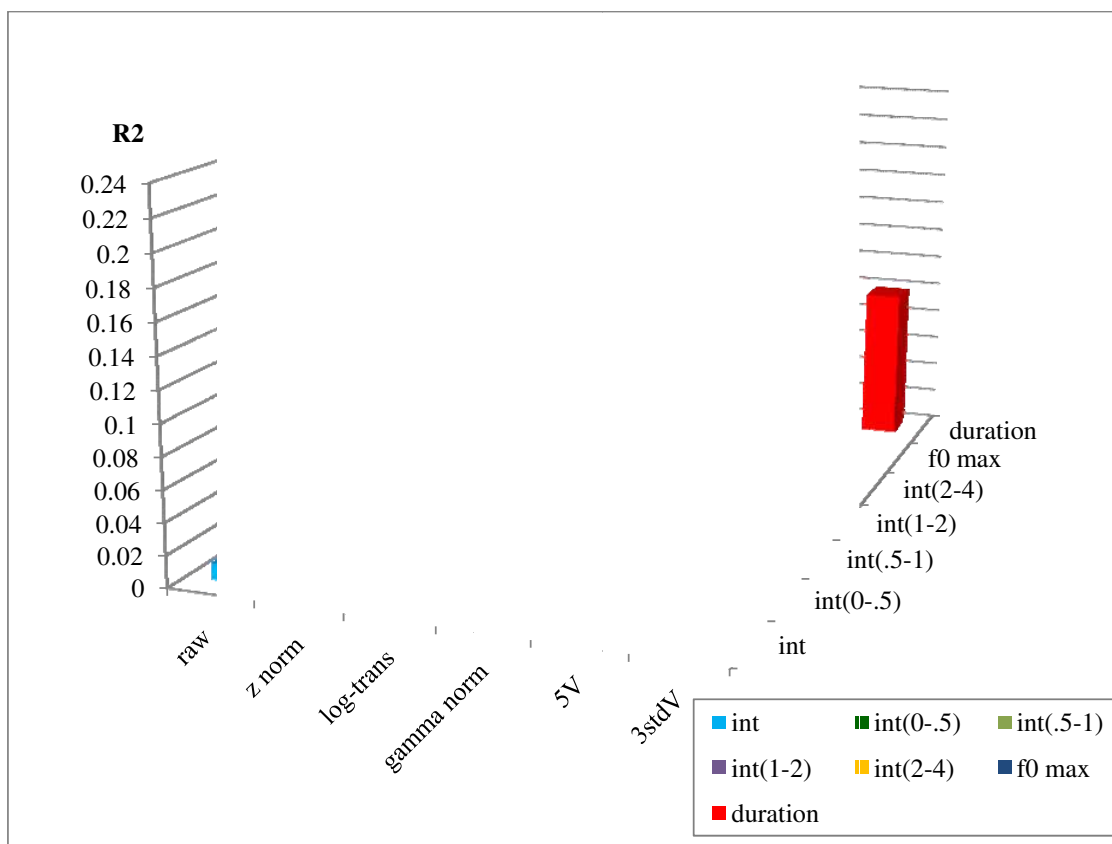


Figure 7.4: Contribution of each acoustic measure (duration, local $F_{0,max}$, overall intensity, subband intensities in 0–500, 500–1000, 1000–2000, and 2000–4000 Hz) to predicting ordinary listeners' response to prosodic boundary: with raw acoustic measures, z -normalized acoustic measures by phone, z -normalized log-transformed acoustic measures by phone, gamma-normalized acoustic measures by phone, syntagmatically normalized acoustic measures over a dynamic window of 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words, in order from left to right.

duration measure employed. (4) On the other hand, other acoustic measures contribute little to signaling prosodic boundary for ordinary listeners.

7.4 Summary and Discussion

The present study evaluates how ordinary listeners identify the presence and the location of prosodic features, relying on changes in the patterns of acoustic parameters. More specifically, it tests whether listeners attend to raw acoustic information or to acoustic information relative to a locally defined context, and, in the latter case, which kind of normalization listeners employ. This study also evaluates whether the perception of prosodic prominence is underlyingly different from that of prosodic boundaries.

The findings from the current study demonstrate that ordinary listeners perceive both prosodic prominence and prosodic boundary by attending to acoustic information. However, although the perception of prosodic prominence is, in some ways, similar to boundary perception, there are important differences. First, ordinary listeners' judgments regarding the presence or absence of both prosodic prominence and boundary of a word rely on the raw duration of the lexically stressed vowels. Comparing vowel duration measures as cues for prosodic features, the largest variation in B-scores is explained when including raw vowel duration as a predictor in multiple linear regression analyses. This suggests that ordinary listeners may not be sensitive to relative speech rate change in the local context nor to phone-specific vowel duration, but rather attend to the absolute duration of lexically stressed vowels. In other words, the relative length of lexically stressed vowels may not be important compared to other instances of the same vowel produced by the same speaker or compared to its neighboring vowels, when ordinary listeners perceive both prosodic prominence and prosodic boundary.

In addition, regression models of both prominence and boundary perception show similar fits to the data, with all the acoustic parameters as predictors (27.4% in perceived prosodic prominence and 29.0% in perceived prosodic boundary), and have similar predictive power of the top 3 acoustic measures (23.4% in prominence perception and 25.0% in boundary perception). The 23.4% prediction rate (based upon overall intensity normalized in a dynamic window of 5 adjacent vowels, raw vowel duration, and local F_0 maximum in a dynamic window of 3 stressed vowels) is noticeably high when compared to the total variation ($r^2 = 12.5\%$) that was accounted for in prominence perception on the basis of a larger number of acoustic parameters including vowel duration, overall intensity, bandpass filtered subband intensities in four different frequency bands, local F_0 maximum, F_0 at the right edge of the vowel, F_1 , and F_2 , as discussed in Chapter 4. These findings suggest that in modeling ordinary listeners' perception of prosodic features in spontaneous conversational speech, it is more important to make comparisons in an appropriate window than to include a large number of acoustic measures. The fact that the role of overall intensity as the strongest cue for prominence appears only when it is syntagmatically normalized with an appropriate window further evidences that the choice of normalization is critical in establishing regression models of prominence perception.

Prominence perception does, however, differ from boundary perception. First, variation in listeners' perception of prosodic prominence is explained by changes in the patterns of a combination of all three kinds of acoustic measures including duration, intensity, and fundamental frequency, while listeners mainly rely on various normalized vowel durations for boundary perception. Other acoustic measures contribute little to guiding ordinary listeners to indicate prosodic boundary. Moreover, the findings from this study suggest that overall intensity accounts for the largest variability of perceived prosodic prominence when normalized in a dynamic window of 5 adjacent vowels, which alone explains 10.4% of the variation in perceived promi-

nence. The contribution of the measures of vowel duration (except raw measure) in the perception of prosodic prominence is relatively small, compared to its contribution to the perception of prosodic boundary. Second, syntagmatically-normalized acoustic measures take into account a larger amount of variation in the regression models of perceived prosodic prominence, while raw measures of acoustic parameters better predict listeners' perception of prosodic boundary.

Another difference arises from the kinds of acoustic measures that are included in the multiple linear regression models. For prosodic prominence perception, the syntagmatically-normalized overall intensity in a dynamic window of 5 adjacent vowels is the best predictor, followed by the raw duration of the lexically stressed vowel, and the local F_0 maximum normalized over a 3 adjacent stressed vowels. The contributions of these three acoustic measures are considerable, with r^2 values of 0.104, 0.094, and 0.036 respectively. In the regression models of perceived boundary, on the other hand, a single acoustic measures, that is, raw vowel duration accounts for most of the variation ($r^2 = 0.23$) in perceived prosodic boundary (79.3%, 0.23/0.29) and the next two top acoustic measures do not contribute much to cueing boundary. These findings suggest that the underlying perception mechanisms of prosodic prominence are different from those of prosodic boundary. When ordinary listeners perceive prosodic boundary, they attend to absolute vowel duration, disregarding local speech rate or phone identity. Yet, the perception of prosodic prominence by ordinary listeners must be characterized by a more complicated model. Instead using one particular form of acoustic parameters, listeners rely on acoustic information from different forms of acoustic measures. That is, in order to best model ordinary listeners' perception of prosodic prominence, changes in overall intensity should be tracked over a relatively large, dynamic window of 5 adjacent vowels, and changes in the local F_0 maximum should be compared with the F_0 maxima of the two neighboring stressed vowels. Listeners do not normalize vowel duration when using it as a cue

for prosodic boundary.

The current study, however, does not exclusively evaluate various methods of normalization. In particular, the sizes of comparison domain attested in the current study are relatively small: 3 adjacent vowels, 5 adjacent vowels, 3 adjacent stressed vowels, and 3 adjacent words. However, it is possible that ordinary listeners employ a much larger dynamic window for normalization. Therefore, in future research, various normalization methods should be compared for modeling of prosodic features.

7.5 Conclusion

Findings from the present study indicate that it is not always necessary to normalize the measures of acoustic parameters in prosody perception. Ordinary listeners normalize the measures of some acoustic parameters, and, even when normalizing them, they employ different normalization methods or domains depending on the kinds of acoustic parameters. It is also shown that the underlying perception mechanisms of prosodic prominence are different from those of boundary perception; in boundary perception, listeners primarily rely on the raw duration of the word-final lexically stressed vowel, while in prominence perception, changes in overall intensity and local F_0 maximum in the local context as well as raw vowel duration are utilized. Findings from the current study further suggest that in the perception of prosodic features in spontaneous conversational speech, ordinary listeners may attend to acoustic information, applying a different normalization method for each acoustic parameter after evaluating all possible methods of comparison rather than employing a single method across different acoustic parameters. Therefore, in order to understand how ordinary listeners identify prosodic structure in everyday speech communication, researchers should explore a variety of acoustic forms instead of uniformly normalized acoustic measures.

Chapter 8

How do Ordinary Speakers Signal Prosodic Features? Speaker-Dependent VS. Speaker-Independent Models

8.1 Introduction

In the previous Chapter, it was shown that ordinary listeners evaluate acoustic variation in appropriate forms of acoustic parameters in appropriate comparison domains when perceiving prosodic features. Turning our attention now to the speaker's point of view, the present chapter asks how ordinary speakers signal prosodic features in spontaneous conversational speech. This study further asks whether there is any uniform acoustic model of prosody across speakers, or whether each speaker employs different acoustic parameters to signal prosodic features. Overall, the central goals of the current study are to gauge the extent of speaker-induced variability in the acoustic implementation of prosodic structure, to identify common or individual acoustic patterns that speakers use to signal prosody in spontaneous conversational speech, to evaluate the contribution of acoustic cues to prosody perception, and to establish optimal statistical models of the acoustic cues to prosody as perceived by ordinary listeners in spontaneous conversational speech.

Most prior studies have attempted to characterize the prosodic structure of an "ideal" speaker as identified by an "ideal" listener, following the generative frame-

work and disregarding variability in the production and the perception of prosodic structure. At the same time, many previous studies have indicated that there are many sources (e.g., speech style, dialect, gender, phonological and phonetic context, and individual variation) which contribute to acoustic variation in the production of prosody (Byrd, 1994; Dilley et al., 1996; Kohler, 1994; Shevchenko and Skopintseva, 2004; Umeda, 1975; Yaeger-Dror, 1996). Among various sources of acoustic variation in prosody production, the current study focuses on acoustic variation induced by individual speakers. In most prior studies, speaker-dependent acoustic variation in prosody production has been observed (e.g., Beckman and Edwards, 1994; Beckman et al., 1992; Cho, 2005, 2006; Cole et al., 2008; Fant et al., 2000a), but few studies have directly examined such variation (Byrd, 1994; Dilley and Shattuck-Hufnagel, 1995; Dilley et al., 1996; Mozziconacci, 1998; Peppe et al., 2000; Redi and Shattuck-Hufnagel, 2001).

Looking into variation in the realization of glottalization in a variety of locations within normal utterances (phrase-medial vs. phrase-final) in American English, Redi and Shattuck-Hufnagel (2001), for example, observed a wide range of variability in the rates of glottalization and in the preferred acoustic parameters associated with boundary-related glottalization. In their study of Southern British English, Peppe and her colleagues (2000) also examined cross-speaker variability of prosody production, in particular, the implementation of intonation using PEPS-Profiling Elements of Prosodic Systems, for which 90 native speakers of a southern variety of British English produced sentences. In the results, they reported qualitative cross-speaker differences in prosodic realization even within a single dialectal speech community, although they did not find any significant quantitative effects of gender and age on prosody production. However, Redi and Shattuck-Hufnagel (2001) only looked at one facet of the acoustic implementation of prosodic boundary, glottalization, and Peppe et al. (2000) showed qualitative differences of prosodic implementation of prosody across speakers.

No prior study, as yet, has systematically investigated speaker-dependent variation in the production of prosodic features across many acoustic parameters.

In the current study, speaker-dependent acoustic variation in prosody production will be investigated, looking at changes in the patterns of various acoustic parameters including vowel duration, intensity, fundamental frequency, and formant structure. Furthermore, this study examines the contribution of each acoustic parameter towards explaining prosody perception, and eventually creates acoustic models of prosody production as indicated by ordinary listeners.

8.2 Acoustic measurements

Acoustic measurements were performed on the lexically stressed vowels from the set of 54 speech excerpts that had been prosodically annotated by groups of ordinary listeners (11–20 transcribers per each speech excerpt) in RPT to measure perceived prominence and the lexically stressed word-final vowels to do the same for boundary. The acoustic measures included the following: vowel duration, intensities (overall and subband in 0–500, 500–1000, 1000–2000, 2000–4000 Hz), local F_0 maximum, F_0 at the right edge of the lexically stressed vowel, formant values (F_1 and F_2), and silent pause. The extracted acoustic measures were then z -normalized by phone within each speaker.

8.3 Results

In this section, the results of speaker-independent and speaker-dependent multiple linear regression analyses of P- and B-scores with the various acoustic measures are summarized.

8.3.1 Prosodic prominence

Figure 8.1 illustrates the distribution of variation in ordinary listeners' perception of prosodic prominence, obtained from simple multiple linear regression analyses of P-scores with acoustic measures by speaker (blue, on the left) and by phone (red, on the right). As can be seen, each speaker's total variation (r^2) of perceived prosodic prominence that can be accounted for on the basis of acoustic measures ranges from 0.120 to 0.536, with the average variation of 0.314, while the total variation in listeners' response to prosodic prominence explained by acoustic measures ranges from 0.149 to 0.398, with the average of 0.220, when reported by phone across all speakers. The results show that speaker-dependent multiple linear regression models generally account for a greater proportion of the variation in prominence perception than do speaker-independent, phone-specific regression models. Yet, greater variability across regression models is observed when looking at multiple regression models of perceived prosodic prominence by speaker (the range of the total variation: 0.416) than when looking at regression models by phone across speakers (the range of the variation: 0.249). Looking closely, the variation in listeners' perception of prosodic prominence in only one speaker-dependent regression model (S01) of perceived prominence is below 0.200, while 5 speaker-independent regression models of prominence perception account for less than 0.200 of the total variability of listeners' prominence perception.

Stepwise multiple linear regression analyses of P-scores with acoustic measures are illustrated in Figure 8.2. The results indicate several interesting findings. First, similar to the findings from speaker-independent acoustic models of prominence perception discussed in Chapter 4, there is no single acoustic measure that is included for/present in regression models of perceived prosodic prominence across speakers. Speakers signal the presence or absence and the locations of prosodic prominence by manipulating the patterns of different acoustic parameters. In addition, in most speaker-dependent multiple regression models of perceived prosodic prominence, vari-

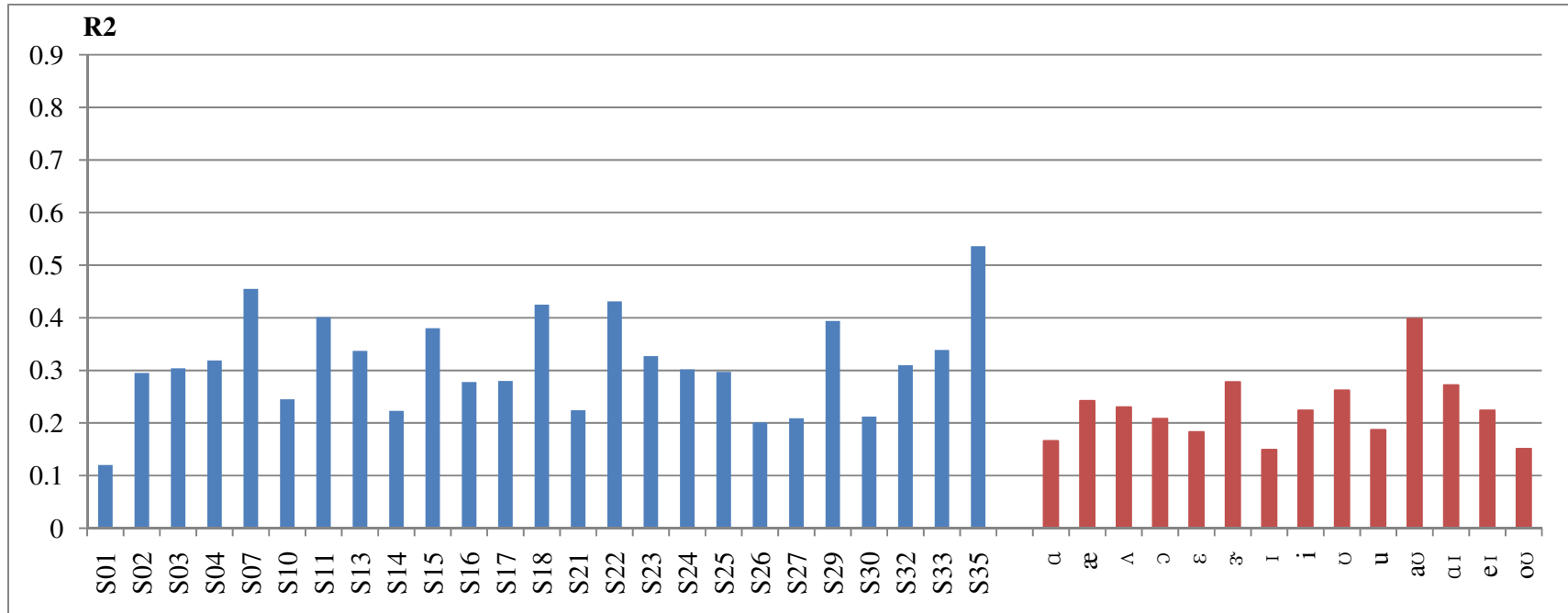


Figure 8.1: The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic prominence by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of P-scores with the acoustic measures

ation in ordinary listeners' response to each speaker's prosodic prominence is accounted for on the basis of acoustic information from a combination of acoustic measures.

Thirdly, and interestingly, the role of intensity measures as cues for prosodic prominence appears to be quite important in many speaker-dependent models of perceived prominence. This has not been found when looking at speaker-independent regression models of perceived prominence, in which vowel duration is considered to be the most reliable cue for prosodic prominence across vowels. 20 out of 25 acoustic regression models of perceived prosodic prominence include intensity measures as one of the predictors, followed by 16 that include temporal measures, 14 that include F_0 measures, and 7 that include formant measures. More interestingly, some speakers' prosodic prominence as identified by a group of ordinary listeners is modeled only through variation in intensity measures: S11, S13, S24, and S27. In other speakers' regression models of perceived prominence, changes in intensity measures are a primary acoustic correlate of perceived prominence, although there are other acoustic measures which aid listeners to perceive prosodic prominence: S22, S23, and S35.

Lastly but most importantly, although ordinary speakers signal prosodic prominence through the implementation of a combination of acoustic measures, speakers vary in which acoustic parameters they primarily manipulate in order to signal prosodic prominence in spontaneous conversational speech. As discussed above, while some speakers (S11, S13, S22, S23, S33, and S35) rely heavily upon intensity parameters in the acoustic implementation of prosodic prominence, others (S07, S10, S14, S15, and S16) mostly utilize fundamental frequency measures to signal prosodic prominence. Only three speakers' regression models (S02, S25, and S29) contain temporal measures as major predictors for prominence, which was shown to be the most reliable acoustic cue for prosodic prominence in the speaker-independent regression models of perceived prosodic prominence. In two regression models (S17 and S18),

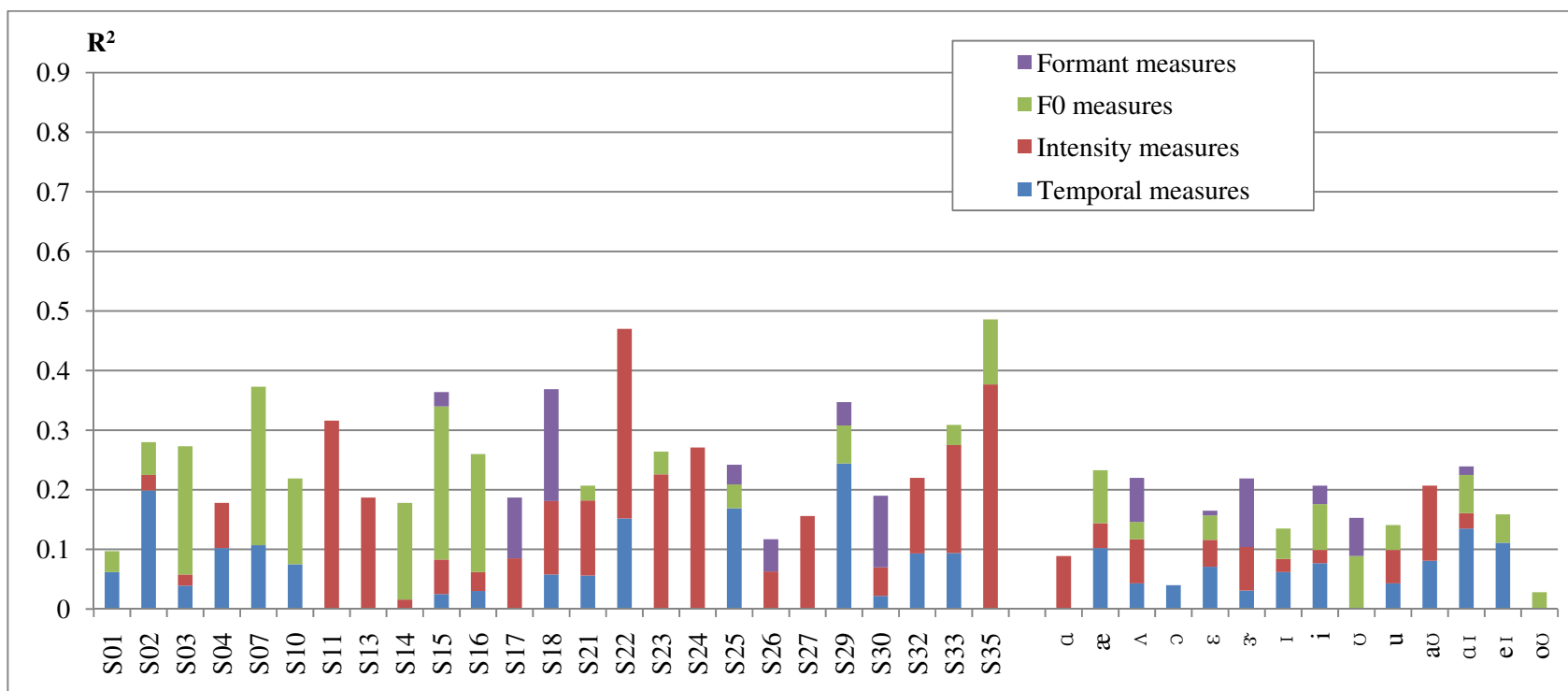


Figure 8.2: The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic prominence by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of P-scores with the acoustic measures

formant measures by themselves take into account more than 50% of speaker-specific variation in listeners' perception of prosodic prominence.

In sum, speaker-dependent acoustic regression models of perceived prominence better account for variation in ordinary listeners' response to prosodic prominence than speaker-independent models. Speakers vary in the set of cues used to encode prosodic prominence: Some speakers rely more on F_0 , duration, and formant measures, while others rely primarily or exclusively on intensity measures. That is, speakers vary in their acoustic encoding of prosodic prominence in terms of both the kinds of acoustic cues and the contribution of each acoustic cue to listeners' perception of prominence.

8.3.2 Prosodic boundary

Figure 8.4 illustrates the distribution of the total variation (r^2) in ordinary listeners' response to prosodic boundary, as explained on the basis of acoustic measures including vowel duration, overall and subband intensities in four frequency bands (0–500, 500–1000, 1000–2000, 2000–4000 Hz), F_0 values (local F_0 maximum and F_0 at the right edge of the vowel), formant values (F_1 and F_2), and silent pauses. In Figure 8.3, the blue bars on the left represent the variation in the speaker-dependent regression models of perceived prosodic boundary, and the red bars on the right represent the same in the speaker-independent, phone-specific regression models of prosodic boundary. Speaker-dependent regression models explain 23.7%–90.1% of the variation for perceived boundary, with an average of 52.2%, while the variations in listeners' response to prosodic prominence that is explained from the acoustic measures in speaker-independent models range from 31.6% to 62.5%, with an average of 45.2%. This generally shows that speaker-dependent regression models are better able to account for listeners' responses to prosodic boundary in spontaneous conversational speech of American English than are speaker-independent regression models. On the

other hand, the range of variation for perceived boundary in speaker-dependent models (variation range: 0.664) is greater than variation range ($r^2 = 0.309$) indicated in speaker-independent models.

Subsequent stepwise multiple linear regression analyses reveal that no matter which regression model (speaker-dependent or speaker-independent) of perceived prosodic boundary is employed, the majority of variation in listeners' perception of prosodic boundary is accounted for based solely upon temporal variation. Only one speaker-dependent stepwise regression model (S22) of perceived prosodic boundary does not include temporal measures as primary cues for prosodic boundary, where changes in F_0 measures exclusively cue prosodic boundary. Yet, in terms of the amount of variation in boundary perception that is/can be explained in regression models, the stepwise regression model of speaker 22 accounts the least for variation ($r^2 = 0.133$) in listeners' perception of prosodic boundary, compared to the other speaker-dependent regression models.

In sum, the results of speaker-dependent multiple linear regression models of B-scores with acoustic measures evidence that speaker-dependent regression models better account for variation in listeners' perception of prosodic boundary. However, at the same time, speaker-dependent models indicate greater variation depending on the individual speaker: some speakers are better at signaling prosodic boundary than others. In some speaker-dependent regression models of perceived boundary, about 90% of ordinary listeners' perception of prosodic boundary can be predicted solely according to acoustic measures-namely, temporal measures. Different from perceived prosodic prominence, prosodic boundary is signaled through the modulation of temporal characteristics of the lexically stressed vowels and the following silent pause.

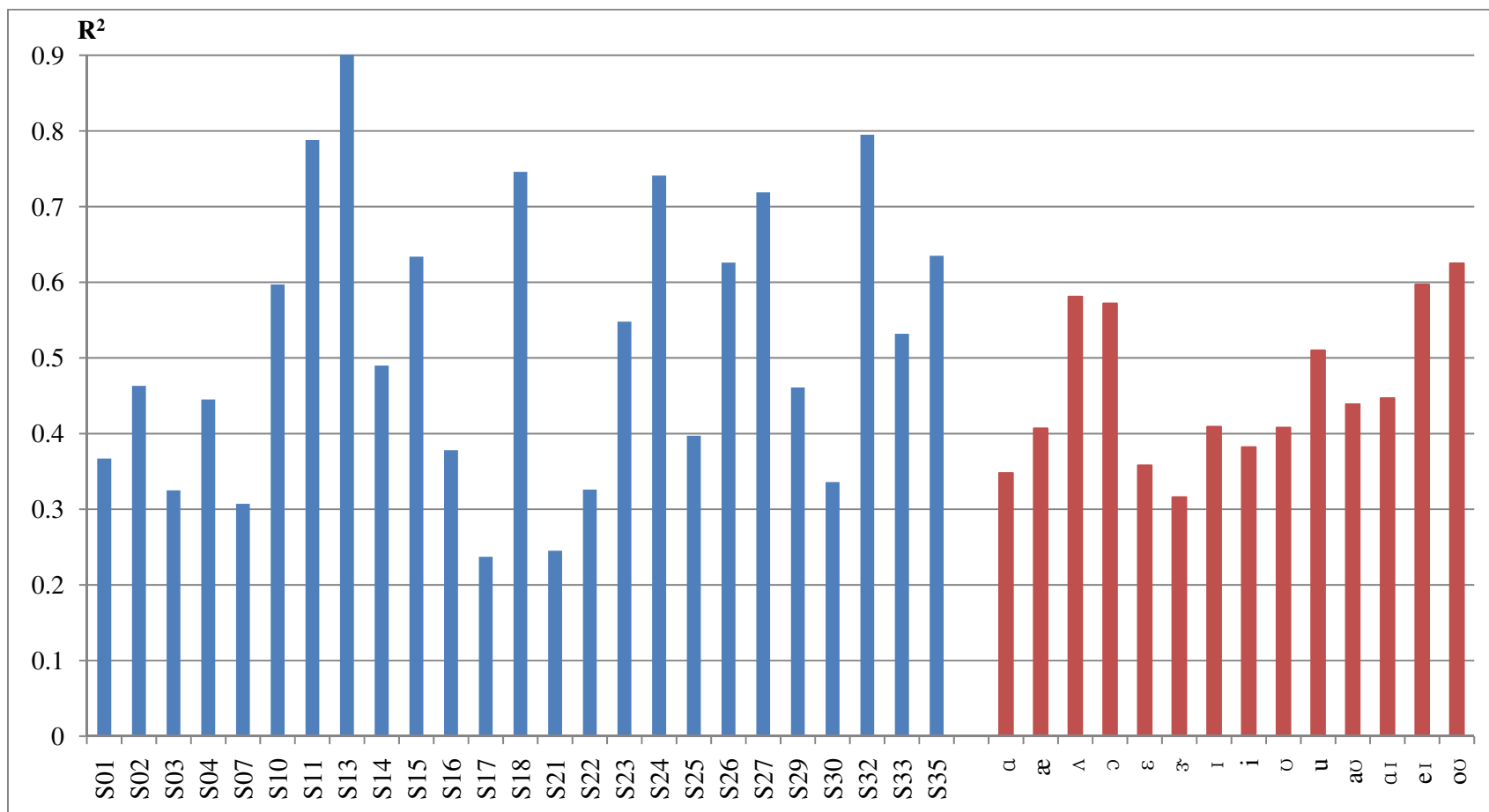


Figure 8.3: The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic boundary by speaker (blue) and by phone (red), obtained from multiple linear regression analyses of B-scores with the acoustic measures

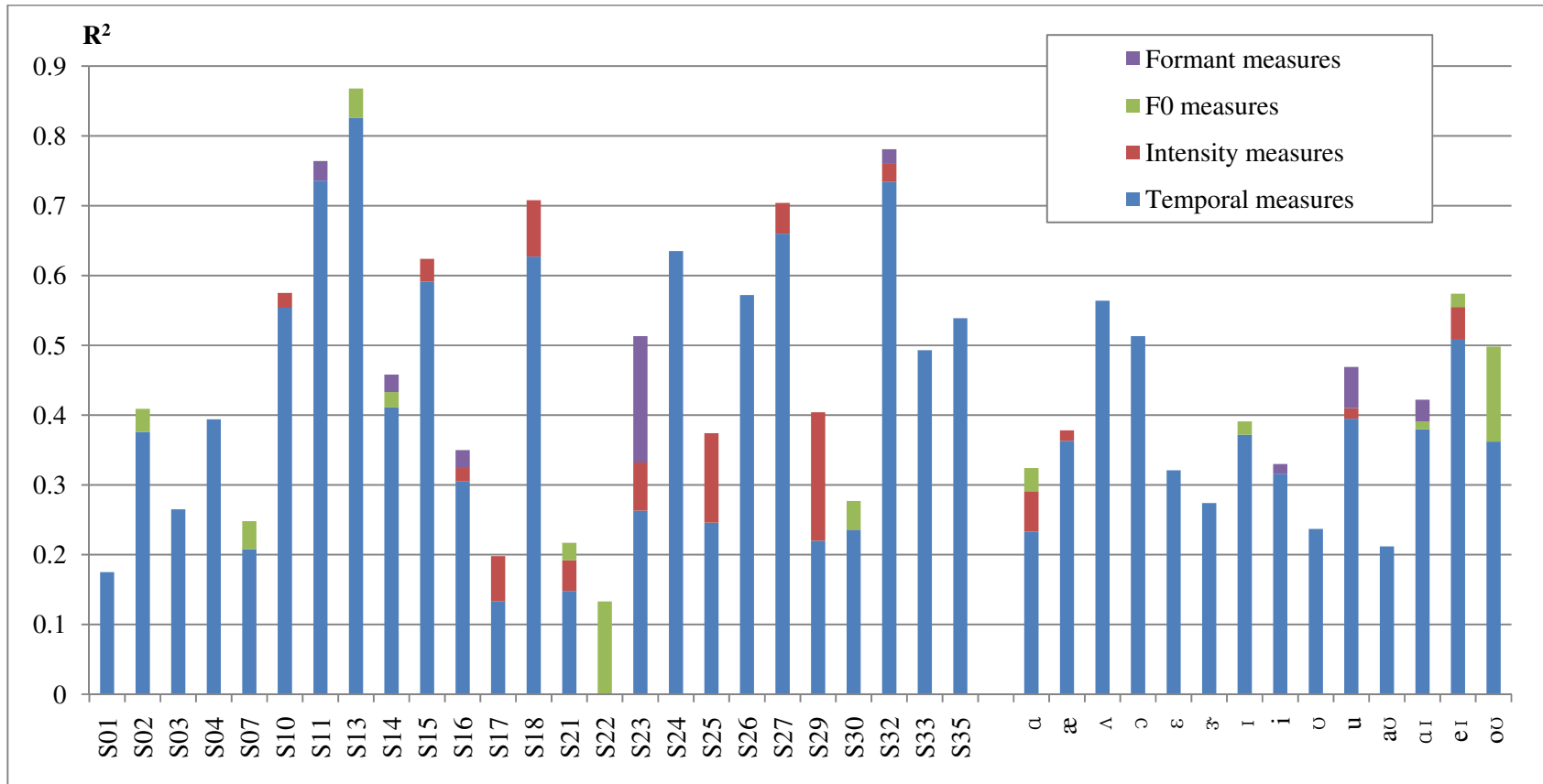


Figure 8.4: The distribution of the total variation (r^2) of the ordinary listeners' perception of prosodic boundary by speaker (on the left) and by phone (on the right), obtained from stepwise multiple linear regression analyses of B-scores with the acoustic measures

8.4 What are the best acoustic models of prosody?

It is shown that ordinary listeners employ different forms of acoustic measures in their free variation in order to decode prosodic structures in the speech signal in Chapter 7 and the previous sections in Chapter 8 demonstrates that speakers modulate a subset of selected acoustic parameters in order to signal prosodic structure in their speech and speaker-dependent variability is great in the acoustic encoding of prosody. If it is true that each speaker select a different subset of acoustic parameters from the set of acoustic parameters that are correlated with prosodic features and different acoustic forms and normalization domains are adopted for prosody perception depending on the characteristics of acoustic parameters, then I should be able to establish the best acoustic models of prosody with the consideration of variance by speakers and by acoustic measures. In the following sections, I report findings from multiple regression analyses of the acoustic encoding prosody and create the best acoustic models of prosody comparing speaker-independent vs. speaker-dependent variability as well as forms of acoustic parameters in various comparison domains.

8.4.1 Best acoustic models of prosodic prominence and boundary

Figure 8.5 illustrates the distribution of the total variation (r^2) in the speaker-independent acoustic models of prosody, indicating that 31.4 to 99.8% of the variation (avg. 52.3%) and 54.1 to 98.4% of the variation (avg. 69.8%) in the acoustic encoding of prosodic phrase boundary in the acoustic encoding of prosodic prominence is explained on the basis of all acoustic measures that are normalized in various domains. On the other hand, Figure 8.6 displays the distribution of the total variation (r^2) in the speaker-dependent acoustic models of prosody. The variation in the acoustic

encoding of prosodic prominence by speakers range from 37.1 to 98.0% (avg. 70.2%), while 54.6 to 100% (avg. 86.1%) of the variation in the acoustic implementation of prosodic phrase boundary is explained on the basis of the same acoustic measures that are used for the establishment of speaker-independent models. These results are consistent with findings from the previous chapters. In light with findings from Chapter 7, the employment of acoustic measures normalized in various ways as independent variables increases the total variation of perceived prominence and boundary. Consistent with findings from the previous sections of Chapter 8, speaker-dependent acoustic models of prosodic prominence account for larger variability in ordinary listeners' perception than speaker-independent models.

The acoustic models of prosody in Figures 8.5 and 8.6 include a very large number of independent variables in the regression models, and the increase of the number of independent variables in regression models inflates the total variation of models. To correct this problem, regression analysis provides adjusted total variation, which takes into account the relations between the degree of freedom and the number of observations, and is considered as a more appropriate goodness-of-fit measure. Therefore, I perform multiple linear regression analyses using backward selection procedures in order to reduce the number of independent variables by eliminating some of independent variables which do not contribute to regression models of prosody and to obtain the largest adjusted total variation in prosody models. The following figures illustrate the distribution of the adjusted total variation in prosody models: Figure 8.7 for speaker-dependent acoustic models of prosody and Figure 8.8 for speaker-independent models of the acoustic encoding of prosody.

A comparison of a set of the total variation in the regression models of prosody illustrated in Figures 8.7 and 8.8 with that in Figures 8.1 and 8.3 reveals that speaker-dependent regression models of prosody with acoustic measures normalized in various comparison domains are far better than both speaker-independent regression models

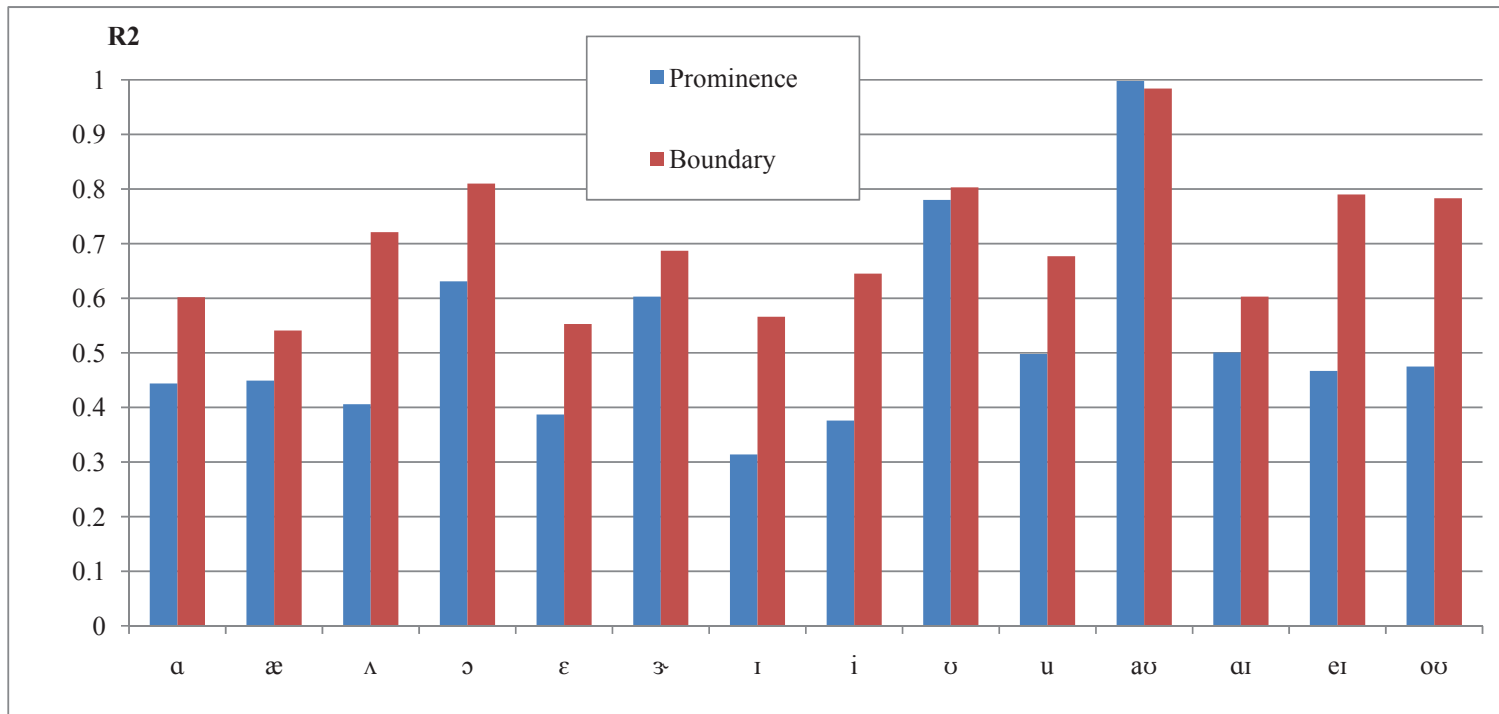


Figure 8.5: The distribution of the total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners

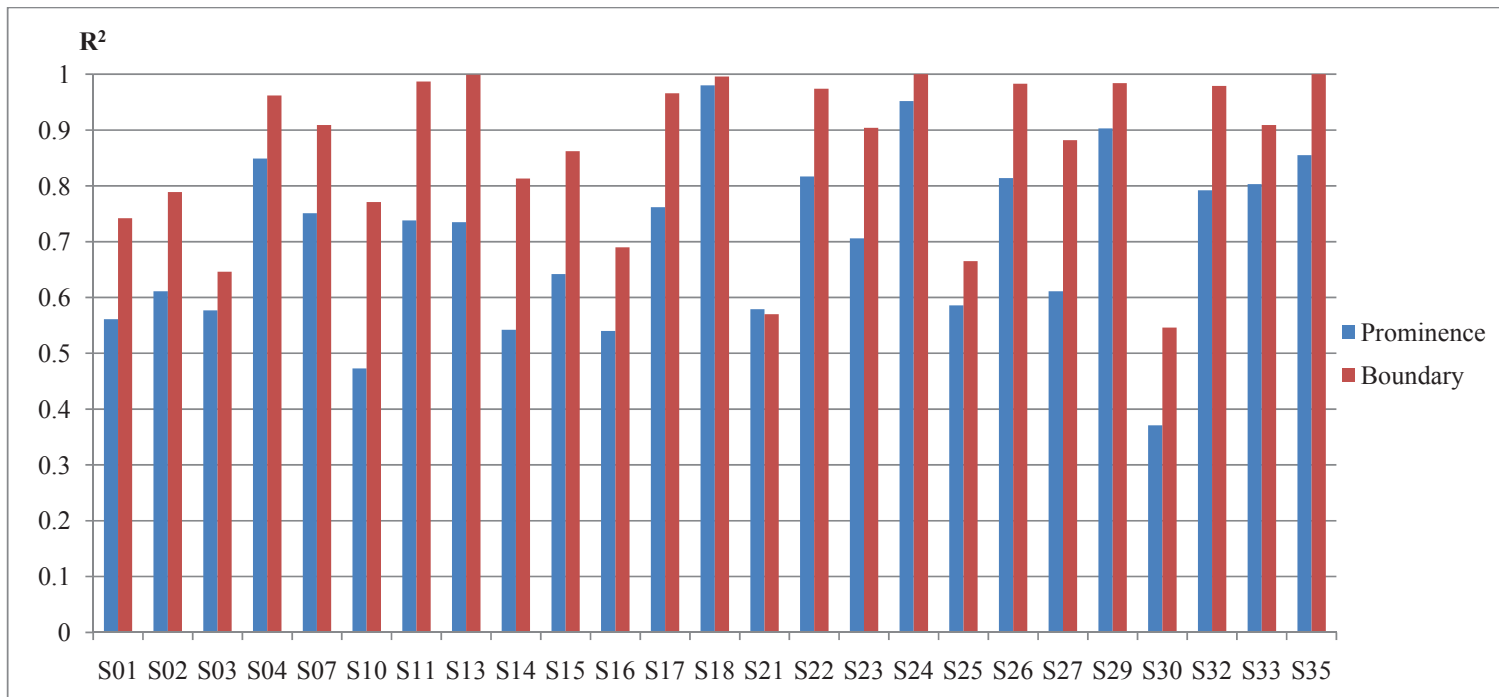


Figure 8.6: The distribution of the total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners

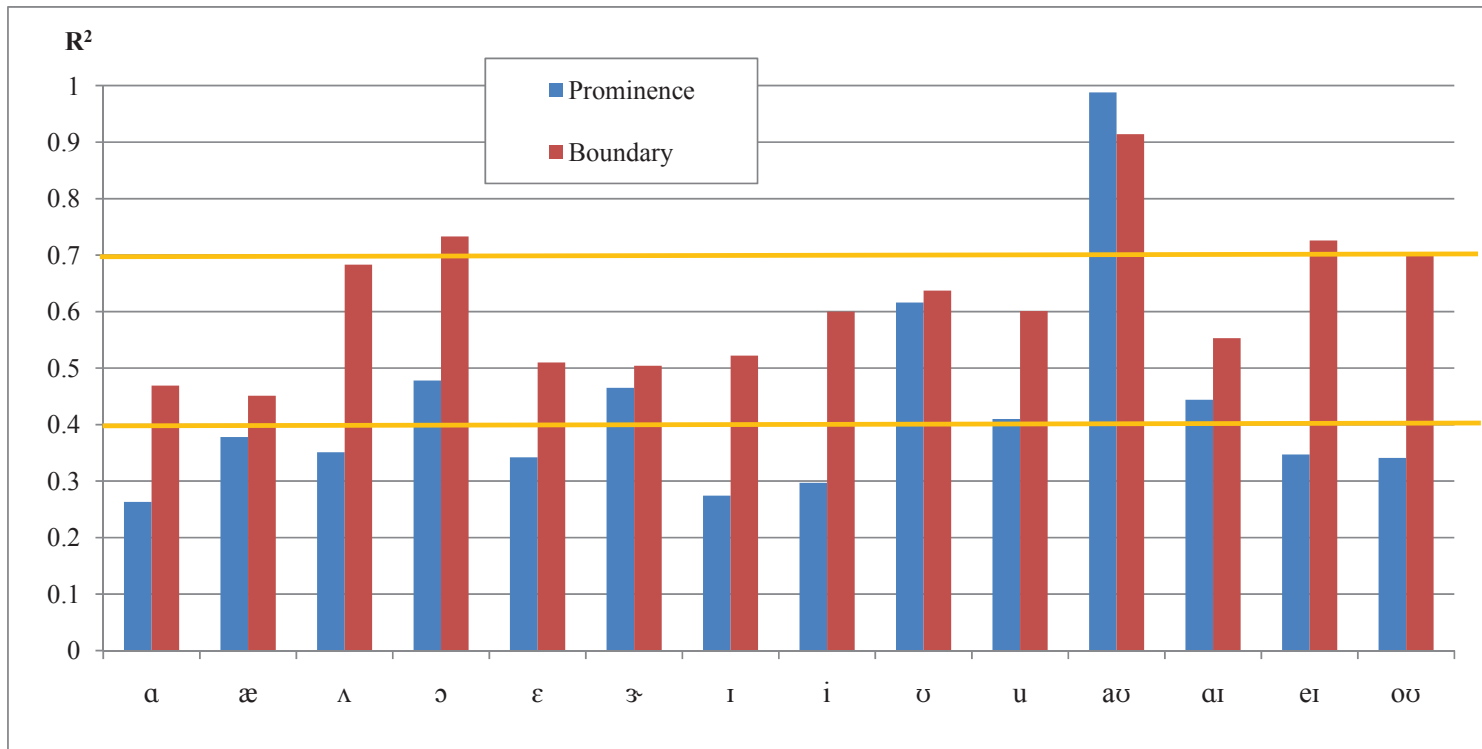


Figure 8.7: The distribution of the adjusted total variation (R^2) in the speaker-independent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners

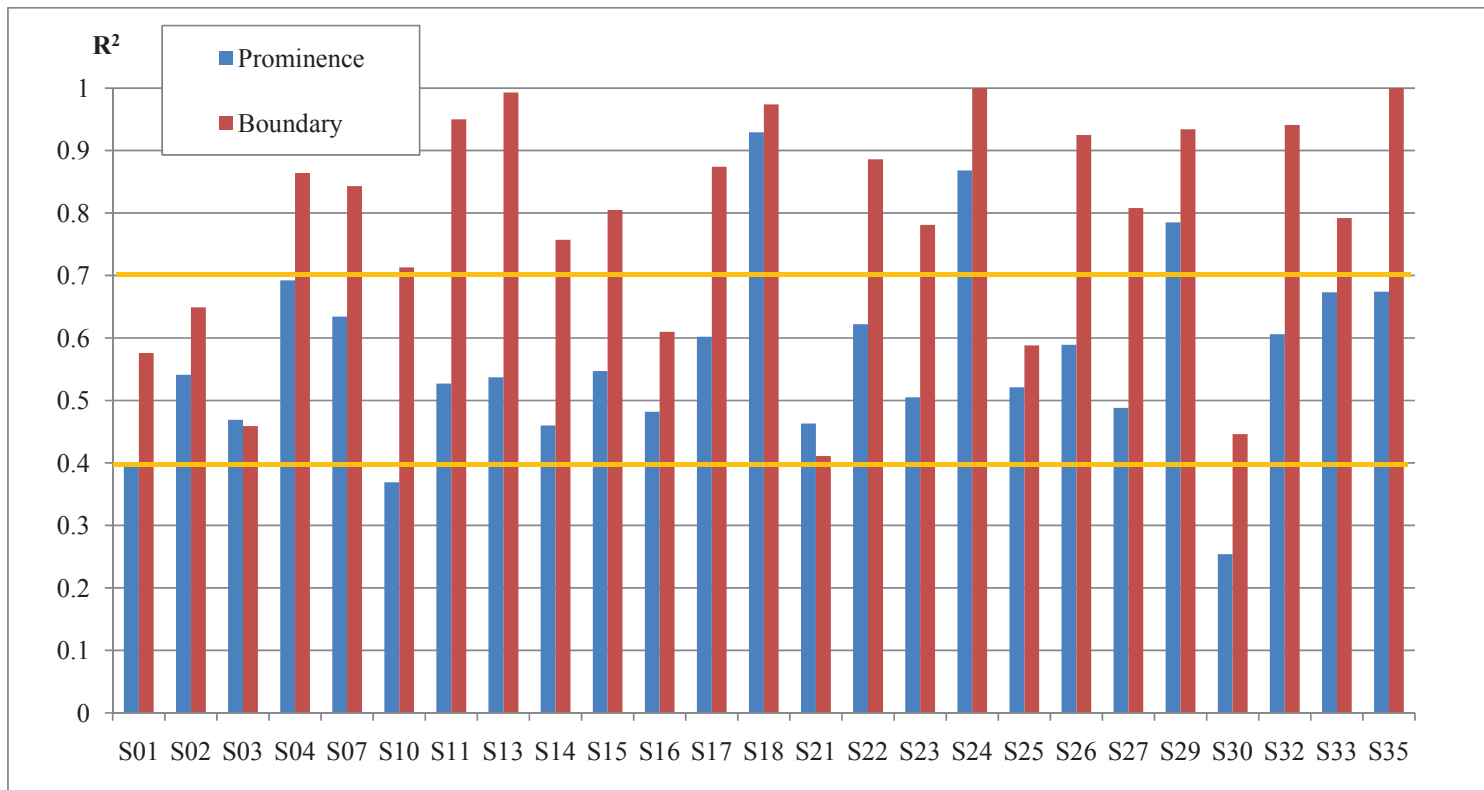


Figure 8.8: The distribution of the adjusted total variation (R^2) in the speaker-dependent acoustic models of prosodic prominence (blue) and prosodic phrase boundary (red) as determined by ordinary listeners

and speaker-dependent regression models with paradigmatically normalized acoustic measures. No speaker-independent regression models and 4 out of 25 speaker-dependent regression models of prosodic prominence account for over 40% of variation in prosodic prominence as determined by a group of ordinary listeners, but 6 out of 14 speaker-independent regression models with acoustic measures normalized in various domains and 22 out of 25 speaker-dependent regression models with acoustic measures normalized in various domains explain over 40% of variation.

The acoustic encoding of prosodic phrase boundary is also better modeled in regression models illustrated in Figures 8.7 and 8.8 than models in Figures 8.1 and 8.3. With paradigmatically normalized acoustic measures, no speaker-independent regression models and 6 out of 25 speaker-dependent models of prosodic boundary account for over 70% of variation in prosodic phrase boundary as determined by ordinary listeners. On the other hand, with acoustic measures normalized within various comparison domains, 4 out of 14 speaker-independent regression models and 18 out of 25 speaker-dependent regression models of prosodic phrase boundary explain over 70% of variation in prosodic boundary as indicated by ordinary listeners. These results demonstrate that the acoustic encoding of prosody is better modeled when employing a selection of acoustic forms normalized in appropriate comparison domains.

Comparing the speaker-dependent regression models of prosody in Figure 8.8 with the speaker-independent models in Figure 8.7, the results show that the speaker-dependent regression models of the acoustic encoding of prosody account much larger variability in prosodic features as determined by ordinary listeners than the speaker-independent models. Looking at acoustic models of prosodic prominence, the comparisons of the speaker-dependent acoustic regression models of prosodic prominence with the speaker-independent models show that 25.4 (S30) to 92.9% (S18) (avg. 56.9%) and 26.3 to 98.8% (avg. 42.8%) of variation in the acoustic encoding of prosodic prominence are explained in the speaker-dependent and the speaker-independent acoustic

models respectively. With a random guideline at the total variation of 40% in yellow on Figures 8.7 and 8.8, it is shown that only 5 out of 14 speaker-independent regression models can take into account over 40% of variation in the acoustic encoding of prosodic prominence but 22 out of 25 speaker-dependent models can. In 15 speaker-dependent models of prosodic prominence, over 50% of variation in prosodic prominence as determined by ordinary listeners is explained based only on acoustic information.

The adjusted total variation in the acoustic encoding of prosodic phrase boundary is also investigated in Figures 8.7 and 8.8. While the speaker-independent acoustic regression models of prosodic phrase boundary accounts for 45.1 to 91.4% (avg. 61.5%) of variation in the acoustic encoding of prosodic phrase boundary, 41.1 to 100% (avg. 78.3%) of variation in boundary implementation is accounted for by the speaker-dependent acoustic regression models. In addition, with a random guideline at the total variation of 70% in yellow on Figures 8.7 and 8.8. I evaluate which regression models of the acoustic encoding of prosodic phrase boundary are better. Similar to the comparisons of speaker-dependent and speaker-independent regression models of prominence encoding, the greater number of the speaker-dependent acoustic models (18 out of 25 regression models) of prosodic phrase boundary take into over 70% of variation in prosodic boundary as indicated by ordinary listeners than the speaker-independent models (3 out of 14 regression models). In addition, four speaker-dependent acoustic regression models of prosodic phrase boundary explain almost 100% of variation in boundary encoding, showing that the locations of prosodic phrase boundaries produced by these three speakers (S13, S18, S24, and S35) can completely be signaled through the modulations of a set of acoustic parameters taken in this study and ordinary listeners can completely recover them on the basis of such acoustic information only.

8.5 Summary and Discussion

The present study examines whether speakers encode prosodic features through the same acoustic implementation, or whether they vary in their acoustic encoding of prosody, as determined by ordinary listeners. In particular, the present study evaluates which acoustic cues each speaker employs when signaling prosody, and how much each acoustic cue contributes to cueing prosody for listeners.

The findings from simple and stepwise multiple linear regression analyses of prosody scores with acoustic measures reveal that each speaker employs different kinds of acoustic parameters to signal prosodic features including prosodic prominence but in boundary production all speakers primarily rely on temporal measures, namely vowel lengthening and silent pause, to signal prosodic boundary. Such differences in the acoustic encoding of prosodic prominence and boundary suggest that the production of prosodic prominence has underlyingly different acoustic mechanisms from boundary production, further confirming the findings from Chapters 4 and 5.

Speakers also vary in the effectiveness of cueing prosodic structure for listeners. That is, some speakers are better at signaling prosodic features through modulations of acoustic patterns, while others are not. For example, in the acoustic regression models of perceived prominence for 5 out of 25 speakers, over 40% of the variation in perceived prominence is taken into account based solely upon variation in the acoustic measures taken in the current study, while only around 10% of variation in prominence perception in S01 is accounted for. As for perceived boundary, in 9 speaker-dependent regression models of perceived boundary, over 60% of speaker-dependent variability in listeners' perception of prosodic boundary is accounted for based on changes in the patterns of acoustic parameters, yet two speaker-independent models explain less than 30% of that same variation. In sum, some speakers are three or four times better at acoustic encoding of prosody than are others. These findings suggest that, given speaker-dependent variability in the acoustic encoding of prosody, rather than only

taking into account the competence of an ideal speaker, more attention should be given to each individual speaker's performance when trying to understand the nature of prosody in everyday communication.

It is not the case, however, that speakers who effectively signal prosodic prominence are also good at acoustically cueing prosodic boundaries. In order to evaluate whether particular speakers are better at signaling prosodic structure through the implementation of acoustic parameters than others, Spearman's non-parametric correlation analyses between each speaker's total variations of P- and B-scores were performed. The results show that speaker-dependent variability in prominence perception is not significantly correlated with that of perceived boundary ($\rho = 0.147$, $p = 0.304$). It is not the case, however, that speakers who effectively signal prosodic prominence are also good at acoustically cueing prosodic boundaries. In order to evaluate whether particular speakers are better at signaling prosodic structure through the implementation of acoustic parameters than others, Spearman's non-parametric correlation analyses between each speaker's total variations of P- and B-scores were performed. The results show that speaker-dependent variability in prominence perception is not significantly correlated with that of perceived boundary.

Another interesting finding stems from the observation that speakers vary in the kind of acoustic parameters used to encode prosodic prominence, and these parameters' contributions as cues for prosodic prominence. The results show that speakers employ a combination of various acoustic measures to signal prosodic prominence in general. However, a closer look reveals that each individual speaker does not manipulate all acoustic parameters when signaling prosodic prominence, but rather primarily relies on a few specific acoustic parameters. In particular, prosodic prominence in four speakers is encoded solely through intensity pattern changes, not aided by any other acoustic parameters. In other words, such speakers make words that carry prosodic prominence louder in their spontaneous conversational speech, and listeners are sensi-

tive to such variation when listening for prosodic prominence. These findings are quite different from those stemming from the speaker-independent models, in which almost all the acoustic parameters are included for modeling perceived prominence, and the contribution of each is only slight. This is because speaker-independent regression models of perceived prominence include acoustic measures from speakers who do not use such acoustic measures as cues for prominence as effectively as other speakers do. Therefore, it is inevitable that speaker-independent regression models explain a smaller amount of the variation in ordinary listeners' response to prominence, but include more acoustic parameters as cues for prosodic prominence.

As in speaker-independent models, prosodic boundary is, on the other hand, signaled through variation in the patterns of temporal parameters in speaker-dependent regression models of perceived boundary. Although boundary encoding is aided by other acoustic cues in some speakers' regression models, a majority of variation in listeners' perception of prosodic boundary is accounted for by temporal information, suggesting that ordinary speakers signal prosodic boundary through the implementation of temporal measures, and ordinary listeners identify the locations of prosodic boundary by relying on the lengthened vowel duration and the presence of a silent pause following a word. However, speakers do not always produce silent pauses to signal a juncture between phrases. In these cases, listeners less reliably identify prosodic boundaries than when speakers produce silent pauses as a cue for prosodic boundary. For example, two speakers (S17 and S22) in this study did not provide any silent pause in their speech excerpts, and listeners were still able to perceive prosodic boundary in such speech. In terms of the effectiveness of signaling prosodic boundary, however, those two speakers' regression models account for the least amount of variation in listeners' perception of prosodic boundary, confirming that the presence of a silent pause following a word is a primary cue to prosodic boundary.

After revealing the benefits of the employment of appropriately normalized forms

of acoustic parameters in the regression models as well as of the speaker-dependent regression models, I establish the best regression models of the acoustic implementation of prosody in this chapter. First, the employment of appropriately normalized acoustic measures in the speaker-independent regression models greatly increases the adjusted total variation in prosody implementation. Secondly, I obtain more of the adjusted total variation explained by establishing the speaker-dependent regression models of prosody encoding instead of the speaker-independent models. In other words, the largest variability in the acoustic encoding of prosody is accounted for by the speaker-dependent regression models with acoustic parameters that are normalized in different ways depending on the characteristics of the acoustic parameters.

8.6 Conclusion

The findings from the first part of this chapter reveal the speaker-dependent variability in signaling prosodic structure in everyday conversational speech, and listeners' attention to speaker-dependent variability in the perception of prosodic features. This study further shows that the regression models of prosody as determined by listeners must be tailored to take into account such speaker-induced variability by selecting a different subset of acoustic parameters depending on speakers. In the second part of the chapter, I propose the best regression models of the acoustic encoding of prosody, with the consideration of the speaker-dependent variability and the variability depending on acoustic parameters. Altogether, the findings from the current study emphasize speakers' active role in signaling prosody with the selection of acoustic parameters in their free variation and listeners' role in responding to the acoustic information as implemented by speakers in spontaneous, conversational speech.

Chapter 9

What Other Factors Affect Ordinary Listeners' Perception of Prosody?

9.1 Introduction

By now, the current study has shown that ordinary speakers encode prosodic structure through acoustic implementation, and that ordinary listeners perceive prosodic features by relying on variation in acoustic parameters in spontaneous conversational speech of American English. First, this study has demonstrated that the acoustic characteristics of lexically stressed vowels are phonetically enhanced under prosodic prominence, and that word-final, lexically-stressed vowels are temporally lengthened with the reduction of other acoustic characteristics before a prosodic boundary. Secondly, it has also been shown that prosodic structure influences the internal temporal structure of monosyllabic CVC words. Yet, this study has also demonstrated speaker-dependent variability in the acoustic encoding of prosody as well as in the parameter-specific normalization domain in listeners' perception of prosody. The focus of the current study has so far been the relationship between prosodic structure and acoustic variation. There are, however, many other factors that influence the acoustic realization of speech utterances, e.g., syntactic structure, discourse structure, word probability, speech rate, etc.

In the present chapter, attention is given to these additional factors that influence acoustic variation, evaluating whether the acoustic variation borne out in the previous

chapters can be attributed to prosodic structure, or if it arises from non-prosodic factors such as syntactic and discourse structure and word probability. In other words, the current study is comprised of two parts. The first section will examine whether ordinary listeners identify the locations of prosodic features, solely responding to their analysis of syntactic structure of given speech utterances, or if any acoustic information aids ordinary listeners in identifying prosodic features in spontaneous conversational speech. The first section will also further examine how much syntactic structure alone can contribute to listeners' perception of prosody, and, if any, how much acoustic variation in the speech signal can additionally contribute to prosody perception. In the second part, the relationship between acoustic variation in the speech signal and the likelihood of lexical items in discourse, and the listeners' prior experience of words in spontaneous conversational speech is investigated. In other words, the main goal of this study is to answer the following research questions: (1) do ordinary listeners perceive prosodic prominence, relying on the likelihood that a word occurs in a certain discourse environment or in a certain language—"American English" (2) do they respond only to acoustic information in the speech signal?, or (3) is there any interaction between acoustic variation and word probability in discourse and in language for the perception of prosodic prominence?

Many prior studies have shown that syntactic structure influences the formation of prosodic structure, but prosodic structure is not always isomorphic to syntactic structure. Prior researchers have proposed theories about the syntax-driven assignment of prosodic structure to speech utterances: the assignment of prosodic phrases and prosodic prominence (Abercrombie, 1964; Liberman and Prince, 1977; Nespor and Vogel, 1983, 1986; Selkirk, 1986 among others). In such theories, researchers claim that the syntactic structure of an utterance is a primary factor in determining its prosodic structure, determining where a prosodic boundary often coincides with a syntactic juncture, and that prosodic prominence is assigned to the most metrically

salient element within a phrase. Yet, many researchers have also observed that speakers vary in the assignment of prosodic structure in their speech, and that the prosodic structure of an utterance is often misaligned with its syntactic structure (e.g., Gee and Grosjean, 1983; Watson and Gibson, 2004). In previous perception studies, the ambiguity in the assignment of syntactic structure and its interpretation has been shown to be resolved with the aid of appropriate prosodic information (Kjelgaard and Speer, 1999; Kraljic and Brennan, 2005 among others).

There are multiple factors than syntactic structure of an utterance that affect the assignment of prosodic structure other. Discourse structure affected by pragmatics and semantics influences the prosodic structure of an utterance (Arnold, 2008; Calhoun, 2006; Cutler et al., 1997; Dahan et al., 2002; Nakatani, 1997). For example, Arnold (2008) demonstrated that listeners tend to perceive words that are acoustically prominent as referring to new entities, while unaccented nouns that are acoustically reduced or not prominent, are perceived as anaphoric referents. As cited in Arnold (2008), in their instruction-giving experiment, Watson et al. (2005) found that prominence rating as well as acoustic measures including intensity, mean pitch, and duration decreases as a function of the accessibility of information of a word, although the accenting of the word was not affected. In Calhoun's doctoral thesis (2006), she also examined how prosodic structure is utilized for the implementation of information structure in discourse. She claimed that as a primary constraint of the assignment of prosodic structure, information structure together with all other factors including syntactic and rhythmical structure determines prosodic structure, including the prosodic phrasing and prominence of a speech utterance. Notably, she found that focal accents are often located at nuclear prominence positions.

In addition, the assignment of prosodic structure is also influenced by other information structures-namely, the predictability of a word in a discourse or in a language. Studies have found that words that are predictable from the surrounding context, or

that frequently occur in a language, are not likely to be prominent (e.g., Aylett and Turk, 2004; Bell et al., 2003). Aylett and Turk (2004) investigated how prosodic structure and language redundancy as a measure of word predictability due to lexical, syntactic, semantic and pragmatic factors are related to the production of words, measured as in duration. They found that word duration is inversely related to predictability of a word, and that prosodic prominence is used to implement such redundancy differences in spontaneous conversational speech. Yet, they also indicated unexplained influences of language redundancy as well as prosodic prominence.

As discussed above, a large body of prior research has indicated that the acoustic realization of a unit in speech utterances is affected by many factors including prosodic, syntactic, semantic, pragmatic, and predictability information, and that such acoustic variation affects the interpretation of speech utterances. Looking at untrained, non-expert ordinary listeners' annotation of prosodic features, the primary objectives of the present study are to see whether listeners' syntactic interpretation of given speech excerpts without the corresponding sounds can predict ordinary listeners' perception of prosody in everyday conversational speech, and, if so, how well they can predict prosodic structure without hearing speech. This study also examines how word predictability in discourse and in the language affects the acoustic realization of a word in isolation from prosodic effects. Overall, this study examines whether ordinary listeners' perception of prosodic features is signaled through the acoustic variation independent of word predictability, or if prosody perception is guided by the predictability of lexical items.

9.2 Does syntactic information fully predict ordinary listeners' perception of prosody?

9.2.1 Data collection

A group of 15 participants, naïve in terms of the phonetics and phonology of prosody transcription, but who have participated in this project for a semester, were recruited from the same undergraduate courses at the University of Illinois at Urbana-Champaign as in Experiments 1 and 2. As a part of a semester long prosody annotation project carried out in a computer laboratory, they participated in a single prosody transcription task. In this prosody annotation task, participants marked the locations of prosodic prominence and boundary for the same 36 speech excerpts used in Experiment 1, as discussed in Chapter 2. However, during prosody annotation, they were not provided with any sound files, and were asked to mark prosodic features on the basis of their expectations, assuming that they were the speakers of the given speech excerpts. As in Experiments 1 and 2, the printed orthographic transcripts of all speech excerpts did not contain any punctuation or capitalization, and were presented in random order. The same 5-minute introduction and simple definitions of prosodic prominence and boundary were provided before the transcription task (see section 2.3.2 in Chapter 2). A group of 15 participants transcribed prosodic prominence first, followed by boundary annotation and the other 15 participants did so in the reverse order.

9.2.2 Results

After having collected prosody annotations, transcriptions were pooled together, and each word was assigned a probabilistic prominence (P-score) and boundary scores (B-score) depending on the number of transcribers who marked that word as prominent

or as followed by a prosodic boundary. Then these transcriptions were compared with those done while listening to sound files, as demonstrated in Figure 9.1 and Figure 9.2. Figure 9.1 displays the distribution of probabilistic prominence scores with and without listening to the relevant sound files for each word in a sample utterance from Speaker 26. In Figure 9.2, the distribution of probabilistic boundary scores with and without sound files in the same sample utterance is illustrated.

These figures show the overview of prosodic structure obtained from listeners' syntactic and semantic expectations, compared with the prosodic structure identified with the aid of acoustic information in the speech signal. When comparing both figures, it seems as though listeners have quite consistent expectations about the locations of prosodic prominence and boundary in a given utterance, and these expectations are well-matched with the actual locations of prosodic features in the speech signal, as identified by other ordinary listeners. For example, some of the long content words such as "personalities" and "independent" are expected to be more prominent than others, and they are indeed perceived as prominent when listeners hear the speech excerpt. Some words in the same excerpt, such as "though" and "independent", are expected to be followed by a prosodic boundary, and listeners indeed perceive a boundary following such words when provided with the audio recording of the utterance. On the other hand, the same figures also demonstrate that there exists some variability in the assignment of prosodic features on the basis of transcribers' expectations contrary to the perception of the established prosodic structure by speakers. For example, the first word, "I", is actually perceived as prominent by over 30% of transcribers when aided by audio files, but is not at all expected to be prominent when they do not hear the audio files. Such variation in the assignment of prosodic features is observed in boundary assignment too. The word "just" is heard as followed by a prosodic boundary by over 70% of transcribers, while none of the participants expect this word to be followed by a boundary. Therefore, it

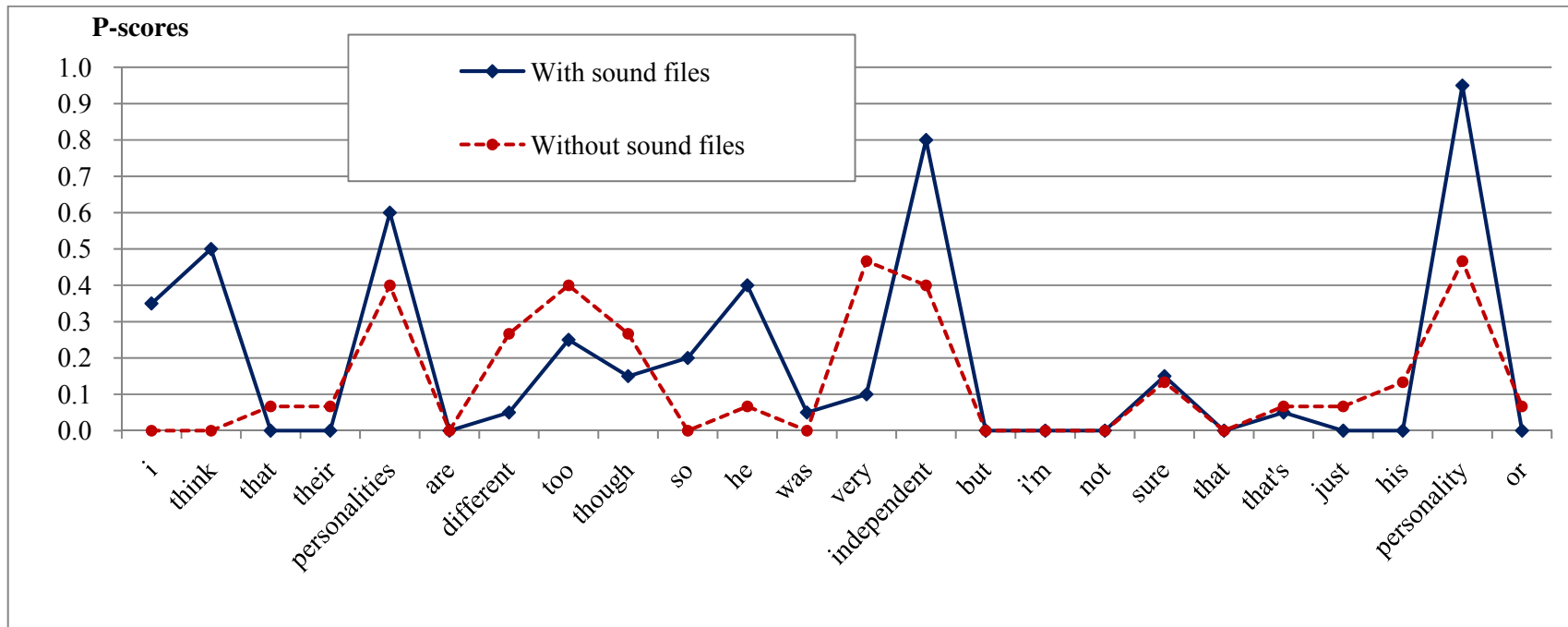


Figure 9.1: The distribution of probabilistic prominence scores (P-scores) of a word in a sample utterance from Speaker 26, with (solid line) and without (dotted line) hearing sound files

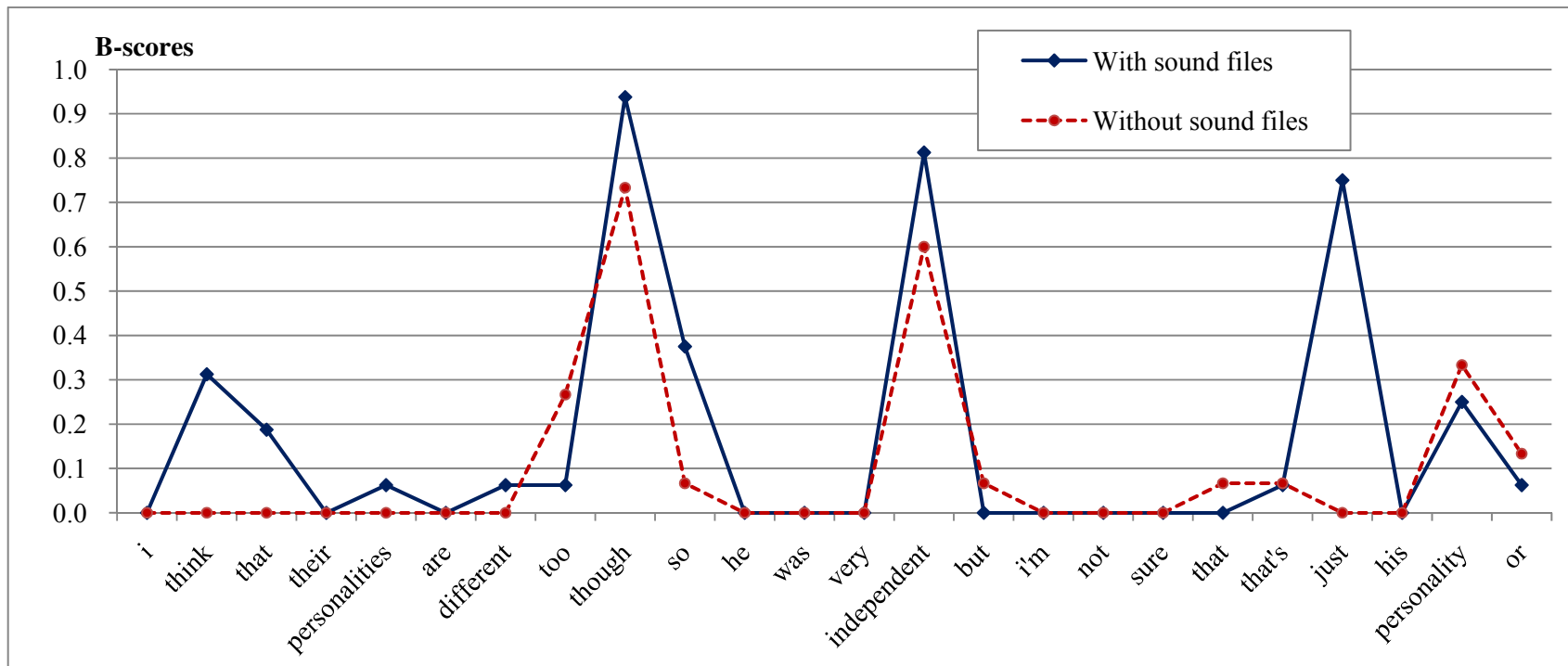


Figure 9.2: The distribution of probabilistic boundary scores (B-scores) of a word in a sample utterance from Speaker 26, with (solid line) and without (dotted line) hearing sound files

is important to statistically evaluate whether prosodic structure assigned to speech excerpts on the basis of transcribers' expectations is different from the actually established prosodic structure, and, if they are related to one another, whether acoustic information contributes to the perception of prosody at all.

This study first evaluated whether ordinary listeners' assignment of prosodic features without audio files is consistent across listeners by using Fleiss' kappa inter-transcribers' agreement scores. Fleiss' kappa coefficients (κ) on the prosody annotation from listeners' expectation are 0.287 ($p < 0.001$) for prosodic prominence and 0.371 ($p < 0.001$) for prosodic boundary. This result tells us that ordinary listeners have fairly consistent and statistically reliable expectations regarding where they would posit prominence and boundary within an utterance. However, these findings do not demonstrate whether prosodic assignment to speech utterance by expectation is the same as prosodic structure produced in spontaneous conversational speech or not.

Therefore, the second test determined whether prosodic structure assigned on the basis of ordinary listeners' expectations is similar to prosodic structure produced in spontaneous conversational speech. That is, this test investigates whether the expected prosodic features for each word is the same as that which is actually assigned. Based on a confusion matrix as shown in Table 9.1, where "No" is assigned to words to which nobody assigned prosodic prominence or boundary, and "Yes" is assigned to words that at least one person assigned either of the two prosodic features, Cohen's kappa agreement scores were calculated by obtaining 0.454 for prosodic prominence and 0.592 for prosodic boundary. This suggests that words for which ordinary listeners expect to assign prosodic prominence are moderately in agreement with those words for which ordinary speakers assigned prosodic prominence, as identified by other ordinary listeners, and that words after which ordinary listeners expect to assign a prosodic boundary are well-matched with those after which ordinary speakers put a

		Prosodic Prominence					Prosodic Boundary		
		No	Yes	Total			No	Yes	Total
Without	With				Without	With			
No		603	342	945	No		1150	222	1372
Yes		195	848	1043	Yes		138	478	616
Total		798	1190	1988	Total		1288	700	1988

Table 9.1: Confusion matrix of prosodic feature assignment based upon listeners’ expectation and prosodic features assigned by speakers in the speech signal: prosodic prominence and prosodic boundary

boundary.

Next, the words for which all ordinary listeners agreed to locate and not to locate a prosodic feature were counted. First, there is only 1 word that all ordinary listeners agree should be prominent, and 942 words to which they agree to assign non-prominence. As for prosodic boundary, only two words are expected to precede a prosodic boundary, and 1501 words in expected to be phrase-medial by all listeners. On the other hand, when sound files were provided, all ordinary listeners agree that there are 10 prominent words, 35 words followed by a boundary, 1012 non-prominent words, and 1148 phrase-medial words. This finding suggests that ordinary transcribers as prospective speakers have different expectations about the locations of prosodic features than others, but, once the utterances in question are produced, prosodic structure as intended by a speaker is more reliably perceived by a listener with the aid of acoustic information.

Lastly, admitting that there is great variation in the assignment of prosodic features in speech utterances by ordinary transcribers as prospective speakers, the mean prominence and boundary scores with and without sound files were compared to determine whether ordinary speakers outperform or underperform ordinary listeners’ expectations about the assignment of prosodic features. With the 95 confidence interval, the mean P-scores when hearing sound files (P-score = 0.151) are significantly

lower than those without aid of auditory input (P-score = 0.178). On the other hand, the mean B-score with the aid of auditory input (B-score = 0.133) is significantly higher than the mean B-score without hearing sound files (B-score = 0.101). That is, the mean P-score is higher but the mean B-score is lower when not hearing sound files, while the reverse is true with sound files.

9.2.3 Discussion

The current study examined whether the assignment of prosodic structure is exclusively determined on the basis of ordinary listeners' interpretations of the syntactic and semantic structure of speech utterances, or whether ordinary listeners' perception of prosody is aided by acoustic information available in the speech signal. The findings from Fleiss' kappa inter-transcriber agreement tests reveal that people generally have the same or similar expectations about the assignment of prosodic features to speech utterances. That is, ordinary transcribers tend to agree about which words within an utterance should be projected as prominent as well as about which words precede prosodic juncture. However, a closer look shows that ordinary transcribers' expectations regarding assigning prosodic boundaries more reliably agree with each other than do expectations of prosodic prominence. These findings suggest that there are more possible ways to assign prosodic prominence than prosodic breaks in speech utterances. There are several ways to interpret this result. It is possible that determining where prominence lands within a phrase is dependent on two separate constraints: information status and speech rhythm. According to Calhoun's information structure-based proposal (2006), only nuclear prominence plays an important role in signaling the informational status of a word, and pre-nuclear prominences are placed according to rhythmical criteria, possibly in combination with criteria related to a word's information-status. Therefore, a word that carries important discourse information, marked as focused or discourse new, is often placed at a position of nuclear

prominence, and prosodic phrasing is structured to signal syntactic and information structure. If so, transcribers, as prospective speakers, would have more freedom to decide which words are assigned prenuclear prominences than where nuclear prominences should be placed within a phrase. As for their choice of speech rhythm, speakers can have different expectations about the locations of prominence within given speech utterances. On the other hand, as for phrasing, my co-authored study (Cole et al., 2010) demonstrated that prosodic phrasing, as indicated by a group of ordinary listeners, is primarily influenced by the syntactic structure of speech utterances. Therefore, prospective speakers may have less freedom in assigning prosodic phrases to speech utterances. This result can also be interpreted as indicating that the assignment of prosodic prominence is constrained by a large number of competing factors including syntactic, information, and discourse structure, and word predictability, but prosodic phrasing is primarily affected by the syntactic structure of speech utterances. In speech production, speakers determine the locations of prosodic prominence in complicated ways, considering many factors. Such complication results in lower Fleiss' kappa inter-transcribers' agreement scores. The relation between prominence perception and word predictability will be examined in section 9.3.

This potential difference in the number of constraints that affect the assignment of prosodic prominence and of prosodic boundary can explain why ordinary transcribers' expectations lead them to mark more prosodic prominence when not hearing sound files than when hearing them. With the printed form of speech excerpts, transcribers mark as many words as they want as prominent, some of which are marked as prominent due to syntactic reasons, others of which are so marked due to information structure, and so on. On the other hand, when the corresponding sound files are provided, a listener's job is to identify the prosodic structure that is intended by the speaker, and, therefore, to differentiate among the many possible prosodic structures that can be assigned to such speech utterances on the basis of acoustic information

in the speech signal. Therefore, the number of prominent words should be a subset of all possible prominent words. Another possible reason why the mean P-score is lower when sound files are provided is time limit. In the prosody annotation tasks with sound files, listeners must mark prosodic prominence while they are listening to the sound files in realtime. While they are marking a word or a phrase as prominent, the next several words may pass by with loose auditory attention on the part of the listener. Due to such time constraints, an ordinary listener may underperform in the auditory perception of prosodic prominence when compared to the expectation-based prominence annotation.

On the other hand, the results show that the mean B-score with sound files is higher than that without. Given fewer possibilities in the assignment of prosodic phrases to speech utterances, and thus fewer words that are possibly followed by a prosodic boundary, the auditory signal reinforces listeners' perception of prosodic boundaries. Another factor which comes into play is the fact that ordinary listeners identify more boundaries when hearing sound files. Disfluency is often associated with the presence of a silent pause. Although speech excerpts were selected which only minimally contain disfluencies, they contain quite a few long silent pauses in the speech signal, such as hesitations and filled pauses. As a primary cue for prosodic boundary, the presence of a long, disfluency-related silent pause may force listeners to mark prosodic boundaries. In the end, ordinary listeners identify more boundaries when hearing the speech signal than otherwise.

As discussed above, listeners' lower agreement on prominence annotation and differences in the mean P-scores between trials with and without sound files may result from the fact that, in the assignment of prosodic prominence, many factors come into play in a complex way. Therefore, the following section will investigate whether word predictability in discourse level (word repetition) as well as in the language (token frequency) is one of constraints in determining the locations of prosodic prominence

that influences ordinary listeners' perception of prosodic prominence.

9.3 How does word predictability relate to the perception of prosodic prominence?

9.3.1 Analysis

The “long” set of prosodically annotated speech excerpts as described in section 2.3 were used for regression analysis between P-scores and word repetition in discourse and word token frequency, P-scores and acoustic measures, and acoustic measures and word repetition in discourse and word token frequency. The vowel duration, overall intensity, and subband intensities in four separate frequency bands were included for acoustic analysis.

9.3.1.1 Word predictability measures as correlates of perceived prominence

In the current study, two measures of word predictability were evaluated: word repetition in discourse and word token frequency in the language. First, the token frequency of a word in the speech excerpts was estimated with the log frequency of the same word in the Switchboard Corpus of spontaneous conversational speech of American English in which much longer phone conversations (over 240 hours of phone conversations from 500 speakers, Godfrey et al., 1992) were recorded than in the Buckeye Corpus (around 40 hours of interviews from 40 speakers). The other measure of word predictability was how many times a word had appeared within one discourse segment. As discussed in sections 2.2 and 2.3, a set of relatively long speech excerpts (31–58 seconds) was prepared for this word repetition measure. That is, the first mention of a word is indexed as 1, the second mention as 2, the third mention as 3, and

the fourth and any subsequent mention as 4. In addition, two additional correlation analyses of P-scores, function words, and frequently reduced words identified from a list of about 80 items by Huddleston and Pullum (2002) (including many pronouns, determiners, auxiliary verbs, prepositions, and conjunctions) were performed, because function words and frequently reduced words have high token frequencies and, thus, tend to be frequently repeated in a discourse.

Table 9.2 summarizes the results of Spearman's non-parametric correlation analyses between P-scores and two word predictability measures (word repetition in discourse and word token frequency) in three different data sets: all words from a set of "long" excerpts, all words excluding frequently reduced words, and all words excluding function words in a set of "long" excerpts. These results show a negative correlation between P-scores and log-frequency of words, suggesting that ordinary listeners tend to perceive a word as prominent when the word does not frequently appear in the language. Subsequent regression analyses of the P-scores and the log-frequency of words reveal that about 18.7% of variation in listeners' perception of prosodic prominence is accounted for on the basis of token frequency information when all the words in the speech excerpts were included for analysis. Comparing the variations in perceived prominence that is explained by information about word token frequency, the total variation decreases when removing frequently reduced words and all function words. This finding suggests that the strong negative correlation of word token frequency with P-scores is attributed to the fact that words that frequently appear in the language are mostly perceived as not prominent. However, even without function words that have high token frequencies and are very likely to be perceived as non-prominent, P-scores are significantly negatively correlated with log-frequency words, confirming that ordinary listeners tend to perceive a word as prominent if that word less frequently appears in the language.

Regarding word repetition effects on prominence perception, Table 9.2 also sum-

Data set	Number of words	Spearman's ρ		R^2	
		Log-frequency	Word-repetition	Log-frequency	Word-repetition
Long excerpts	1725	-0.456 (p < 0.001)	-0.175 (p < 0.001)	0.187	0.025
Long excerpts removing frequently reduced words	1134	-0.379 (p < 0.001)	-0.065 (p = 0.029)	0.124	0.003
Long excerpts removing function words	1040	-0.341 (p < 0.001)	-0.050 (p = 0.108)	0.107	0.002

Table 9.2: Spearman's non-parametric correlation and linear regression analyses of P-scores and log-frequency, and P-scores and word repetition (1, 2, 3, and 4 more), from words in three data sets: all long excerpts, long excerpts minus frequently reduced words, and long excerpts minus function words.

marizes the results from Spearman’s non-parametric correlation and linear regression analyses between P-scores and word repetition. The results demonstrate that word repetition in discourse is only negatively correlated with P-scores when looking at all the words in the “long” excerpt contexts. This can be explained because function words are not likely to show reduction effects as the word repetition index increases in discourse. Although correlation and regression analyses do not indicate interesting findings, the distribution of the mean P-scores suggests that a word mentioned for the first time is more likely to be perceived as prominent than its subsequent mentions as shown in Figure 9.3. The mean P-scores of words that are introduced in a discourse for the first time (P-score = 0.2464) are higher than those of words mentioned for the second time (0.1868) and for those mentioned for the third time as well (0.2097). However, when the word is reintroduced, presumably in another discourse, the likelihood that ordinary listeners perceive this word as prominent is raised: The mean P-score of words that are mentioned four or more times is 0.2506. In other words, ordinary listeners are more likely to perceive a word as prominent when a word appears for the first time in discourse, but they are less likely to perceive a word as prominent when the word is spoken a second or third time. Yet, when the word reenters into discourse, which presumably comprises another discourse segment, as the fourth or subsequent mention, listeners tend to perceive it again as prominent.

In sum, correlation analyses between two measures of word predictability and of perceived prominence show that the perception of prosodic prominence in spontaneous conversational speech by ordinary listeners is correlated with how easily a word can be predicted from a local context (within discourse) or a global context (in the language). Words that are more predictable because they either frequently appear in the language or repetitively occur within discourse, are less often perceived as prominent. In terms of the extent of their effects on prominence perception, word token frequency is more strongly correlated with perceived prominence, and almost

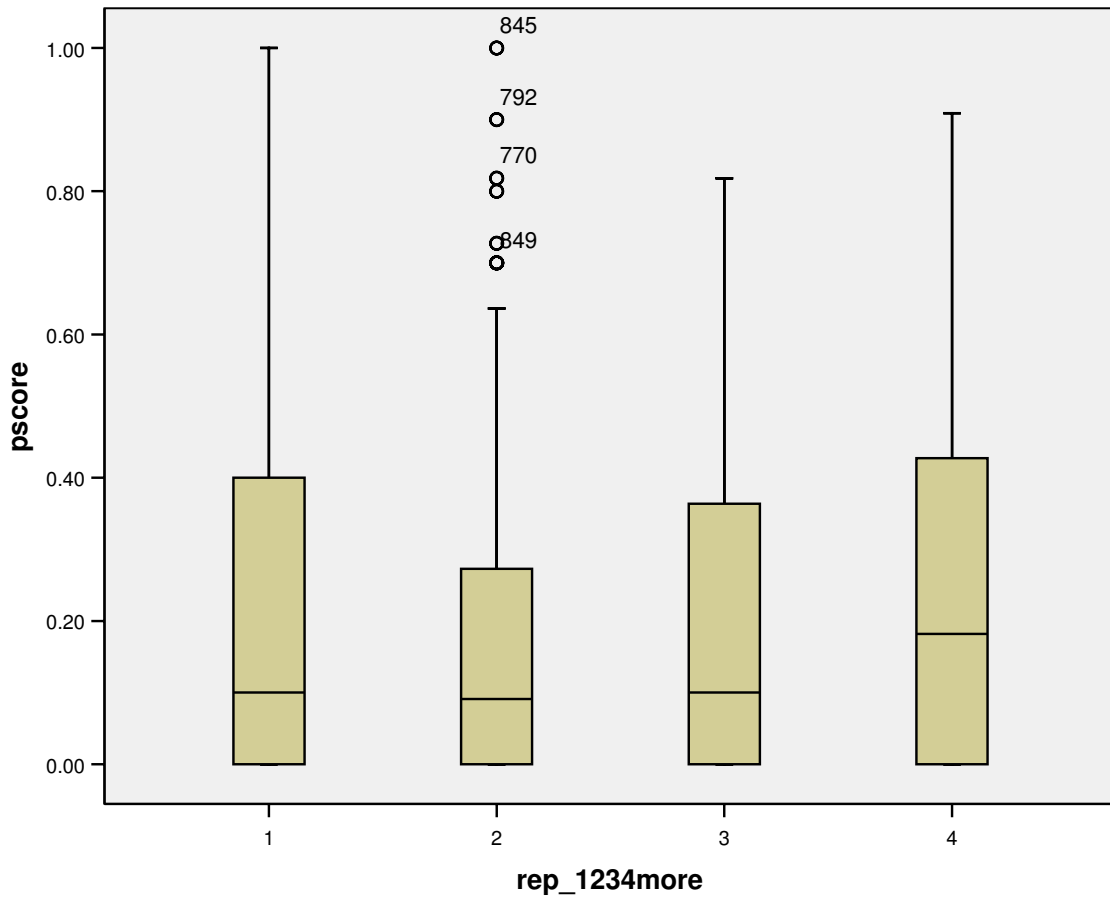


Figure 9.3: Boxplots of P-scores for words in “long” excerpts, grouped by repetition index. These plots include only words that occur with at least two instances within a discourse segment: (1) first mention, (2) second mention, (3) third mention, and (4) fourth or more mention.

20% of listeners' responses to prosodic prominence can be predicted based upon word frequency information, while word predictability within discourse does not predict a great extent of listeners' perception of prosodic prominence. In the following section, the relationship between word predictability and acoustic measures will be investigated in order to see whether strong correlations between perceived prominence and word predictability measures are due to acoustic variation associated with word frequency and repetition, or whether there are any independent effects of word predictability on perceiving prosodic prominence from acoustic effects.

9.3.1.2 Word predictability and acoustic variation as independent cues for prosodic prominence

In this section, the results from simple and stepwise multiple linear regression analyses of perceived prominence with acoustic and word predictability measures are reported as illustrated in Figure 9.4. Around 27% of the total variation (r^2) in ordinary listeners' responses to prosodic prominence is explained from word predictability information as well as acoustic information. More specifically, among all predictors, the contribution of word token frequency is 18.7%, followed by vowel duration (5.8%), and subband intensity in 1000–2000 Hz (2.1%). Other measures' contributions are negligible in modeling listeners' perception of prosodic prominence. This result demonstrates that when listeners judge the presence or absence and the location of prosodic prominence, they employ information from both word predictability and acoustic parameters.

9.3.2 Discussion

As for factors other than syntactic structure influencing the assignment of prosodic features on speech utterances, the current study examines whether listeners' perception of prosodic prominence is influenced by word predictability, and, if so, whether

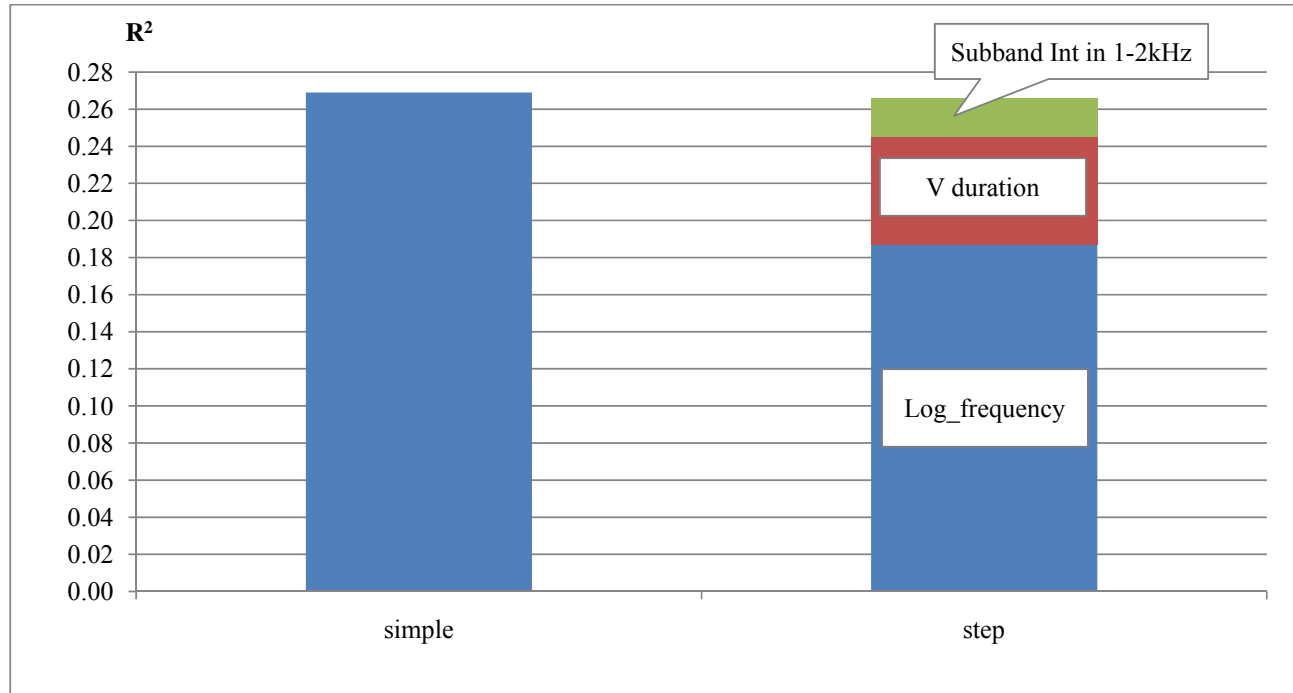


Figure 9.4: The total variation (r^2) in ordinary listeners' perception of prosodic prominence from multiple linear regression analyses on the left and each major correlate's contribution to modeling listeners' perception of prosodic prominence on the right

the effects of word predictability is independent from the effects of acoustic variation on prominence perception. Findings from Spearman’s non-parametric correlation and multiple linear regression analyses indicate that when words are predictable because they have higher token frequency in the language or they are repeatedly mentioned within a small discourse, they tend to be less prominent than words that are neither globally nor locally predictable, as indicated by ordinary listeners. These findings are consistent with those from prior studies in which a word is more likely to be prominent when introduced to the discourse for the first time, and is perceived as more prominent or as referring to a new entity (e.g., Arnold, 2008; Watson et al., 2005), as well as in studies in which more frequent words have more variation in pronunciation, and are often times reduced and not fully pronounced in spontaneous speech (e.g., Bell et al., 2003; Jurafsky, 2002).

This study further evaluates the independence of word predictability and acoustic information as cues to prosodic prominence in spontaneous conversational speech. If these two kinds of measures are not independent of one another, then the results should show that the variation in perceived prominence explained by word predictability is fully accounted for on the basis of acoustic variation, or that the variation in perceived prominence that is explained by acoustic variation is a subset of the variation that word predictability can account for. However, it was found that the two measures related to a word’s predictability and acoustic measures combine to contribute to listeners’ perception of prosodic prominence, and that neither word predictability nor acoustic parameters completely explain the variation in listeners’ perception of prominence. This suggests that the prosodic structure of a speech utterance is not fully mediated through acoustic encoding, or at least, from a listener’s perspective, acoustic information is not the only source of information about the prosodic structure of a given utterance, but listeners must rather take into account many other factors such as word predictability in their perception of prosodic features.

9.4 Conclusion

The current chapter is comprised of two different studies. In the first part, the role of listeners' expectation regarding prosodic structure in the perception of prosody was investigated, while the second part investigated the effects of word predictability on prosody perception. The main findings from this study are (1) listeners have reliably similar expectations about the assignment of prosodic features to a given speech utterance, (2) the expected assigned prosodic structure is similar to their perceived prosodic structure, and (3) information about word predictability influences listeners' perception of prosody independently from acoustic information. However, these factors are not independent of one another, nor do they perfectly covary with one another. Many factors come into play when signaling prosodic structure in speech, and a listener must attend to various factors to recover the prosodic structure intended by the speaker.

Chapter 10

Conclusions

This concludes the examination of the acoustic nature of prosody production in spontaneous conversational speech of American English, and its perception by untrained, non-expert ordinary listeners. The following sections summarize the main findings in this work and propose statistical models of prosody production and perception from the listeners' and speakers' point of view. The implications and contribution of this work to other fields in linguistics, psycholinguistics as well as to speech sciences are discussed. This section will be wrapped up by suggesting future research directions.

10.1 Summary of findings

In this work, the nature of production and perception of prosody has been investigated with spontaneous conversational speech of American English, focusing on acoustic variation arising from prosodic context and its interaction with other factors such as word predictability and listener's syntactic and semantic expectations. This dissertation demonstrates that ordinary speakers implement prosody through the modulation of acoustic information in the speech signal of spontaneous conversational speech, and ordinary listeners recover prosodic structures as intended by speakers, relying on acoustic information with the aid of various other factors including word predictability as well as their syntactic and semantic expectations on prosodic structures. On the basis of these findings, my dissertation proposes the best statistical models of the acoustic encoding of prosodic structures as determined by

ordinary listeners.

Chapter 2 introduces a new prosody annotation method (Rapid Prosody Transcription, RPT), which is economical in terms of time and monetary commitment and is also a good approximation of prosody perception in everyday communication. In RPT, a large group of untrained, non-expert ordinary listeners marked the locations of prosodic features, either prominence or boundary, while listening in real-time to speech excerpts. Prosody annotation in this task is solely based upon listeners' auditory impressions, not aided by any visual display of speech. Each word in the speech excerpts is assigned a probabilistic prosody score, which represents the relative strength of the corresponding prosodic feature. The reliability of ordinary listeners' prosody annotations has been evaluated, showing that ordinary listeners' annotations of prosodic features in spontaneous conversational speech are not only reliable and consistent across listeners, but are also comparable to those of trained, expert transcribers. However, this study also indicates variation in agreement scores as well as in the mean intervals between prosodic features by speaker and by listener, allowing the examination of variability in prosody production and perception.

Chapter 3 summarizes findings from the investigation of the relationship between prosodic scores and vowel identity. Showing that the distribution of prosodic scores is not uniform by vowel identity, I have looked at acoustic variation arising from prosodic context in Chapters 4 and 5. In Chapter 4, findings from the examination of the acoustic encoding of prosodic prominence are reported. First, the study demonstrates that the acoustic characteristics of lexically stressed vowels are enhanced under prosodic prominence in all acoustic dimensions. The target vowels are temporally lengthened and louder, and have an increased local F_0 maximum. Moreover, they have a distinct formant structure in frequency, reflecting more open and peripheral articulation, and in amplitude, reflecting energy concentration. However, stepwise multiple linear regression analyses have also shown that there is no single dominant

acoustic cue to prosodic prominence as indicated by ordinary listeners, but prosodic prominence is signaled through a combination of various acoustic cues, suggesting that ordinary listeners attend to any available acoustic cues for prosodic prominence in everyday communication rather than they rely on one particular acoustic cue.

The findings from the examination of acoustic variation associated with prosodic boundary are presented in Chapter 5 and are compared with those in relation to prosodic prominence. Similar to prosodic prominence, prosodic boundary influences the acoustic characteristics of speech elements-in particular the word-final lexically-stressed vowels in the current study. These word-final lexically-stressed vowels in the preboundary condition are lengthened and are often followed by a silent pause longer than 20 ms. However, they are not acoustically enhanced, but rather are reduced in some acoustic dimensions, particularly spectral properties. That is, they have reduced overall and subband intensity in low frequency bands. Subsequent stepwise multiple linear regression models of prosodic boundary are proposed, revealing that prosodic boundary, as identified by ordinary listeners, is signaled primarily based upon the information related to speech rate, e.g., vowel length variation and the presence/absence of a silent pause following a word.

The findings in Chapters 4 and 5 suggest that the underlying mechanisms of the acoustic encoding of prosodic prominence are different from those of prosodic boundary. In prominence production, ordinary speakers determine the assignment of prosodic prominence while they speak and select any subset of acoustic parameters from a set of all acoustic parameters that are correlated with prosodic prominence. Then the determined prosodic prominences are signaled through the active modulation of the selected subset of acoustic parameters, making the elements of speech acoustically enhanced. As for boundary production, on the other hand, speakers slow their speech in the vicinity of a prosodic boundary, and this slower speech tempo is reflected in the lengthened duration and the lower spectral energy, as well as a

following silent pause.

Taking into account the fact that both prosodic prominence and boundary elongate the duration of lexically stressed vowels, Chapter 6 examines whether the lengthening effects of prosodic prominence are identical to boundary effects, looking at the effects of prosodic prominence and phrase boundary on the internal temporal structure of monosyllabic CVC words. Revealing that regardless of the position within a word, all subsyllabic components (onset, nucleus, and coda) of monosyllabic words are lengthened before a prosodic prominence as well as before a prosodic boundary, this work suggests that the lengthening effects of prosodic features override the intrinsic length difference associated with phone identity. Yet, the relative strength of lengthening effect due to prosodic features varies as a function of prosodic feature and the position within a word. Under prosodic prominence, a nucleus's duration lengthens the most, followed by the duration of onsets and codas, while before a prosodic phrase boundary, the lengthening of codas is the second largest, although a nucleus duration lengthens the most similar to prominence effects. In other words, under both prosodic prominence and boundary, the metrically most salient and important lexically-stressed vowels are lengthened the most. On the basis of these findings, I propose that the temporal effects of prosodic prominence are not identical to boundary effects when looking at the internal temporal structure of the monosyllabic CVC words. More specifically, I further propose that under prosodic prominence, a nucleus takes over a coda proportion, making the CV in a prominent word more distinct from that in other, non-prominent words, while before a prosodic boundary, the VC duration increases more, reflecting the fact that lengthening associated with boundary is more likely due to the slow-down of the articulators, and that such slow-down may be concentrated at the lexically stressed vowels. It is further shown that a considerable amount of ordinary listeners' perception of prosodic features can be explained based solely upon changes in the raw durations of a word, suggesting that ordinary listeners

may not rely on changes in the relative durations of words in any comparison domains but rather attend to variation in raw durations.

The subsequent chapter further examines whether ordinary listeners attend to acoustic variation in any comparison domains, and, if so, whether all the acoustic parameters are compared in the same domain. The findings from multiple linear regression models of prosody scores with acoustic measures normalized in various domains have shown that listeners' responses to prosodic prominence are better explained when employing appropriately selected normalization methods for each acoustic parameter. That is to say, in prominence perception, ordinary listeners utilize the duration information in raw measures, the intensity information compared with intensity information of the adjacent speech elements, and changes of F_0 maximum in the local context. This suggests that ordinary listeners employ, if necessary, an optimal normalization method for each acoustic parameter. Yet, listeners' boundary perception is not sensitive to the way in which acoustic measures are normalized, implying that the absolute duration of a lexically-stressed vowel as well as the presence of a following silent pause are primary cues to prosodic boundary, and that these measures are insensitive to normalization.

In the first part of Chapter 8 speaker-dependent and speaker-independent models of the acoustic encoding of prosody are evaluated and in the second part the best statistical models of prosody implementation are proposed. Indicating the speaker-dependent variability, it is shown that the speaker-dependent models can generally better account for the variation in the acoustic encoding of prosody as determined by ordinary listeners than the speaker-independent models. It is further shown that when looking at regression models of prosody by speaker, speakers vary in the effectiveness of the acoustic implementation of prosodic features in the speech signal, as well as in the selection of acoustic parameters that are modulated to signal prosodic features.

On the basis of the finding that some speakers are better at signaling prosodic

features than others, I propose that prosody must be modeled by individual speaker with the careful considerations on each speaker's selection of normalization methods depending on characteristics of acoustic parameters. The best statistical models of the acoustic implementation of prosody account for almost 60% of the variation in prosodic prominence and 80% in prosodic phrase boundary in average.

Lastly, the present study examines factors other than acoustic parameters that interrelate with acoustic variation and influence the production and the perception of prosodic features in spontaneous conversational speech. In particular, Chapter 9 discusses the relationship between listeners' expectations regarding the assignment of prosodic features and speakers' production of prosodic structures and between listeners' expectations and listeners' actual perception of prosodic structures. In addition, the effects of word predictability within a discourse and in the language on listeners' perception of prosody have also been investigated. According to the findings from this study, listeners greatly agree on where they would locate prosodic prominences and boundaries in the given speech excerpts, and their expectations do not differ greatly from where listeners actually marked prosodic prominences and boundaries. This suggests that ordinary listeners as prospective speakers do have similar expectations about prosodic structure, but that their perception of prosodic features is also directly shaped by acoustic information.

Word predictability has also been shown to influence listeners' perception of prosodic features. If a word is more predictable within a discourse or the language, ordinary listeners are prone to hear such words as non-prominent. Although predictability information is encoded as acoustic variation in the speech signal, step-wise regression models have demonstrated that word predictability information is not encoded solely through the modulation of acoustic parameters, but that word predictability, in particular, word token frequency contributes to listeners' perception of prosodic prominence. These findings confirm that multiple factors relating to listen-

ers' expectations and word predictability in addition to acoustic information interplay with one another in prosody perception.

10.2 Conclusion

Through a wide range of evidence, this research has shown that (1) prosodic structure is signaled through acoustic encoding, and, more specifically through speaker-specific acoustic encoding, and (2) such acoustic encoding of prosody is perceived by listeners in the optimal comparison domains. Furthermore, allowing the variability in speakers' acoustic implementation of prosody as well as in listeners' normalization of acoustic information, this research proposes statistical models of prosody in spontaneous, conversational speech of American English. This research has further demonstrated that (3) in addition to acoustic variation, other factors including word predictability in discourse as well as in the language, and expectations from syntactic and semantic structure of speech utterances interplay with one another to cue prosodic structure. These findings have implications for investigating prosody-related phenomena in other disciplines of speech sciences including psycholinguistics, speech pathology, and speech technology, as well as linguistics. For example, acoustic variation obtained by treating acoustic parameters appropriately in an optimal domain can aid the development of automatic speech recognizers (ASR) with human-like performance. In developing ASRs, speaker-specific acoustic variation in relation to prosody should be taken into account, and ASR systems should be sensitive to such speaker-specific acoustic variation. In speech pathology, the production and the perception of prosody by speakers with speech disorders must be understood in relation to a wide range of acoustic variation in normal speakers' prosody implementation.

My dissertation emphasizes the importance of variability in producing and perceiving prosody in everyday conversational speech, which is generally disregarded

in research following the generative framework. Revealing that there is no uniform prosodic structure determined by linguistic rule, this work shows that variability resides in both speakers' and listeners' sides. The variability is attributed to the assignment of prosody, the selection of a set of acoustic parameters, and their acoustic implementation in the speakers' side and the selection of acoustic parameters and normalization windows in the listeners' side. My dissertation also introduces a new prosody annotation method, Rapid Prosody Transcription (RPT) as a way to approximate ordinary listeners' perception of prosodic features as produced by ordinary speakers in everyday speech communication. This method is not only a reliable and consistent, but is also an economical method for obtaining prosody annotation. This new method can be employed in prosody production and perception research. For example, the relative easiness of transcription tasks allows researchers to obtain prosody annotation from young children with low cognitive ability or from adults with various speech impairments.

In future research, with a larger prosodically annotated corpus (about 5.6 hrs of speech), the examination of the nature of prosodic boundary in spontaneous conversational speech will be extended in various ways. First, the production and perception of prosody will be further examined, eliminating possible prominence effects from boundary effects, and the nature of prosodic prominence, removing possible boundary effects, which was not taken into consideration for most of the current study. Secondly, the nature of nuclear and prenuclear prominences will be examined. In these preliminary results, it has been shown that ordinary listeners more greatly agree on nuclear prominences than prenuclear prominences, suggesting that nuclear prominence may be at a separate status from prenuclear prominence in its acoustic implementation due to many reasons such as informational structure. Such high agreement regarding nuclear prominences may arise from factors including more distinct acoustic information or distinct information status in a discourse. Thirdly,

in future research, the temporal modulation of subsyllabic components of disyllabic words with lexical stress on the first and on the second syllable will be examined to test whether there is a uniform lengthening domain, or multiple targets of lengthening as proposed by Turk and Shattuck-Hufnagel (2007). Lastly, like speaker-dependent prosody models, I would like to evaluate whether there are any listener-independent prosody models, or if there exists listener-dependent variability in prosody perception.

Appendix A

Buckeye Perception Projects

A.1 Subject PowerPoint[®] and Transcript examples




A.1.1 A sample Microsoft PowerPoint[®] page (chunk boundaries) constructed for the sound file presentation for each transcriber

Mark the location of “Chunk boundaries”

Sound 1			Sound 6		
Sound 2			Sound 7		
Sound 3			Sound 8		
Sound 4			Sound 9		
Sound 5					

A.1.2 A sample Microsoft PowerPoint® page (prominence) constructed for the sound file presentation for each transcriber

Mark the location of “Prominence”

Sound 1			Sound 6		
Sound 2			Sound 7		
Sound 3			Sound 8		
Sound 4			Sound 9		
Sound 5					

A.1.3 A sample corresponding orthographic word transcript

BS01PS06.L1.DOCX (page 1 of 6)

A. Mark the location of "Chunk boundaries"

1. Practice Session

i have a project i work on not as much as i probably should but it's the swan cleaners project and what we do we have little houses donation boxes in all the swan cleaners

2. Test Session

1) I've lived in columbus my entire life thirty four years um born and raised on the west side of columbus um <IVER> um i'm a um it's kind of a unique position i i guess i'm a um my job title is a senior research associate um i received my um bachelors masters p h d and nurse practitioner all from here from o s u and i'm working down at dodd hall in a research position i'm also a nurse practitioner down there that um sees patients in follow up and in clinic um they sort of created the position for me about a year ago after i had my little boy

2) <IVER> um working on three different um research projects with three different physicians two are l- um specific to the traumatic brain injury service looking at long term outcomes in people that have sustained traumatic brain injuries the other one is a drug study um looking at control of agitation and the other one is a um spinal cord injury research study <IVER> it's a no it's a uh clinical study where we're seeing how treadmill training effects ambulation in people with spinal cord injury <IVER> so they're kind of spread out twenty five twenty five ten percent that's kind of how it's divided out between the projects

3) a nurse practitioner has more authority i guess to diagnose and see patients um the state of ohio is i think the only state that has a given um nurse practitioners prescriptive authority where we can write our own prescriptions so we can see patients do the physicals we can uh suggest medications i guess per se but then the physician has to come in confirm our findings and then he actually writes the prescriptions

4) yes and that's what i like <IVER> correct <IVER> exactly <IVER> exactly <IVER> but we're working on it we've gotta couple house bills in right now and see how things go <IVER> be exciting <IVER> yes <IVER> i love ohio state but i don't i i guess from a career perspective they say it's not a good idea to get all your degrees from one place but <IVER> i don't intend to ever leave here so

5) i love it <IVER> yes <IVER> yes <IVER> yes i uh um uh lordy um grew up on the westside i went to **** my husband went to **** um proximity wise is probably within a mile of each other we were kind of high school sweethearts and the whole bit um his dad still lives in grove city my mom lives still at our old family house there on the westside and we moved um also on the westside probably couple miles from my mom

6) for me personally um family being close uh both sides of the family are here um for me familiar familiarity i don't like a lot of change a lot of new things um for me uh i was involved in i i swam for o s u swimming an got real involved with the o s u sports an just the sporting programs an i love ohio state football so i i like the sports there a lot of activities to do to now with my little boy going to cosi and the zoo and just knowing where things are and how to get around <IVER> so family um familiarity and um just the community i think

7) crime i think it's going up and up um traffic is horrendous but the one d- but i mean you need to go through that right now because columbus is expanding so much that we need the orange brailles and um i guess the lack of publicity nationally i i we don't have yknow like the national football teams and now with the hockey coming i i think were getting more recognition but when people talk about columbus they then i don't think they know columbus ohio or columbus georgia or columbus

8) i don't know that much about it to be quite honest i did read about it in the paper and i know the whole thing with um chief jackson and the whole um mayoral candidates how they're getting in debates over that an i know this is something new on top of that um i know it was the front page on sunday's paper i looked at it briefly um

9) i think there must be some truth to it i don't know the extent <IVER> oh it's possible but i can't imagine it would not knowing how big it's going to get i guess i can't imagine <IVER> that it's going to well it might make national news i just <IVER> it was on national news you see i'm not that informed on that um

10) there's so much negativity on the national news i don't know if that'll just be blended in or if people will associate that with columbus <IVER> right <IVER> uh <IVER> two and a half <IVER> yes <IVER> yes <IVER> cosi's going to be opening up on the sixth um looking forward to doing that we had a um uh seasons pass to wyandotte lake spent a lot of time at the pools um different swim centers hunter different rec centers the y programs are excellent even here at ohio state they've got the um um recreational sports on sunday afternoons where they can go um we go to the zoo a lot libraries have story times reading hours those type of things um just all kind of community things easter egg hunts at the rec centers <IVER> um

11) up until march he was at home with my aunt with family watching him in our home an then my grandfather who's ninety five got um ill and my aunt had to choose between my son and my grandfather and um she's watching actually my grandfather's doing quite well now but she um watched him so we now have him in a day care um in hilliard

12) again with the news reports with uh um different bombings and the i think the jewish school caught my attention and now that he is in daycare i mean we did a pretty good screening but you still always kind of wonder what's going on he's done quite well he's i mean he's learned how to use scissors how to color he's socially learning how to play with other kids those were all things he wasn't doing at home potty training's been a big deal because all the other kids in class are doing it and

13) he talks about biting even though he's never bit and um there's one particular boy right now when his family leaves he just has complete temper tantrums and he's tried that out a few times too he's one you can leave and it doesn't phase him he's very comfortable with other people but like last month or so he's tried that a few times and i think he picked that up at day care seeing the other kids do that

14) it's quite possible but we pick- the day care that i thought was closest to well the one that was open that had the closest values that we had um i personally would like a church based program but uh there aren't any well there weren't any open when we were looking for one um we did have the option of going here to o s u to their day care and it has a wonderful reputation an everybody that i've talked to that has had children there is very pleased with the program but what i didn't like was so that if they weren't um imparting values on the children they don't celebrate any holidays at all so they just kind of ignore everything where he is now i still like getting my mother's day card in

15) i mean they're doing trick or treat thing on thursday they're going door to door they get to dress up or he wouldn't have gotten any of that at o s u and there pros and cons to to both of those but <IVER> i guess almost at o s u i wish they would celebrate every holiday so then they're exposed to all of it instead of none of it

B. Mark the location of "Prominence"

1. Practice Session

they have both their children in a catholic prep school and yeah kindergarten and preschool and it's a yes it's it's a catholic school environment

2. Test Session

1) i was born here in columbus so i really haven't known anything else um have a dog named sammy i work in the dorms as a night assistant and i have a girlfriend right now <IVER> yeah <IVER> yeah my dog's at home <IVER> yeah i have two older brothers one's ten years older the other one's fifteen years older <IVER> yeah <IVER> um it didn't seem that much stranger than anyone else <IVER> i mean my the middle brother was always around he wasn't very social in high school so

2) um i'm a pre art major right now hoping to get into computer animation <IVER> freshman <IVER> i went to **** it's the new high school there <IVER> yeah high school a lot of people are really narrow minded and work on really trying to fit in in college people are more open idea like open minded and they're more willing to express their own ideas <IVER> i was more comfortable here than in high school

3) you don't see like the big cliques like the popular abercrombie wearing people or like all the art students hanging out together it's a big mix of people and everyone's doing their own thing <IVER> um freshman sophomore year yes after that i quit wrestling and i just kind of found myself being myself instead of trying to be someone else sports is definitely part of a clique you try to be part of the team and the team tries to do everything the same <IVER> so everybody's trying to be like everyone else

4) it was part of that i started wrestling because a lot of my friends did <IVER> so i mean basically the entire winter i'd either be doing stuff by myself or on the team wrestling with my friends so the whole reason i quit wrestling is because the coach wanted me to cut too much weight i was really starting to feel it physically <IVER> yeah i was cutting twenty pounds in three days

5) you'd run around in three pairs of sweat suits and vinyl jump suit over that under heaters <IVER> you'd sleep in three pairs of sweats so you'd lose about two and a half pounds in your sleep <IVER> it is that's why i quit <IVER> by the end i was pretty much just puking up blood and passing out a lot <IVER> there were a lot of teammates that were cutting that much weight if not more they were pretty much doing it for the team

6) yeah by then it was junior year when we were wrestling varisty and everyone just figured it was two more years left they could just struggle through it <IVER> it definitely did because i would sleep probably the eight hours right after practice until i woke up in the morning <IVER> i delivered the columbus dispatch in the morning too so <IVER> it was pretty much deliver papers go to class wrestling sleep <IVER> almost all the parents did <IVER> but he'd bring in nutritional list or whatever to explain to the parents blah blah blah

7) not that i know of none of the other sports teams were so inclined to like push their athletes to cut weight <IVER> or to do anything extreme to their bodies <IVER> um-hum <IVER> and the lower you are apparently the easier it is or something that was the coach's philosophy anyway <IVER> some of them are athletes they're pretty much just preppy student council i'm good at school but bad outside of school they'd go out and party outside of school but they'd join all like the students against drugs programs <IVER> so they're really hypocritical

8) there's a little bit well quite a bit of it was yeah some of the people in the group didn't play sports but it wasn't all from one sport or it was football wrestlers it was usually like the best football players the best wrestlers <IVER> in high school yeah i was friends with most of them <IVER> but i just didn't hang out with the entire crowd

9) it wasn't so much of a hate thing it was just more these are the people i like to hang out with if you don't like that that's fine <IVER> our school wasn't really a violent one there weren't really any fights or anything <IVER> it was mostly caucasian and very few asian and other ethnicities <IVER> yeah my roommate actually had that teacher and none of the allegations are true <IVER> she was alleging that students were writing racial slurs on their desks and stuff but i had friends in classes of hers and none of them wrote any <IVER> they never saw it they never said anything about it

10) a lot of people switch from being like the really quiet more reserved people to the party animal <IVER> go out and have fun person <IVER> just the number of people that were here <IVER> brother <IVER> brother jed <IVER> yeah <IVER> um i talk to him every now and then <IVER> yeah <IVER> yeah i think every college does <IVER> um i'm a lutheran and a lot of what he says about everyone going to hell <IVER> is really opposing my beliefs so i stand there and pretty much spar with him about it and he usually ends up getting frustrated and tells me to leave

11) well he told a jewish friend of mine that her grandparents that died in the holocaust went to hell because they were jewish <IVER> yeah he says a lot of racial slurs like that he's really homophobic <IVER> where as i don't know lutheran is a little bit more open to <IVER> homosexuality i guess

12) just pretty much that if you ask for forgiveness forgiveness is granted whatever <IVER> um doesn't just because you sin doesn't mean you're going to hell <IVER> and pretty much stuff like that and he'd try to quote the bible and i'd have a bible there with me and i'd look it up and it wouldn't it wouldn't be right at all

13) well when he's telling my friends that their grandparents went to hell because they're of a different religion <IVER> that's pretty extreme <IVER> yeah um i've grown up here all my life so it really hasn't shocked me all that much <IVER> yeah i had about three or four friends that were actually gay <IVER> it wasn't a huge shock it wasn't my first encounter with homosexuality

14) if they wanna get married that's their business i don't think its mine or anyone else to say that they can't <IVER> they're normal people just outside of their sexual well majority of the people prefer heterosexual <IVER> relations but they're normal people outside of they're homosexuality i mean it's nothing different than regular parents <IVER> i really am not educated in that at all so i wouldn't probably say anything since i'd ignorant in the field

A.2 Consent and instruction forms

A.2.1 5-minute instruction for subjects

Instructions for speech transcription experiment

In this experiment you will listen to some excerpts from recorded interviews and your task is to add marks to a transcript indicating the location of two features that you perceive in the speech: *prominence* and *chunking*. Let me describe these two features.

In normal speech, speakers pronounce some word or words in a sentence with more **prominence** than others. The prominent words are in a sense highlighted for the listener, and stand out from other non-prominent words. In some of the excerpts you will hear, you will be asked to mark all prominent words by underlining them.

example: ... word word word ...

Another feature of normal speech that we are interested in is the way speakers break up an utterance into **chunks**. These chunks group words in a way that helps the listener interpret the utterance, and are especially important when the speaker produces long stretches of continuous speech. An example of chunking that is familiar to everyone is the chunking that breaks digit sequences down into sub-groups:

example: 123 4567

For some of the excerpts you will hear, you will be asked to mark the chunks by inserting a vertical line between words that belong to different chunks. It is important for you to know that the boundary between two chunks does not necessarily correspond to the location where you would place a comma, period, or other punctuation mark, so you must really listen and mark the boundary where you here a juncture between two chunks. A chunk may be as small as a single word, or it may contain many words, and speakers can vary quite a bit in the size of the chunks they produce in a given utterance.

show example: ... word word | word ...

In marking the location of prominence and chunk boundaries, there may be differences between transcribers in how they perceive the features of the same utterance. Don't be concerned about how your transcription compares to anyone else's. There is not necessarily a single "correct" transcription, and we're interested in each of your transcriptions, individually.

A.2.2 Subject consent form

University of Illinois at Urbana-Champaign

Department of Linguistics

4088 Foreign Languages Building, MC-168

707 South Mathews Avenue

Urbana, IL 61801-3625

Informed Consent to Participate in Linguistics Experiment Directed by Jennifer S. Cole. Speech Transcription Experiment

You are invited to participate in a study of speech transcription. You will listen through headphones to a series of short excerpts (less than 20 sec. each) of recorded interviews from speakers of American English. As you listen, you will be asked to mark a printed transcript of the speech to indicate words that you hear as prominent and also for the location of boundaries between words that belong to different chunks. The entire experiment will last no more than 40 minutes, and is self-paced: you will control how quickly you progress through the speech files. If you feel uncomfortable or unable to complete the experiment you may stop at any time. When you complete the experiment, the experimenter will collect your transcription. There are no risks or discomfort expected as a result of your participation, beyond those of everyday life. You are free to withdraw from the study at any time for any reason. Your decision either to participate in this research or not to participate in it will in no way affect your evaluation in your linguistics courses now or in the future, and will not affect your present or future academic standing at UIUC.

Your performance in this study will be completely confidential. Your transcript will be identified by initials, age and gender, but no other personal information such as your name will be noted. After the experiment is complete your data will be securely stored in Professor Cole's data archive, identified only by your initials, age, gender and the date of the experiment.

You may not directly benefit from participation though there may be benefits to general knowledge or to society.

You are encouraged to ask any questions that you might have about this study before or after your participation. However, answers that could influence the outcome of the study will be deferred to the end of the experiment. Questions can be addressed to Jennifer Cole (jscole@uiuc.edu, tel. 244-3057). If you have any questions about research subjects' rights, or need to report any research-related injury, please contact the U of I Institutional Review Board at 333-2670 or irb@uiuc.edu.

I understand the above information and voluntarily consent to participate in the experiment described above. I am 18 years of age or older and I have been given a copy of this consent form.

SIGNATURE: _____ Date: _____

Approved by the University of Illinois Department of Linguistics IRB: _____
(Linguistics IRB member)

This consent form is valid through the following date: _____

A.2.3 Subject language survey form

Set No. _____ Date _____ Course Number Ling 100/ Ling 225
 Gender Male/ Female Age _____ Native country _____

1. What is your native language? _____
2. What language(s) did you use in your normal life during your childhood?

3. Have you ever learned other language(s) other than your first native language? If so, what language(s) have you studied? And how long?

	language	Years and months	Age when you first started study
1			
2			
3			
4			
5			
6			

4. Have you ever lived or stayed in a country where language(s) other than your native language is/are spoken for more than one month?

	Country	Years and months	Age when you first arrived in a country
1			
2			
3			
4			
5			
6			

5. Have you ever taken phonetics or phonology courses for English and/ or any other language(s)?

6. Have you ever participated in this experiment? Yes _____/ No _____

Appendix B

Programming Scripts

B.1 Sample scripts for PraatTM version 5.2.03

- Root-Mean-Square intensity with bandwidth filter

```
1 # This script is written by Yoonsook Mo, University of Illinois on March 25, 2008
2 # in order to get RMS intensity information from the Buckeye corpus sound files
3
4 form Configuration
5     sentence Directory c:\research\buckeye\perception_study_longset1\sound-tgrid-
6 longset2\
7     sentence Outputfile_pref buckeye_perception_longset2_phones_intensity
8     integer Bw_start 0
9     integer Bw_end 500
10 endform
11
12 #clearinfo
13
14
15 Create Strings as file list... list_wavs 'directory$'*wav
16 n_files = Get number of strings
17
18 outputfile$ = "'outputfile_pref$'_bw_start'-'bw_end'.txt"
19 filedelete 'directory$'outputfile$'
20
21 fileappend 'directory$'outputfile$' 'directory$'newline$'
22 fileappend 'directory$'outputfile$' bandwidth = 'bw_start'-'bw_end'newline$'
23 'newline$'
24
25 for t from 1 to n_files
26     select Strings list_wavs
27     filename$ = Get string... t
28     filename$ = filename$ - ".wav"
29
30     Read from file... 'directory$'filename$.wav
31     Filter (pass Hann band)... bw_start bw_end 0.1
32
33     Read from file... 'directory$'filename$.TextGrid
34
35     n_intervals = Get number of intervals... 2
36     tmin = Get starting point... 2 1
```

```

37     tmax = Get end point... 2 n_intervals
38
39     #fileappend 'directory$' 'outputfile$' 'filename$' 'n_intervals' 'tmin' 'tmax'
40 'newline$'
41     #print 'filename$', 'newline$'
42
43     for i from 1 to n_intervals
44         phone$ = Get label of interval... 2 i
45         start = Get starting point... 2 i
46         end = Get end point... 2 i
47
48         select Sound 'filename$'_band
49         rms_intensity = Get root-mean-square... start end
50         db = 20 * log10 (rms_intensity / 0.00002)
51         fileappend 'directory$' 'outputfile$' 'filename$' 'i'/'n_intervals'
52 'phone$' 'start' 'end' 'db' 'newline$'
53
54         select TextGrid 'filename$'
55     endfor
56     plus Sound 'filename$'_band
57     plus Sound 'filename$'
58     Remove
59
60 endfor
61 select Strings list_wavs
62 Remove

```

- Pitch values from sound files

```

1 # This script is written by Yoonsook Mo, University of Illinois on January 17, 2007
2 # in order to get Pitch values from the Buckeye corpus sound files
3
4 form Configuration
5     sentence Directory c:\research\buckeye\wav_textgrid\s15-s24\
6     sentence Outputfile buckeye_pitch_s15-s24.txt
7 endform
8
9 #clearinfo
10
11
12 Create Strings as file list... list_wavs 'directory$'*wav
13 n_files = Get number of strings
14
15 filedelete 'directory$' 'outputfile$'
16
17 for t from 1 to n_files
18     select Strings list_wavs
19     filename$ = Get string... t
20     filename$ = filename$ - ".wav"
21
22     Read from file... 'directory$' 'filename$'.wav
23     tmin = Get starting time
24     tmax = Get finishing time

```

```

25
26     Read from file... 'directory$' 'filename$'.TextGrid
27     n_intervals = Get number of intervals... 2
28     fileappend 'directory$' 'outputfile$' 'filename$' 'n_intervals' 'tmin' 'tmax'
29 'newline$'
30 #   print 'filename$', 'n_intervals', 'tmin', 'tmax' 'newline$'
31
32     plus Sound 'filename$'
33     Edit
34     editor TextGrid 'filename$'
35                                     ## interval tier1
36     Zoom... tmin tmin+5
37     Select next tier
38     Move cursor to... tmin
39     Select next interval
40     Select previous interval
41
42     for i from 1 to n_intervals
43         phone$ = Get label of interval
44         start = Get starting point of interval
45         end = Get end point of interval
46
47         time1 = start + (end - start) / 4
48         time2 = start + (end - start) / 2
49         time3 = start + (end - start) * 3 / 4
50
51         if (phone$ == "a" | phone$ == "aa" | phone$ == "aan" | phone$ == "ae" |
52 phone$ == "aen" | phone$ == "ah" | phone$ == "ahn" | phone$ == "an" | phone$ ==
53 "ao" | phone$ == "aon" | phone$ == "aw" | phone$ == "awn" | phone$ == "ay" |
54 phone$ == "ayn" | phone$ == "e" | phone$ == "eh" | phone$ == "ehn" | phone$ ==
55 "el" | phone$ == "em" | phone$ == "en" | phone$ == "eng" | phone$ == "er" |
56 phone$ == "ern" | phone$ == "ey" | phone$ == "eyn" | phone$ == "i" | phone$ ==
57 "id" | phone$ == "ih" | phone$ == "ihn" | phone$ == "iy" | phone$ == "iyn" |
58 phone$ == "ow" | phone$ == "own" | phone$ == "oy" | phone$ == "oyn" | phone$ ==
59 "uh" | phone$ == "uhn" | phone$ == "uw" | phone$ == "uwn")
60
61         Move cursor to... time1
62         pitch1$ = Get pitch
63
64         Move cursor to... time2
65         pitch2$ = Get pitch
66
67         Move cursor to... time3
68         pitch3$ = Get pitch
69
70         Move cursor to... start
71
72         fileappend 'directory$' 'outputfile$' 'filename$' 'i'/'n_intervals'
73 'phone$' 'start' 'pitch1$' 'pitch2$' 'pitch3$' 'newline$'
74         #print 'filename$' 'n_intervals' 'i' 'phone$' 'start' 'pitch1$'
75 'pitch2$' 'pitch3$' 'newline$'
76
77     else
78         fileappend 'directory$' 'outputfile$' 'filename$' 'i'/'n_intervals'

```

```

79 'phone$' 'start''newline$'
80
81     endif
82
83     #print 'filename$' 'i'/'n_intervals' 'phone$' 'start''newline$'
84     Select next interval
85
86     endfor
87
88     Close
89     Remove
90
91 endfor
92

```

B.2 Sample scripts for PythonTM version 2.6

- Z-normalization of intensity measures

```

1 # This script is written by Yoonsook Mo, University of Illinois on March 25, 2008
2 # in order to calculate normalized intensities measured from the Buckeye corpus
3
4
5 import sys, glob, string, math
6
7 def vowelcheck(phone):
8     if 'a' in phone or 'e' in phone or 'i' in phone or 'o' in phone or 'u' in phone:
9         return 1
10    else:
11        return 0
12
13 input_file = ['C:\\research\\buckeye\\perception_study_longset1\\sound-tgrid\\
14 formants-intensities\\buckeye_perception_longset1_phones_intensity_0-500.txt',
15             'C:\\research\\buckeye\\perception_study_longset1\\sound-tgrid-
16 longset2\\formants-intensities\\buckeye_perception_longset2_phones_intensity_
17 0-500.txt']
18
19 #input_file = ['C:\\research\\buckeye\\perception_study_longset1\\sound-tgrid\\
20 formants-intensities\\buckeye_perception_longset1_phones_intensity.txt',
21 #             'C:\\research\\buckeye\\perception_study_longset1\\sound-tgrid-
22 longset2\\formants-intensities\\buckeye_perception_longset2_phones_intensity.txt']
23
24 vowel_list = {}
25
26 for file in input_file:
27
28     input = open(file, 'r')
29     lines = input.readlines()
30     input.close()
31
32

```

```

33     for aline in lines:
34         line = aline.strip().split()
35         if len(line) == 6 and line[5] != '--undefined--':
36             phone = line[2]
37             intensity = float(line[5])
38
39             if vowelcheck(phone) and len(phone) <= 2:
40                 #print aline
41                 if phone in vowel_list:
42                     vowel_list[phone].append(intensity)
43                 else:
44                     vowel_list[phone] = [intensity]
45
46
47 #ccount = 0
48 stat_list = {}
49 for phone in vowel_list:
50     #ccount += len(vowel_list[phone])
51
52     max = 0.0
53     sum = 0.0
54     for value in vowel_list[phone]:
55         sum += value
56         if value > max:
57             max = value
58     average = sum / len(vowel_list[phone])
59
60     sum2 = 0.0
61     for value in vowel_list[phone]:
62         sum2 += (value - average) * (value - average)
63
64     if len(vowel_list[phone]) >= 2:
65         stdev = math.sqrt(sum2 / len(vowel_list[phone]))
66     else:
67         stdev = 0.0
68
69     stat_list[phone] = (average, stdev, max)
70     #print id, phone, i, average, stdev
71
72
73
74 for file in input_file:
75
76     input = open(file, 'r')
77     lines = input.readlines()
78     input.close()
79     output_file = file.replace('.txt', '_zscore.txt')
80     output = open(output_file, 'w')
81
82     for aline in lines:
83         line = aline.strip().split()
84         if len(line) == 6 and line[5] != '--undefined--':
85             phone = line[2]
86             intensity = float(line[5])

```

```

87
88         if vowelcheck(phone) and len(phone) <= 2:
89             if phone in stat_list:
90                 #print intensity, stat_list[phone][2], stat_list[phone][1]
91                 if stat_list[phone][1] == 0:
92                     zscore = 0.0
93                 else:
94                     zscore = (math.log(intensity) - math.log(stat_list[phone]
95 [2])) / math.log(stat_list[phone][1])
96
97                 for i in range(0, len(line)):
98                     output.write(line[i] + ' ')
99                     output.write(str(zscore) + '\n')
100             else:
101                 output.write(aline)
102         else:
103             output.write(aline)
104     else:
105         output.write(aline)
106
107 output.close()

```

- Pitch Median filtering and Interpolation

```

1 # This script is written by Yoonsook Mo, University of Illinois on July 13, 2008
2 # in order to get interpolated Pitch values following median filtering
3
4 import string, sys, wave, glob
5 from scipy import signal, interp
6
7 # read pitch information from the PRAAT result
8 def read_pitch(file):
9
10     lines = open(file).readlines()
11     pitch_x = []
12     pitch_y = []
13
14     tmin = 1.0
15     tmax = 0.0
16
17     for line in lines:
18         items = line.strip().split()
19         time = float(items[0])
20         if time <= tmin:
21             tmin = time
22         if time >= tmax:
23             tmax = time
24
25         pitch = float(items[1])
26         pitch_x.append(time) # change time into frame number
27         pitch_y.append(pitch)
28
29     return pitch_x, pitch_y, tmin, tmax

```



```

30
31
32 def read_textgrid(file):
33
34     lines = open(file).readlines()
35
36     pos1 = 0
37     pos2 = 0
38     pos3 = 0
39
40     output_data = []
41     output_data2 = []
42
43     for line in lines:
44         words = line.strip().split()
45         if 'item [1]' in line:
46             pos1 = 1
47         if 'item [2]' in line:
48             pos1 = 2
49
50         if pos1 == 0:
51             if len(words) == 3 and words[0] == "xmin":
52                 xmin = float(words[2])
53             elif len(words) == 3 and words[0] == "xmax":
54                 xmax = float(words[2])
55
56         if pos1 == 1:
57             if 'intervals: size' in line:
58                 pos2 = 1
59             if pos2 == 1:
60                 if len(words) == 3 and words[0] == 'xmin':
61                     start = float(words[2])
62                 elif len(words) == 3 and words[0] == 'xmax':
63                     end = float(words[2])
64                 elif len(words) == 3 and words[0] == 'text':
65                     text = words[2].strip(' "')
66                     output_data.append((start, end, text))
67
68         elif pos1 == 2:
69             if 'intervals: size' in line:
70                 pos3 = 1
71             if pos3 == 1:
72                 if len(words) == 3 and words[0] == 'xmin':
73                     start = float(words[2])
74                 elif len(words) == 3 and words[0] == 'xmax':
75                     end = float(words[2])
76                 elif len(words) == 3 and words[0] == 'text':
77                     text = words[2].strip(' "')
78                     output_data2.append((start, end, text))
79
80     return output_data, output_data2, xmin, xmax
81
82
83 #directory="f:\\pitch_median_interpolate\\input_data\\"

```

```

84 directory="c:\\research\\buckeye\\pitch_median_interpolate\\input_data\\"
85
86 PITCH_INTERVAL = 1 # 1ms
87 TIER_GAP = 0.005
88 MEDIAN_FRONT = 0.01 # 10ms range before consonant-vowel split point
89 MEDIAN_REAR = 0.02 # 20ms range after consonant-vowel split point
90 MEDIAN_FILTER_ARRAY_SIZE = 13
91
92 textgrid_list = glob.glob(directory + "*.TextGrid")
93
94
95 for filename in textgrid_list:
96     pitch_file = filename.lower().replace(".textgrid", ".pitch")
97     wav_file = filename.lower().replace(".textgrid", ".wav")
98
99     wav = wave.open(wav_file, 'r')
100     params = wav.getparams()
101     wavRate = params[2]
102     nframes = params[3]
103
104     base = filename.replace(directory, '')
105     output_file = filename.lower().replace(".textgrid", "_pitch.txt")
106     output = open(output_file, 'w')
107
108
109     # read words and phones information from a TextGrid file
110     (words_list, phones_list, tmin, tmax) = read_textgrid(filename)
111
112     # read pitches from a calculated pitch file from praat
113     # x in frame number, y in Hertz
114     (pitch_x, pitch_y, pitch_x_min, pitch_x_max) = read_pitch(pitch_file)
115
116     original_pitch = []
117     for pitch in pitch_y:
118         original_pitch.append(pitch)
119
120     n_pitches = len(pitch_y)
121     pitch_x_min_ms = int(round(pitch_x_min*1000))
122
123     print base, tmin, n_pitches
124
125     # median filtering
126     phone_status = 0 # 0 for UNDECIDED
127     pre_phone_status = 0 # 1 for VOWEL
128                                     # 2 for CONSONANTS
129
130     vowels = []
131     consonants = []
132
133     for tuple in phones_list:
134         start = tuple[0]
135         end = tuple[1]
136         phone = tuple[2]
137
138         if 'NOISE' not in phone and 'SIL' not in phone and 'EXCLUDE' not in

```

```

138 phone:
139         if 'a' in phone or 'e' in phone or 'i' in phone or 'o' in phone or
140 'u' in phone:
141             for word in words_list:
142                 i = 0
143                 if start >= word[0] - TIER_GAP and end <= word[1] +
144 TIER_GAP:
145                     i += 1
146                     selected_word = word[2]
147                 if i > 1:
148                     selected_word = "DOUBLEY_MATCHED"
149
150                 phone_status = 1 #vowel
151                 vowels.append((start, end, phone, selected_word))
152                 #print base, start, phone, selected_word
153
154                 if pre_phone_status == 2:
155                     start_rounded = int(round(start*1000)) # nearest pitch time
156 (ms) for starting point of the phone
157                     start_window = start_rounded - int(round(MEDIAN_FRONT*1000))
158                     end_window = start_rounded + int(round(MEDIAN_REAR*1000))
159
160                     #print start, start_rounded, start_window, end_window
161
162                     pre_filtered = []
163                     median_filtered = []
164
165                     if start_window < 0:
166                         print "negative start_window in ", base, start, phone,
167 selected_word
168
169                     else:
170                         for t in range(start_window, end_window + 1, PITCH_
171 INTERVAL):
172                             pre_filtered.append(pitch_y[t-pitch_x_min_ms])
173
174                             median_filtered = signal.medfilt(pre_filtered, MEDIAN_
175 FILTER_ARRAY_SIZE)
176                             #print start, phone, median_filtered, pre_filtered
177
178                             for i in range(0, len(median_filtered), PITCH_INTERVAL):
179                                 #print start_window+i-pitch_x_min_ms, i, pitch_y
180 [start_window+i-pitch_x_min_ms], median_filtered[i], pitch_y[start_window+i-
181 pitch_x_min_ms] - median_filtered[i]
182                                 pitch_y[start_window-pitch_x_min_ms+i] = median_
183 filtered[i]
184                                 #print start_window+i-pitch_x_min_ms, i, pitch_y
185 [start_window+i-pitch_x_min_ms], original_pitch[start_window+i-pitch_x_min_ms],
186 pitch_y[start_window+i-pitch_x_min_ms] - original_pitch[start_window+i-pitch_x_
187 min_ms]
188
189                 else:
190                     for word in words_list:
191                         i = 0

```

```

192         if start >= word[0] - TIER_GAP and end <= word[1] + TIER_GAP:
193             i += 1
194             selected_word = word
195         if i > 1:
196             selected_word = "DOUBLEY_MATCHED"
197
198         phone_status = 2
199         consonants.append((start, end, phone, selected_word))
200         #print base, start, phone, selected_word
201
202     pre_phone_status = phone_status
203
204
205     med_output_filename = filename.lower().replace(".textgrid", "_all_pitch_
206 median.txt")
207     med_output = open(med_output_filename, 'w')
208
209     # med_output.write(base+' (median filtered)\n')
210     # med_output.write('sampling rate = '+str(wavRate)+'\n')
211     # med_output.write('no. of frames = '+str(nframes)+'\n')
212     # med_output.write('start = '+str(tmin)+'\n')
213     # med_output.write('end = '+str(tmax)+'\n\n')
214
215
216     for j in range(0, n_pitches):
217         med_output.write(str(pitch_x[j])+'\t'+str(pitch_y[j])+'\t'+str(original_
218 pitch[j])+'\n')
219
220     med_output.close()
221
222
223     # interpolation
224     # count zero pitch values in the front and back of file
225
226     front_zero = 0 # count no. of front zero pitches
227     nonzero_pitch = 0 # 1 if pass by non-zero pitch value
228     for i in range(0, n_pitches):
229         if pitch_y[i] == 0 and nonzero_pitch == 0:
230             front_zero += 1
231         elif pitch_y[i] != 0:
232             nonzero_pitch += 1
233
234     back_zero = 0 # count no. of back zero pitches
235     nonzero_pitch = 0 # 1 if pass by non-zero pitch value
236     for i in range(n_pitches-1, -1, -1):
237         if pitch_y[i] == 0 and nonzero_pitch == 0:
238             back_zero += 1
239         elif pitch_y[i] != 0:
240             nonzero_pitch += 1
241
242     n_interp_pitches = n_pitches - front_zero - back_zero
243
244     # prepare pitch list by removing all zero pitches
245

```

```

246     strip_pitch_x = []    # list of pitch values excluding zero values in between
247     strip_pitch_y = []
248
249     for i in range(0, n_pitches):
250
251         if pitch_y[i] != 0:
252             strip_pitch_x.append(pitch_x[i])
253             strip_pitch_y.append(pitch_y[i])
254
255
256     # reference x-list removing only front/back zero pitches
257
258     interp_pitch_count = 0
259     interp_pitch_x = []
260
261     for i in range(0, n_pitches):
262
263         if i >= front_zero and interp_pitch_count < n_interp_pitches:
264             interp_pitch_x.append(pitch_x[i])
265             interp_pitch_count += 1
266
267
268     # do the interpolation
269
270     interp_pitch_y = interp(interp_pitch_x, strip_pitch_x, strip_pitch_y)
271
272     for i in range(0, front_zero):
273         output.write(str(pitch_x[i])+'\t'+str(original_pitch[i])+'\t'+str(pitch_
274 y[i])+'\t'+ '---'+'\n')
275
276     for i in range(0, n_interp_pitches):
277         output.write(str(pitch_x[i+front_zero])+'\t'+str(original_pitch[i+front_
278 zero])+'\t'+str(pitch_y[i+front_zero])+'\t'+str(interp_pitch_y[i])+'\n')
279
280     for i in range(0, back_zero):
281         output.write(str(pitch_x[i+front_zero+n_interp_pitches])+'\t'+str
282 (original_pitch[i+front_zero+n_interp_pitches])+'\t'+str(pitch_y[i+front_zero+n_
283 interp_pitches])+'\t'+ '---'+'\n')
284
285
286     output.close()
287
288
289

```

- Normalized maximum and minimum pitch values in word-last-phones

```

1 # This script is written by Yoonsook Mo, University of Illinois on August 27, 2009
2 # in order to get normalized max/min pitch values in word-last phones
3
4 import sys, glob, string, math
5
6 def vowelcheck(phone):

```

```

7         if 'a' in phone or 'e' in phone or 'i' in phone or 'o' in phone or 'u' in
8 phone:
9             return 1
10        else:
11            return 0
12
13
14        dir = 'e:\\research\\buckeye\\pitch_median_interpolate\\'
15        #files = glob.glob(dir + 'input_data_longset\\*.txt')
16        #input_file = dir + 'phones_longset_results.txt'
17
18        #output_file = dir + 'last_phones_LS_pitch_normalized_results.txt'
19
20        files = glob.glob(dir + 'input_data\\*.txt')
21        input_file = dir + 'phones_set1+2_results.txt'
22        output_file = dir + 'last_phones_SS_pitch_normalized_results.txt'
23
24
25        output = open(output_file, 'w')
26        #output.write("speaker"+'\t'+ "word"+'\t'+ "wmin"+'\t'+ "wmax"+'\t'+ "phone"+'\t'+
27 "pmin"+'\t'+ "pmax"+'\t'+ "min_pitch_t"+'\t'+ "min_pitch"+'\t'+ "min_pitch_z"+'\n')
28        output.write("speaker"+'\t'+ "phone"+'\t'+ "pmin"+'\t'+ "pmax"+'\t'+ "word"+'\t'+
29 "wmin"+'\t'+ "wmax"+'\t'+ "min_pitch_t"+'\t'+ "min_pitch"+'\t'+ "min_pitch_z"+'\n')
30
31        pitch_files = []
32
33
34        #### gather pitch file names
35
36        for file in files:
37            if '_pitch.txt' in file:
38                pitch_files.append(file)
39
40        #### read pitch values
41        for file in pitch_files:
42
43            names = file.split('\\')
44            base = names[len(names)-1].replace('_pitch.txt', '')
45
46            if '-1-1' in base:
47                base = base.replace('-1-1', '-1')
48            elif '-2-1' in base:
49                base = base.replace('-2-1', '-2')
50
51            #print base, "is started"
52
53            pitches = []
54            input2 = open(file, 'r')
55            lines2 = input2.readlines()
56            input2.close()
57
58            t_min = 0.0
59            t_max = 0.0
60

```

```

61     for line in lines2:
62         items = line.strip().split()
63         if len(items) == 4:
64             if items[3] != '---':
65                 pitch = float(items[3])
66                 pitches.append((time, pitch))
67                 time = float(items[0])
68         if t_min == 0.0 or time <= t_min:
69             t_min = time
70         if t_max <= time:
71             t_max = time
72
73
74
75     ##### read phones information
76
77     input = open(input_file, 'r')
78     lines = input.readlines()
79     input.close()
80
81     prev_word = ""
82     prev_xmin = -100.0
83     prev_line = ""
84
85     for line in lines:
86         items = line.strip().split()
87
88         if len(items) != 0 and base in items[0]:
89             if len(items) < 8:
90                 print items
91
92             else:
93                 word = items[5]
94                 xmin = float(items[6])
95
96             ##### find the last phone of the word
97             if prev_word != word and prev_xmin != xmin and prev_xmin != -100.0:
98
99                 #print prev_line, line, prev_xmin
100                items2 = prev_line.strip().split()
101
102                phone = items2[2]
103                pmin = float(items2[3])
104                pmax = float(items2[4])
105
106                #max_pitch = 0.0
107                min_pitch = 0.0
108
109                #max_pitch_x = -100.0    # arbitrary negative number
110                min_pitch_x = -100.0
111
112            ##### find the min max pitches in the phone duration
113            for pitch in pitches:
114                if pitch[0] >= pmin and pitch[0] <= pmax:

```

```

115         #if pitch[1] >= max_pitch:
116         #     max_pitch = pitch[1]
117         #     max_pitch_x = pitch[0]
118         if min_pitch == 0.0 or pitch[1] <= min_pitch:
119             min_pitch = pitch[1]
120             min_pitch_x = pitch[0]
121
122     norm_min = min_pitch_x - 0.2
123     norm_max = min_pitch_x + 0.2
124
125     if norm_min < t_min:
126         norm_min = t_min
127         norm_max = t_min + 0.4
128     elif norm_max > t_max:
129         norm_min = t_max - 0.4
130         norm_max = t_max
131
132     norm_pitches = []
133     norm_sum_pitch = 0.0
134     norm_sum_pitch_diff_sq = 0.0
135
136     ##### normalization around min_pitch_x (400 ms window)
137     for pitch in pitches:
138         if pitch[0] >= norm_min and pitch[0] <= norm_max:
139             norm_pitches.append(pitch[1])
140             norm_sum_pitch += pitch[1]
141
142
143     ##### find out average and stdev of pitches in normaliztion window
144
145     if len(norm_pitches) == 0:
146         norm_avg_pitch = 0
147     else:
148         norm_avg_pitch = norm_sum_pitch / len(norm_pitches)
149
150     for value in norm_pitches:
151         norm_sum_pitch_diff_sq += (value - norm_avg_pitch) * (value
152 - norm_avg_pitch)
153
154     if len(norm_pitches) == 0:
155         norm_std_pitch = 0.0
156     else:
157         norm_std_pitch = math.sqrt(norm_sum_pitch_diff_sq / len(norm_
158 pitches))
159
160
161
162     ##### calculated standard values of max and min pitch values
163
164     if norm_std_pitch == 0.0:
165         #max_pitch_z = 0.0
166         min_pitch_z = 0.0
167
168     else:

```



```
169         #max_pitch_z = (max_pitch - norm_avg_pitch) / norm_std_pitch
170         min_pitch_z = (min_pitch - norm_avg_pitch) / norm_std_pitch
171
172
173
174         ##### print out calculated pitch values
175         for a in items2:
176             output.write(a+'\t')
177
178             output.write(str(min_pitch_x)+'\t'+str(min_pitch)+'\t'+str(min_
179 pitch_z)+'\n')
180
181
182         ##### memorized previous line information
183         prev_word = word
184         prev_xmin = xmin
185         prev_line = line
186
187         print base, "is done!!"
188         output.write('\n\n')
```

References

- D. Abercrombie. Syllable and quantity and enclitics in english. In D. Abercrombie, editor, *In honor of Daniel Jones: papers contributed on the occasion of his eightieth birthday*. Longman, London, 1964.
- P. Adank, R. Smits, and R. van Hout. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5):3099–3107, 2004.
- L. Aguilar, A. Bonafonte, F. Campillo, and D. Escudero. Determining intonational boundaries from the acoustic signal. In *The proceedings of Interspeech*, Brighton, UK, 2009.
- S. Ananthakrishnan and S. Narayanan. Fine-grained pitch accent and boundary tone labeling with parametric F0 features. In *The proceedings of IEEE international conference of Acoustics, Speech, and Signal Processing*, pages 4545–4548, 2008.
- J. E. Arnold. THE BACON not the bacon: How children and adults understand accented and unaccented noun phrases. *Cognition*, 108:69–99, 2008.
- J. E. Atkinson. Correlation analysis of the physiological factors controlling fundamental voice frequency. *Journal of the Acoustical Society of America*, 63:211–222, 1978.
- M. Aylett and A. Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56, 2004.
- M. Beckman and G. Ayers. Guidelines for ToBI labeling (version 3.0). Manuscript and accompanying speech materials, 1997. URL http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.
- M. E. Beckman. *Stress and non-stress*. Foris Publications, Dordrecht, The Netherlands, 1986.
- M. E. Beckman and J. Edwards. Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston and M. E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, pages 152–178. Cambridge University Press, Cambridge, 1990.

- M. E. Beckman and J. Edwards. Articulatory evidence for differentiating stress categories. In P. A. Keating, editor, *Papers in laboratory phonology III: Gesture, segment and prosody*, pages 7–33. Cambridge University Press, Cambridge, 1994.
- M. E. Beckman and S.-A. Jun. K-ToBI (Korean ToBI) labeling conventions, 1996. URL <http://www.linguistics.ucla.edu/people/jun/ktobi/k-tobi-V2.html>. version 2.1, revised November 1996.
- M. E. Beckman and J. Pierrehumbert. Intonational structure in English and Japanese. *Phonology yearbook*, 3:255–310, 1986.
- M.E. Beckman, J. Edwards, and J. Fletcher. Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty and D. R. Ladd, editors, *Papers in Laboratory Phonology II: Gesture, segment, prosody*, pages 68–86. Cambridge University Press, Cambridge, 1992.
- A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *Journal of the Acoustical Society of America*, 113(2):1001–1024, 2003.
- R. Berkovits. Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, 21:155–180, 1993a.
- R. Berkovits. Progressive utterance-final lengthening in syllables with final fricatives. *Language and Speech*, 36:89–98, 1993b.
- R. Berkovits. Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, 37:237–250, 1994.
- P. Boersma and D. Weenink. *Praat: doing phonetics by computer*, 5.0.30 edition, 2005.
- D. Bolinger. A theory of pitch accent in English: accent, morpheme, order. In I. Abe and T. Kanekiyo, editors, *Forms of English: accent, morpheme, order*. Harvard University Press, Cambridge, MA, 1958.
- G. Bruce, B. Granstrom, K. Gustafson, and D. House. Interaction of F0 and duration in the perception of prosodic phrasing in Swedish. In B. Granstrom and L. Nord, editors, *Nordic Prosody VI: Papers from a symposium*, pages 7–22. Almqvist & Wiksell International, 1993.
- I. Bulyko and M. Ostendorf. Joint prosody prediction and unit selection for concatenative speech synthesis. In *The proceedings of ICASSP*, pages 781–784, Salt Lake City, Utah, 2001.
- D. Byrd. Relations of sex and dialect to reduction. *Speech Communication*, 15:39–54, 1994.

- D. Byrd. Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, 57(1):3–16, 2000.
- D. Byrd and D. Rigg. Locality interactions with prominence in determining the scope of phrasal lengthening. *Journal of the International Phonetic Association*, 38:187–202, 2008.
- D. Byrd and E. Saltzman. Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, 26(2):173–200, 1998.
- D. Byrd and E. Saltzman. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2):149–180, 2003.
- D. Byrd, J. Krivokapic, and S. Lee. How far, how long: On the temporal scope of prosodic boundary effects. *Journal of the Acoustical Society of America*, 120(3):1589–1599, 2006.
- S. Calhoun. *Information structure and the prosodic structure of English: A probabilistic relationship*. PhD thesis, University of Edinburgh, 2006.
- T. Cambier-Langeveld. The domain of final lengthening in the production of Dutch. In J. Coerts and H. D. Hoop, editors, *Linguistics in the Netherlands*, pages 13–24. John Benjamins, Amsterdam, 1997.
- T. Cambier-Langeveld and A. E. Turk. A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics*, 27(3):255–280, 1999.
- T. Cambier-Langeveld, M. Nespors, and V. J. van Heuven. The domain of final lengthening in production and perception in Dutch. In *The proceedings of Eurospeech*, Rhodes, Greece, 1997.
- K. Carlson, C. Jr. Clifton, and L. Fraizer. Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45:58–81, 2001.
- R. Carlson and M. Swerts. Relating perceptual judgments of upcoming prosodic breaks to F0 features. *Phonum*, 9:181–184, 2003a.
- R. Carlson and M. Swerts. Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials. In *The proceedings of the International Congress of Phonetic Sciences*, Barcelona, Spain, 2003b.
- R. Carlson, J. Hirschberg, and M. Swerts. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46(3/4):326–333, 2005.
- S. Chavarria, T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson. Acoustic differentiation of ip and IP boundary level: Comparison of L- and L-LIn *The proceedings of Speech Prosody*, Nara, Japan, 2004.

- T. Cho. *The effects of prosody on articulation in English*. Routledge, New York, 2002.
- T. Cho. Manifestation of prosodic structure in articulatory variation: evidence from lip kinematics in English. In L. M. Goldstein, D. H. Whalen, and C.T. Best, editors, *Papers in Laboratory Phonology VIII: Varieties of phonological competence*, pages 519–548. Mouton de Greyter, Berlin, 2006.
- T. Cho and P. A. Keating. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, 29(2):155–190, 2001.
- T. H. Cho. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of (a,i) in English. *Journal of the Acoustical Society of America*, 117(6):3867–3878, 2005.
- T. H. Cho. Prosodic strengthening in transboundary V-to-V lingual movement in American English. *Phonetica*, 65(1–2):45–61, 2008.
- J. Y. Choi, M. Hasegawa-Johnson, and J. Cole. Finding intonational boundaries using acoustic cues related to the voice source. *Journal of the Acoustical Society of America*, 118(4):2579–2587, 2005.
- J. Cole, H. Kim, H. Choi, and M. Hasegawa-Johnson. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from radio news speech. *Journal of Phonetics*, 35:180–209, 2007.
- J. Cole, L. Goldstein, A. Katsika, Y. Mo, E. Nava, and M. Tiede. Perceived prosody: Phonetic bases of prominence and boundaries. In *Presented at 156th Meeting of the Acoustical Society of America*, Miami, FL, 2008.
- J. Cole, Y. Mo, and S. Baek. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, 25(7–9):1141–1177, 2010.
- R. Collier. Physiological correlates of intonation patterns. *Journal of the Acoustical Society of America*, 58(1):249–255, 1975.
- W. E. Cooper, S. J. Eady, and P. R. Mueller. Acoustical aspects of contrastive stress in question answer contexts. *Journal of the Acoustical Society of America*, 77(6): 2142–2156, 1985.
- T. H. Crystal and A. S. House. Segmental durations in connected speech signals: Current results. *Journal of the Acoustical Society of America*, 83(4):1553–1573, 1988a.
- T. H. Crystal and A. S. House. Segmental durations in connected speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, 83(4):1574–1585, 1988b.

- A. Cutler, D. Dahan, and W. vanDonselaar. Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2):141–201, 1997.
- D. Dahan, M. K. Tanenhaus, and C. G. Chambers. Accent and reference resolution in spoken language comprehension. *Journal of Memory and Language*, 47:292–314, 2002.
- L. Dilley and S. Shattuck-Hufnagel. Variability in glottalization of word onset vowels in American English. In *The proceedings of the 8th International Congress of Phonetic Sciences (ICPhS)*, pages 586–589, Stockholm, Sweden, 1995.
- L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24:423–444, 1996.
- L. C. Dilley, M. Breen, E. Gibson, M. Bolivar, and J. Kraemer. A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (rhythm and pitch) and ToBI (tones and break indices). In *The proceedings of the International Conference on Spoken Language Processing*, Pittsburg, PA, 2006.
- S. J. Eady and W. E. Cooper. Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, 80:402–415, 1986.
- W. Eefting. The effect of information value and accentuation on the duration of Dutch words, syllables, and segments. *Journal of the Acoustical Society of America*, 89(1):412–424, 1991.
- Y. Erikson and M. Alstermark. Fundamental frequency correlates of the grave word accent in Swedish: the effect of vowel duration. *STL-QPSR*, 13(2–3):48–60, 1972.
- G. Fant and A. Kruckenberg. Preliminaries to the study of Swedish prose reading and reading style. *TMH-QPSR*, 2:1–89, 1989.
- G. Fant and A. Kruckenberg. Notes on stress and word accent in Swedish. *STL-QPSR*, 35(2–3):125–144, 1994.
- G. Fant, S. Hertegård, and A. Kruckenberg. Focal accent and subglottal pressure. *TMH-QPSR*, 2(2):29–32, 1996.
- G. Fant, A. Kruckenberg, S. Hertegård, and J. Liljencrants. Sub- and supraglottal pressures in speech. In R. Bannert, M. Heldner, K. Sullivan, and P. Wretling, editors, *The Proceedings of Fonetik 97 (Phonum)*, volume 4, pages 25–28, Umeå, Sweden, 1997.
- G. Fant, A. Kruckenberg, J. Liljencrants, and S. Hertegård. Acoustic-phonetic studies of prominence in Swedish. *TMH-QPSR*, 41(2–3):1–51, 2000a.
- G. Fant, A. Kruckenberg, and J. Liljencrants. Acoustic-phonetic analysis of prominence in Swedish. In A. Botinis, editor, *Intonation: Analysis, modeling and technology*, volume 55–86. Springer, Dordrecht, The Netherlands, 2000b.

- G. Fant, A. Kruckenberg, and J. B. Ferreira. Individual variations in pausing: A study of read speech. *PHONUM*, 9:193–196, 2003.
- F. Ferreira. Creation of prosody during sentence production. *Psychological Review*, 100(2):233–253, 1993.
- L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. In *The proceedings of ICSLP*, Denver, CO, 2002.
- J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.
- C. Fougeron. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, 29(2):109–135, 2001.
- A. Fox. *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford University Press, New York, 2000.
- D. B. Fry. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4):765–768, 1955.
- D. B. Fry. Experiments in the perception of stress. *Language and Speech*, 1(2):126–152, 1958.
- J. P. Gee and F. Grosjean. Performance structures - a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15(4):411–458, 1983.
- R. D. Glave and A. C. M. Rietveld. Is effort dependence of speech loudness explicable on basis of acoustical cues. *Journal of the Acoustical Society of America*, 58(4):875–879, 1975.
- J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *The proceedings of the IEEE International Conference on the Acoustics, Speech and Signal Processing (ICASP)*, pages 517–520, San Francisco, CA, 1992.
- S. Greenberg, H. Carvey, and L. Hitchcock. The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *In Proceedings of ISCA Workshop on Prosody in Speech Processing (Speech Prosody 2002)*, Aix-en-Provence, 2002.
- S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang. Temporal properties of spontaneous speech- a syllable-centric perspective. *Journal of Phonetics*, 31:465–485, 2003.
- M. Grice, M. Reyelt, M. Benzmuller, J. Mayer, and J. Batliner. Consistency in transcription and labeling of German intonation with GToBI. In *The proceedings of the International Conference on Spoken Language Processing*, pages 1716–1719, Philadelphia, PA, 1996.

- C. Gussenhoven and A. C. M. Rietveld. Fundamental-frequency declination in Dutch - testing 3 hypotheses. *Journal of Phonetics*, 16(3):355–369, 1988.
- C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken. The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America*, 102(5):3009–3022, 1997.
- H. M. Hanson. Vowel amplitude variation during sentence production. In *The proceedings of ICASSP*, pages 1627–1630, Munich, Germany, 1997.
- P. Hansson. Prosodic phrasing in spontaneous Swedish. In *Travaux de l'institut de linguistique de Lund 43*, Dept. of Linguistics and Phonetics, Lund University, Sweden, 2003.
- M. Heldner. *Focal accent-F0 movements and beyond*. PhD thesis, Umea University, 2001a.
- M. Heldner. Spectral emphasis as a perceptual cue to prominence. *TMH-QPSR*, 42(1):51–57, 2001b.
- M. Heldner. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, 31(1):39–62, 2003.
- M. Heldner and B. Megyesi. The acoustic and morpho-syntactic context of prosodic boundaries in dialogs. *PHONUM*, 9:117–120, 2003.
- M. Heldner and E. Strangert. To what extent I perceived focus determined by F0-Cues? *The proceedings of Eurospeech. Rhodes, Greece*, pages 875–877, 1997.
- M. Heldner and E. Strangert. Temporal effects of focus in Swedish. *Journal of Phonetics*, 29(3):329–361, 2001.
- R. Herman, M. Beckman, and K. Honda. Suglottal pressure and final lowering in English. In *The proceedings of ICSLP*, Philadelphia, PA, 1996.
- D. J. Hermes and H. H. Rump. Perception of prominence in speech intonation induced by rising and falling pitch movements. *Journal of the Acoustical Society of America*, 96(1):83–92, 1994.
- J. Hirschberg and J. Pierrehumbert. Intonational structuring of discourse. In *The Proceedings of the 24th Meeting of the Association for Computational Linguistics*, pages 136–144, 1986.
- L. Hitchcock and S. Greenberg. Vowel height is intimately associated with stress accent in spontaneous American English. In *The proceedings of Eurospeech*, Aalborg, Denmark, 2001.
- M. Horne, E. Strangert, and M. Heldner. Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In *The proceedings of the International Congress of Phonetic Sciences*, pages 170–173, Stockholm, Sweden, 1995.

- R. Huddleston and G. K. Pullum. *The Cambridge grammar of the English language*. Cambridge University Press, Cambridge, 2002.
- S.-A. Jun. *The phonetics and phonology of Korean prosody*. PhD thesis, Ohio state university, 1993.
- S. A. Jun. The accentual phrase in the Korean prosodic hierarchy. *Phonology*, 15(2): 189–226, 1998.
- S. A. Jun. Prosodic phrasing and attachment preferences. *Journal of psycholinguistic research*, 32:219–249, 2003.
- D. Jurafsky. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, and S. Jannedy, editors, *Probabilistic Linguistics*. MIT Press, Cambridge, 2002.
- E. Kainada. Prosodic boundary effects on durations and vowel hiatus in modern Greek. In *The Proceedings of the International Congress of Phonetic Sciences*, pages 1225–1228, Saabruken, Germany, 2007.
- S. Kang and S. Speer. Prosody and clause boundaries in Korean. In B. Bernard and I. Marlien, editors, *The proceedings of Speech Prosody 2002*, pages 419–421. Aix-en-Provence: Laboratoire Parole et Langage, Université de Provence, 2002.
- S. Kang and S. R. Speer. Prosodic disambiguation of participle constructions in English. In *The proceedings of Speech Prosody*, Nara, Japan, 2004.
- P. A. Keating, T. Cho, C. Fougeron, and C. Hsu. Domain-initial strengthening in four languages; in local, ogden, temple. In *Papers in Laboratory Phonology. 6: Phonetic interpretations*, pages 145–163, Cambridge, 2003. Cambridge University Press.
- H. Kim, T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson. Acoustic differentiation of L- and L-L% in switchboard and radio news speech. In *The proceedings of Speech Prosody*, Dresden, Germany, 2006.
- M. M. Kjelgaard and S. P. Speer. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, 40:153–194, 1999.
- D. H. Klatt. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3:129–140, 1975.
- D. H. Klatt. Linguistic uses of segmental duration in English - acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59(5):1208–1221, 1976.
- D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

- G. Kochanski. Prosody beyond fundamental frequency. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schlieber, editors, *Methods in empirical prosody research*. De Gruyter, New York, 2006.
- G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2):1038–1054, 2005.
- K. J. Kohler. Glottal stops and glottalization in German. *Phonetica*, 51:38–51, 1994.
- K. J. Kohler. The perception of prominence patterns. *Phonetica*, 65:257–269, 2008.
- T. Kraljic and S. E. Brennan. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, 50:194–231, 2005.
- J. Krivokapic. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35(2):162–179, 2007.
- D. R. Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 2nd edition, 2008.
- P. Ladefoged and D. E. Broadbent. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1):98–104, 1957.
- A. Leemann, K. Hirose, and H. Fujisaki. Analysis of voice fundamental frequency contours of continuing and terminating prosodic phrases in four Swiss German dialects. In *The proceedings of Interspeech*, Brighton, UK, 2009.
- I. Lehiste and R. A. Fox. Influence of duration and amplitude on the perception of prominence by Swedish listeners. *Speech Communication*, 13(1–2):149–154, 1993.
- M. Liberman. *The intonational system of English*. PhD thesis, MIT, 1975.
- M. Liberman and J. Pierrhumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language sound structure*, pages 157–233. MIT Press, Cambridge, MA, 1984.
- M. Liberman and A. Prince. Stress and linguistic rhythm. *Linguistic Inquiry*, 8(2): 249–336, 1977.
- P. Lieberman. Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32(4):451–454, 1960.
- P. Lieberman. On the acoustic basis of the perception of intonation by linguists. *Word-Journal of the International Linguistic Association*, 21(1):40–54, 1965.
- P. Lieberman. *Intonation, perception and language*. MIT Press, Cambridge, MA, 1967.

- J. Liljencrants, G. Fant, and A. Kruckenberg. Subglottal pressure and prosody in Swedish. In *The Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing*, pages 1–4, Beijing, China, 2000.
- H.-Y. Lin and J. Fon. Perception of temporal cues at discourse boundaries. In *The proceedings of Interspeech*, Brighton, UK, 2009.
- K. Maekawa. Effects of focus on duration and vowel formant frequency in Japanese. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing prosody: Computational models for processing spontaneous speech*, pages 95–116. Springer-verlag, New York, 1997.
- R. S. McGowan and E. L. Saltzman. Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics*, 23(1–2):255–269, 1995.
- S. Mozziconacci. *Speech variability and emotion: Production and perception*. Technische Universiteit Eindhoven, Eindhoven, 1998.
- S. Nakai, S. Kunnari, A. Turk, K. Suomi, and R. Ylitalo. Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics*, 37:29–45, 2009.
- C. H. Nakatani. Integrating prosodic and discourse modeling. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing prosody: Computational models for processing spontaneous speech*, pages 67–80. Springer-verlag, New York, 1997.
- M. Nespør and I. Vogel. Prosodic structure above the word. In A. Cutler and D. R. Ladd, editors, *Prosody: models and measurements*, pages 123–140. Springer-Verlag, New York, 1983.
- M. Nespør and I. Vogel. *Prosodic Phonology. Studies in generative grammar 28*. Foris Publications, Dordrecht, The Netherlands, 1986.
- S. G. Nootboom. *Production and perception of vowel duration*. PhD thesis, Utrecht University, 1972.
- J. Ogata, M. Goto, and K. Itou. The use of acoustically detected filled and silent pauses in spontaneous speech recognition. In *The proceedings of IEEE international Conference on Acoustics, Speech, and Signal Processing*, pages 4305–4308, 2009.
- J. J. Ohala. Respiratory activity in speech. In W. Hardcastle and A. Marchal, editors, *Speech production and speech modeling*, pages 23–53. Kluwer, Dordrecht, The Netherlands, 1990.
- M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. The boston university radio news corpus, 1995.
- S. Peppe, J. Maxim, and B. Wells. Prosodic variation in Southern British English. *Language and Speech*, 43(3):309–334, 2000.

- J. Pierrehumbert. Perception of fundamental-frequency declination. *Journal of the Acoustical Society of America*, 66(2):363–369, 1979.
- J. Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, MIT, 1980.
- J. Pierrehumbert and M. Beckman. *Japanese Tone Structure, Linguistic Inquiry Monograph 15*. MIT Press, Cambridge, 1988.
- J. Pierrehumbert and J. Hirschberg. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in communications*. MIT Press, Cambridge, MA, 1990.
- J. Pitrelli, M. E. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability. In *The proceedings of the International Conference on Spoken Language Processing*, Yokohama, Japan, 1994.
- M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45:89–95, 2005.
- M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. Buckeye corpus of conversational speech (2nd release). Department of Psychology, Ohio State University (Distributor), 2007. URL <http://www.buckeyecorpus.osu.edu>.
- L. C. W. Pols, H. R. C. Tromp, and R. Plomp. Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53(4):1093–1101, 1973.
- P. Prieto, A. Estrella, J. Thorson, and M. M. Vanrell. Is prosodic development correlated with grammatical development? Evidence from emerging intonation in Catalan and Spanish. *Journal of Child Language*, submitted.
- W. Raymond, M. Pitt, K. Johnson, E. Hume, M. Makashay, R. Dautricourt, and C. Hilts. An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. In *The proceedings of ICSLP 02*, Denver, CO, 2002.
- L. Redi and S. Shattuck-Hufnagel. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29:407–429, 2001.
- A. C. M. Rietveld and C. Gussenhoven. On the relation between pitch excursion size and prominence. *Journal of Phonetics*, 13(3):299–308, 1985.
- I. Roca and W. Johnson. *A course in phonology*. Blackwell Publishers, Oxford, 1999.
- K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.

- H. H. Rump. *Prominence of pitch-accented syllables*. PhD thesis, Eindhoven Technical University, 1996.
- E. Saltzman, H. Nam, J. Krivocapic, and L. Goldstein. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *The proceedings of the Speech Prosody*, Campinas, Brazil, 2008.
- A. Sanderman. *Prosodic phrasing: Production, perception, acceptability and comprehension*. PhD thesis, Eindhoven University of Technology, 1996.
- E. O. Selkirk. *Phonology and Syntax: The relation between sound and structure*. MIT Press, Cambridge, MA, 1984.
- E. O. Selkirk. On derived domains in sentence phonology. *Phonology yearbook*, 3: 371–405, 1986.
- S. Shattuck-Hufnagel, L. Dilley, N. Veilleux, A. Brugos, and R. Speer. f_0 peaks and valley aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. In *The proceedings of Speech Prosody*, Nara, Japan, 2004.
- T. I. Shevchenko and T. S. Skopintseva. Prosody variation in English: Geographical, social, situational. In *The proceedings of Speech Prosody*, Nara, Japan, 2004.
- R. Silipo and S. Greenberg. Automatic transcription of prosodic stress for spontaneous English discourse. In *The proceedings of the 14 th International Congress of Phonetic Sciences*, pages 2351–2354, 1999.
- R. Silipo and S. Greenberg. Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of the NIST Speech Transcription Workshop*, College Park, MD, 2000.
- K. E. A. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg. ToBI: A standard for labeling English prosody. In *The proceedings of ICSLP*, pages 867–870, Alberta, Canada, 1992.
- A. M. C. Sluijter. *Phonetic correlates of stress and accent*. PhD thesis, Holland Institute of Generative Linguistics, 1995.
- A. M. C. Sluijter and V. J. van Heuven. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, 52 (2):71–89, 1995.
- A. M. C. Sluijter and V. J. van Heuven. Acoustic correlates of linguistic stress and accent in Dutch and American English. In *The proceedings of ICSLP 96*, Philadelphia, PA, 1996a.
- A. M. C. Sluijter and V. J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4):2471–2485, 1996b.

- A. M. C. Sluijter and V. J. van Heuven. Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in American English. *Manuscript*, 1996c.
- A. M. C. Sluijter, S. Shattuck-Hufnagel, K. N. Stevens, and V. J. van Heuven. Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in English. In *The proceedings of ICPHS 95*, pages 630–633, Stockholm, Sweden, 1995.
- A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, 1997.
- C. Smith. Topic transitions and durational prosody in reading aloud: Production and modeling. *Speech Communication*, 42:247–270, 2004.
- E. Strangert and M. Heldner. The labeling of prominence in Swedish by phonetically experienced transcribers. In *The proceedings of 8th ICPHS*, pages 13–19, Stockholm, Sweden, 1995.
- E. Strangert and M. Helnder. Labeling of boundaries and prominence by phonetically experienced and non-experienced transcribers. *Phonum*, 3:85–109, 1995.
- H. Strik and L. Boves. A physiological model of intonation. In *Proceeding of the Department of Language and Speech, University of Nijmegen*, volume 16/17, pages 96–105, 1994.
- H. Strik and L. Boves. Downtrend in F₀ and P_{sb}. *Journal of Phonetics*, 23(1–2): 203–220, 1995.
- A. K. Syrdal and J. McGory. Inter-transcriber reliability of ToBI prosodic labeling. In *The proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.
- J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge, 1990.
- S. Tanaka and W. J. Gould. Relationships between vocal intensity and noninvasively obtained aerodynamic parameters in normal subjects. *Journal of the Acoustical Society of America*, 73:1316–1321, 1983.
- J. Terken. Fundamental-frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89(4):1768–1776, 1991.
- J. Terken. Fundamental-frequency and perceived prominence of accented syllables .II. nonfinal accents. *Journal of the Acoustical Society of America*, 95(6):3662–3665, 1994.

- J. Terken and S. G. Nootboom. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, 2:145–164, 1987.
- I. R. Titze. On the relation between subglottal pressure and fundamental frequency in phonation. *Journal of the Acoustical Society of America*, 85(2):901–906, 1989.
- J. Trouvain and M. Grice. The effect of tempo on prosodic structure. In *The proceedings of the International Congress of Phonetic Sciences*, pages 1067–1070, San Francisco, CA, 1999.
- H. Truckenbrodt. *Phonological phrases: Their relation to syntax, focus, and prominence*. PhD thesis, Massachusetts Institute of Technology, 1995.
- A. E. Turk and J. R. Sawusch. The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America*, 99(6):3782–3790, 1996.
- A. E. Turk and J. R. Sawusch. The domain of accentual lengthening in American English. *Journal of Phonetics*, 25(1):25–41, 1997.
- A. E. Turk and S. Shattuck-Hufnagel. Word-boundary-related duration patterns in English. *Journal of Phonetics*, 28:397–440, 2000.
- A. E. Turk and S. Shattuck-Hufnagel. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35:445–472, 2007.
- A. E. Turk and L. White. Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2):171–206, 1999.
- M. Ueyama. An experimental study of vowel duration in phrase-final contexts in Japanese. *UCLA Working Papers in Phonetics*, 97:174–182, 1999.
- N. Umeda. Vowel duration in American English. *Journal of the Acoustical Society of America*, 58(2):434–445, 1975.
- D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12(1):1–23, 1993.
- J. J. Venditti. Prosody in sentence processing. In R. Mazuka, M. Nakayama, and Y. Shirai, editors, *The Handbook of East Asian Psycholinguistics, Volume II: Japanese*, pages 208–217. Cambridge University Press, Oxford, 2006.
- K. Vicsi and G. Szaszak. Prosodic cues for automatic phrase boundary detection in ASR. In *Lecture Notes in Computer Science*, pages 547–553. Springer, Berlin, 2006.
- A. Wagner. Acoustic cues for automatic determination of phrasing. In *The proceedings of Speech Prosody*, Chicago, IL, 2010.

- D. Watson and E. Gibson. The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19(6): 713–755, 2004.
- D. Watson, J. E. Arnold, and M.K. Tanenhaus. Not just given and new: The effects of discourse and task based constraints on acoustic prominence. In *Presented at the 2005 CUNY human sentence processing conference*, Tucson, AZ, 2005.
- P. Welby. Effects of pitch accent position, type, and status on focus projection. *Language and Speech*, 46(1):53–81, 2003.
- C. W. Wightman, S. Shattuckhufnagel, M. Ostendorf, and P. J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91(3):1707–1717, 1992.
- M. Yaeger-Dror. Register as a variable in prosodic analysis: The case of the English negative. *Speech Communication*, 19(1):39–60, 1996.
- L.-C. Yang. Duration, pause and the temporal structure of mandarin conversational speech. In *The proceedings of the International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007.
- T.-J. Yoon. Speaker consistency in the realization of prosodic prominence in the Boston University radio speech corpus. In *Presented at Speech Prosody 2010 Satellite Workshop: Prosodic Prominence Perceptual and Automatic Identification*, Chicago, IL, 2010.
- T.-J. Yoon, S. Chavarria, J. Cole, and M. Hasegawa-Johnson. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of Interspeech*, pages 2722–2732, Jeju, Korea, 2004.
- T.-J. Yoon, J. Cole, and M. Hasegawa-Johnson. On the edge: Acoustic cues to layered prosodic domains. In *The Proceedings of the International Congress of Phonetic Sciences*, pages 1017–1020, Saarbrücken, Germany, 2007.